# Tamil Handwriting Recognition using Subspace and DTW based Classifiers

Niranjan Joshi[1], G. Sita[1], A.G. Ramakrishnan[1]
and Sriganesh Madhvanath[2]

[1] Dept. of Electrical Engg., Indian Institute of Science, Bangalore, INDIA
[2] Hewlett-Packard Laboratories, Bangalore, India
e-mail: {niranjan,sita,ramkiag}@ee.iisc.ernet.in,
srig@hplabs.com

Corresponding Author:

G. Sita
Department of Electrical Engg.
Indian Institute of Science
Bangalore - 560 012, INDIA

# Tamil Handwriting Recognition using Subspace and DTW based Classifiers

Niranjan Joshi[1], G. Sita[1], A.G. Ramakrishnan[1],Sriganesh Madhvanath[2]

[1] Dept. of Electrical Engg., Indian Institute of Science, Bangalore, INDIA
[2] Hewlett-Packard Laboratories, Bangalore, India
e-mail: {niranjan,sita,ramkiag}@ee.iisc.ernet.in, srig@hplabs.com

**Abstract.** In this paper, we report the results of recognition of online handwritten Tamil characters. We experimented with two different approaches. One is subspace based method wherein the interactions between the features in the feature spate are assumed to be linear. In the second approach, we investigated an elastic matching technique using dynamic programming principles. We compare the methods to find their suitability for an on-line form-filling application in writer dependent, independent and adaptive scenarios. The comparison is in terms of average recognition accuracy and the number of training samples required to obtain an acceptable performance. While the first criterion evaluates effective recognition capability of a scheme, the second one is important for studying the effectiveness of a scheme in real time applications. We also perform error analysis to determine the advisability of combining the classifiers.

## 1 Introduction

Handwriting recognition is a desirable attribute for real time operation of hand held systems where the resources are limited and the devices are too small to have full sized keyboards. Cursive English script recognition is already an inbuilt feature in pocket sized Personal Digital Assistants (PDA) with very high recognition accuracies. For a good review of online handwriting recognition, see [1]. Online handwriting recognition is especially very relevant in Indian scenario, as symbols requiring long key stroke sequences are very common in Indian languages. It also eliminates the need to adapt to any complex key stroke sequences and handwriting input is faster compared to any other text input mechanism for Indian languages. Given the complexity of entering the Indian scripts, using a keyboard, handwriting recognition has the potential to simplify and thereby revolutionize data entry for Indian languages.

The challenges posed by Indian languages are different from English. In addition, there has been very little research on machine recognition of Indian scripts. Consequently, exhaustive experimentation is necessary in order to get a good insight into the script from machine recognition point of view. In this paper, we address the problem of online handwriting recognition of Tamil which is a popular South Indian language and also one of the official languages in countries such as Singapore, Malaysia, Sri Lanka.

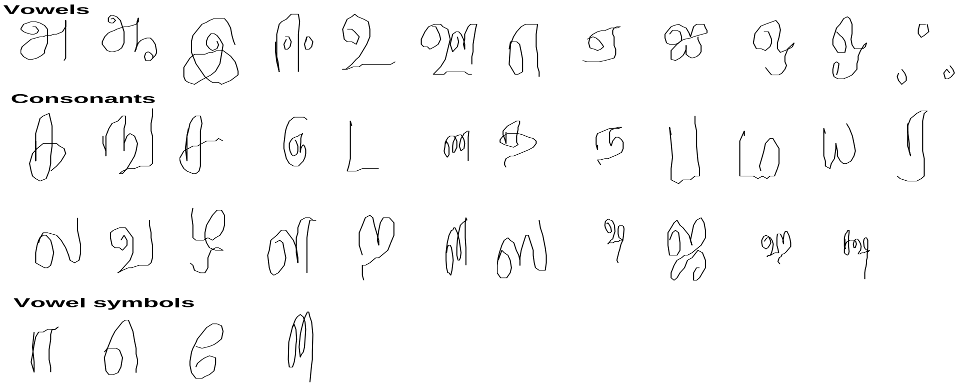


**Fig. 1.** Tamil character set

There are 156 distinct symbols/characters in Tamil of which 12 are pure vowels and 23 are pure consonants. This set of 35 characters are the basic character units of the script and the remaining character classes are vowel-consonant combinations. They are composed of two parts, namely the basic character and a modifier symbol corresponding to each of the basic character. Although in most of the cases, the basic character and the modifier are written in separate strokes, in the present work, we consider them to be written in single stroke. Fig. 1 presents the basic Tamil character set. In south Indian scripts such as Tamil, Malayalam, Kannada, and Telugu, characters are written in isolation. Hence, in the current work, each character is considered as a separate class for recognition. The input is a temporally ordered sequence of $(x, y)$ pen coordinates corresponding to an isolated character obtained from the digitizer.

We experimented with two different approaches for character recognition. One is principal component analysis based method wherein each character class is modeled as a subspace. A consequence of this is that whenever a core pattern and its variations occur, all the linear combinations of these patterns are treated as members of the class. This is equivalent to synthesizing patterns by taking linear combinations. The second approach uses dynamic programming principles for recognition and uses elastic matching . A comparison of both the methods is carried out to find their suitability for an online form filling application in writer dependent, independent and adaptive scenarios. In writer dependent case, the character model is built using the individual writer's data only. In writer independent case, the data of the writer under consideration is not a part of the training set. In writer adaptive case, the training set consists of the data of all other writers in addition to a part of the current writer's data. In all the three cases, the training data is different from the test data.

## 2 Preprocessing

We use a Pocket PC for dynamic capture of the handwritten characters. The input from the digitizer corresponding to a handwritten character is a sequence of points of the form $(x_i, y_i)$ with embedded pen-up and pen-down events when multiple strokes are involved. Pre-processing is required in order to compensate for variations in time and scale, and can be classified into two steps - smoothing and normalization. Smoothing is performed to reduce the amount of high frequency noise in the input resulting from the digitizer or tremors in writing.In our scheme, each stroke is smoothed independently using a 5-tap Gaussian low-pass filter. Normalization is carried out to account for variability in character size and pen velocity. The details of these operations are given in [5].

# 3 Methods

## 3.1 Subspace based classification

It is essentially a linear transformation of the feature space. By selecting the principal directions in which variance is significant, the feature space can be approximated by a lower order space. We use this method to model each character class as a subspace. As a consequence of this, whenever we have a core pattern and its variations, all the linear combinations of these patterns are treated as members of the class. The method is briefly desscribed below. For more details, refer [3].

Let the $N$ training vectors of a particular class be $(\mathbf{x}_1, \cdots, \mathbf{x}_N)$. The correlation matrix is defined as,

$$\mathbf{R}_{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i'$$

For finding the principal components, we solve the eigen value equation,

$$\lambda \mathbf{v} = \mathbf{R}_x \mathbf{v}$$

In this fashion, the basis vectors for each class $k$ are computed as a set of $N$ eigen vectors $\mathbf{v}_j^k, j = 1, \cdots, N$. Each eigen vector is normalized so that the basis is orthonormal. For a given test vector $\mathbf{x}_{test}$, its projection distance to the subspaces spanned by individual character classes is used as a measure to recognize its correct class lable. In this work, subspace spanned by the first 11 eigen vectors is used after experimentation.

An advantage of subspace method is its ability to approximate the feature vector in a low dimensional space which leads to a reduction in the running time in real time applications. This is possible as normally the smallest eigen values correspond to spurious variations in the character. Therefore, selecting a subset of the original subspace increases the accuracy of classification.

## 3.2 DTW based classification

Dynamic time warping (DTW) is a elastic matching technique. It allows nonlinear alignment of sequences and computes a more sophisticated similarity measure. This is especially useful to compare patterns in which rate of progression varies non-linearly which makes similarity measures such as euclidean distance and cross-correlation unusable. Classifiers using DTW-based distances have been shown to be well suited for handwriting recognition task by several researchers [1, 4] .

Suppose we have two time series $Q = (q_1, \cdots, q_n)$ and $C = (c_1, \cdots, c_m)$, of length $n$ and $m$ respectively. To align two sequences using DTW, we construct an $n$-by-$m$ matrix where the $(i, j)^{th}$ element is the Euclidean distance $d(q_i, c_j) = (q_i - c_j)^2$ between the two points $q_i$ and $c_j$. Each matrix element $(i, j)$ corresponds to the alignment between the points $q_i$ and $c_j$. A warping path, $W$ is a contiguous set of matrix elements that defines a mapping between $Q$ and $C$ and is written as $W = w_1, \cdots, w_K$ where $max(m, n) \leq K < m + n - 1$. The warping path is typically subject to several constraints such as, boundary conditions, continuity, monotonicity, and windowing [4]. The DTW algorithm finds the point-to-point correspondence between the curves which satisfies the above constraints and yields the minimum sum of the costs associated with the matchings of the data points. There are exponentially many warping paths that satisfy the above conditions. The path that minimizes the warping cost is,

$$DTW(Q, C) = min\{\sqrt{\sum_{k=1}^{K} w_k/K}$$

The warping path can be found very efficiently using dynamic programming to evaluate the following recurrence which defines the cumulative distance $\gamma(i, j)$ as the distance $d(i, j)$ found in the current cell and the minimum of the cumulative distances of the adjacent elements.

In order to resolve the confusion among character classes and reduce computational time, in this method classification stage is divided into two steps. In the first pre-classification step, we use Euclidean distance as a measure for obtaining the possible character candidates. This is followed by fine classification using $(x, y)$ coordinates as features on the output classes given by pre-classification step. Both the steps use DTW as the distance measure. More details of this method can be found in [6].

## 4 Database

The database is collected from 15 native Tamil writers using a custom application running on a Pocket PC (Compaq iPAQ 3850). It contains vowels, consonants, vowel-consonant combinations and special characters totalling 156 symbols. This set covers all the discrete symbols that make up all characters in Tamil. Ten samples of each of the symbols under consideration were collected from each writer totalling 23400 samples. The ten datasets from each writer are collected at different times to avoid fatigue in the writers which would reflect in the hand writing. In the user interface of the training mode, the character to be written is displayed and the user has to write in a given writing area. In the testing mode, the user writes in boxes obviating the need for character segmentation.

## 5 Experimental Results

The objective of the current investigation is to evaluate the performance of both subspace based and DTW based methods so as to combine the advantages

**Table 1.** Samples used in WD,WI and WA modes

| Mode | Training samples | Test samples |
|---|---|---|
| WI | 19440 | 4860 |
| WD | 1134 | 486 |
| WA | 23814 | 486 |

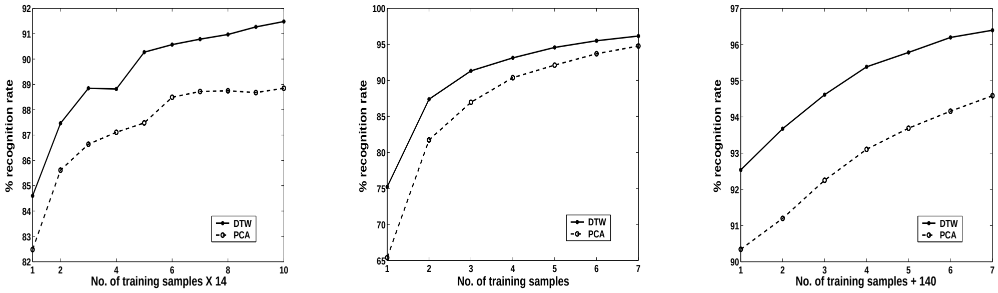
of the two schemes to formulate subsequently a hybrid scheme for handwriting recognition. The comparison is carried out for all the three modes, namely, writer dependent (WD), writer independent (WI) and writer adaptive (WA). The difference in the three modes is only in the training dataset. For writer independent recognition, we use "leave one out" strategy. In this strategy, out of the 15 writers data, data from 12 writers is used for training the recognizer. The remaining 3 writers data is used as the test data. The recognizer is trained to recognize the variations of a specific writer only in the case of WD recognition. Hence, out of the ten data sets of a particular class of a given writer, seven are used as training set to model a class and the remaining three datasets are used for testing. In WA case, the recognizer is trained to incorporate a much larger variability in the writing style by including other writers' data along with the specific writer's data. Table. 1 presents the number of samples used in all the three modes for training as well testing the recognizer's performance. All the experiments are run using a Pentium IV processor with 512 MB RAM.

In subspace method, we experimented using different number of principal components to approximate the chracter classes under conisderation. The mean accuracy of a given classification scheme is computed by averaging the accuracies computed across all writers. This is compared across WD, WI and WA modes. It is found that the gain in recognition accuracy is not very significant if we further increase the number of principal components beyond 11. Hence, in the rest of the investigation, we considered only first 11 principal components for subspace based studies.

The comparison is carried out with respect to the average class recognition accuracy for different number of training samples. Average recognition accuracy is found out by dividing number of correctly recognized test patterns with total number of test patterns. This leads to the minimum number of training samples required for an acceptable recognition performance for each of the schemes. Fig. 2 shows that in each of the modes, the DTW method outperforms subspace method in terms of recognition accuracy. In writer adaptive case, the performance of both methods is good as a larger writing style variability is incorporated in the model by including other writers data in the training set.

**Table 2.** Comparison of error rates for subspace and DTW based methods

| Mode | PCA | DTW | Common |
|---|---|---|---|
| WI | 11.15 | 8.52 | 3.56 |
| WD | 5.23 | 3.30 | 1.47 |
| WA | 5.41 | 3.60 | 1.20 |



(a) (b) (c)

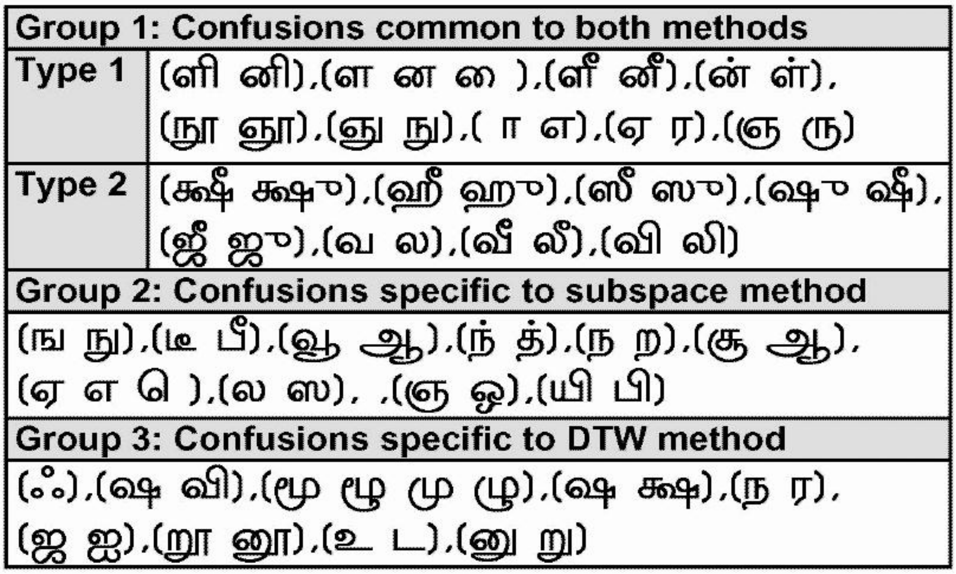
**Fig. 2.** Recognition for (a) WI case (b) WD case (c) WA case

## 6 Error analysis

In this section, we present the analysis of confused characters in both the methods. For this, we manually check the structure of the misrecognized test samples. Only those errors are considered, which occur frequently. Fig. 3 shows set of confusion pairs/triplets/quadruples. *Group 1* confusions are observed with both classifiers. Confusions of type 1 occur due to a ”loop” getting confused with either a ”cusp” or a ”straight line”. Confusions of type 2 occur because of a sharp ”corner” getting confused with a ”curved” part. It is apparent that structures of most of the characters involved in these types of confusions are very much similar. Therefore both the methods which provide global (dis)similarity measure get easily confused within these characters. However this observation provides some important clues for further modifications. Confused characters shown in the figure along with rest of the characters similar to them can be grouped together for first level of classification. Further classification can be performed by giving importance to local features. Rest two groups of errors shown in Fig. 3 are specific only to a certain method. In Table 2, a group-wise error comparison of the methods in each of the modes of operation, namely, WI,WD,WA is presented. The second and third columns of the table present the percentage errors obtained with each of the methods and for different modes. In the other columns, group wise errors are presented. Failure of subspace method is possibly due to estimated subspaces for confused characters are very ”near” or overlapping. Errors in DTW method are mainly because its elastic matching capability overfits the template. However since these set of errors are of non-overlapping in nature a classifier combination scheme combining these two classifiers could prove helpful for improving overall accuracy.

## 7 Conclusions

The suitability of two different schemes, namely subspace based method and dynamioc time warping(DTW) based methods, for online handwriting recognition of Tamil script is investigated in three different writer modes. Although, the performance of DTW based method is marginally better, in terms of speed subspace based method wins over. To overcome the data dependency of subspace method,



| Group 1: Confusions common to both methods | |
|---|---|
| Type 1 | (ளி னி),(ள ன ை ),(ளீ னீ),(ன் ள்),<br>(நூ ஞூ),(ஞு நு),( ா எ),(ஏ ர),(ஞ ரு) |
| Type 2 | (க்ஷீ க்ஷு),(ஹீ ஹு),(ஸீ ஸு),(ஷு ஷீ),<br>(ஜீ ஜு),(வ ல),(வீ லீ),(வி லி) |
| **Group 2: Confusions specific to subspace method** | |
| (ங நு),(டீ பீ),(லூ ஆ),(ந் த்),(ந ற),(சூ ஆ),<br>(ஏ எ ெ ),(ல ஸ), ,(ஞ ஒ),(யி பி) | |
| **Group 3: Confusions specific to DTW method** | |
| (ஂ),(ஷ வி),(மூ ழூ மு ழு),(ஷ க்ஷ),(ந ர),<br>(ஜ ஐ),(றூ ஞூ),(உ ட),(ஞு று) | |

**Fig. 3.** Confusion set

and use the advantage of elastic and nonlinear matching capability of DTW, hierarchical classification schemes are currently being investigated. Although both the methods are studied for a specific real time application, at present recognition speed is not being compared as essentially the objective is to reap the advantages of both the methods. Although DTW based method is computationally expensive, it can be overcome by using prototype selection/reduction methods.

## References

1. Charles C. Tappert, Ching Y. Suen, and Tory Wakahara, "The state of the art in on-line handwriting recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12(8), 1990, pp. 787-808.
2. C. S. Sundaresan and S. S. Keerthi, "A study of representations for pen based handwriting recognition of Tamil characters," *Fifth International Conference on Document Analysis and Recognition*, Sep. 1999, pp.422-425.
3. Deepu V., "On-line writer dependent handwriting character recognition," Master of Engineering project report, *Indian Institute of Science*, India, Jan. 2003.
4. E. Keogh and M. Pazzani, " Derivative dynamic time warping," *First SIAM International Conference on Data Mining (SDM'2001)*, Chicago, USA, 2001.
5. X. Li, D.Y. Yeung, "On-line handwritten alphanumeric character recognition using dominant points in strokes," *Pattern Recognition*, 30(1), 1997, pp. 31-44.
6. Niranjan Joshi, G. Sita, A.G. Ramakrishnan, and Sriganesh Madhvanath, "Comparison of elastic matching algorithms for on-line Tamil handwriting recognition," ICONIP'04, Kolkatta, 2004.