# Elastic Matching of Online Handwritten Tamil and Telugu Scripts Using Local Features

L.Prasanth,V.Jagadeesh Babu, R. Raghunath Sharma,
G.V.Prabhakara Rao,
*Sri Sathya Sai Institute of Higher Learning,*
*Puttaparthy, India*

Dinesh M
*Hewlett-Packard Labs,*
*Bangalore, India*

## Abstract

*This paper describes character based elastic matching using local features for recognizing online handwritten data. Dynamic Time Warping (DTW) has been used with four different feature sets: x-y features, Shape Context (SC) and Tangent Angle (TA) features, Generalized Shape Context feature (GSC) and the fourth set containing x-y, normalized first and second derivatives and curvature features. Nearest neighborhood classifier with DTW distance was used as the classifier. In comparison, the SC and TA feature set was found to be the slowest and the fourth set was best among all in the recognition rate. The results have been compiled for the online handwritten Tamil and Telugu data. On Telugu data we obtained an accuracy of 90.6% with a speed of 0.166 symbols/sec. To increase the speed we have proposed a 2-stage recognition scheme using which we obtained accuracy of 89.77% but with a speed of 3.977 symbols/sec.*

## 1. Introduction

Online handwriting recognition is the recognition of handwriting that is obtained from the movements of the pen or a stylus when the writer writes on a pen-based interface. For small devices like PDAs, Digitizing Tablets etc. the speed of recognition is an important factor. The online handwriting technology has a lot of potential in India that has more than eighteen official languages. Tamil and Telugu are two major Indic scripts that have their origin in South India. While a lot of research is being done on online recognition of Tamil scripts, much research has not yet been done on online Telugu handwriting recognition. Both the scripts are complex from the point of recognition as several symbols are structurally alike with minor differences. Hence, the recognition of these scripts using either template or elastic matching methods requires efficient features that are more discriminative at the local level.

As discussed by Vuurpijl *et.al.* in [1], Dynamic Time Warping (DTW) is an elastic matching approach that has been observed to give better results than the WANDA system that uses HCLUS prototype matching techniques. In [2], the authors have reported writer dependent recognition of Tamil characters involving multiple stages that use DTW with x-y coordinates, quantized slopes and dominant points as features. Belongie *et. al*. [3] have used shape context feature with bipartite matching for finding the correspondence between boundary points of two shapes as part of their proposed method for object recognition. Shape signatures are functions that map the boundary points of a shape to a one-dimensional space. As described in [4] by Zhang & Lu, distance of the boundary point to the centroid of the shape, the tangent angle, curvature etc are examples of signatures. In [5] Tonouchi *et.al*. have proposed an algorithm in which they have used stroke length for matching the symbols stroke wise. Since the method was proposed for Japanese-Kanji characters, where the characters are made up of strokes that are very near to each other and are not as curvy as Indic scripts, stroke length feature was enough to match strokes between two different characters.

In this paper we present the classification results on online Tamil and Telugu handwritten data. The Tamil data used is "HP Labs Isolated Handwritten Tamil Character Dataset" [6] and the Telugu data set was collected by us. We use DTW distance in combination with the Nearest Neighbor classifier for recognition. Section 2 speaks about our data collection of Telugu symbols. In section 3 we describe the preprocessing of the online data followed by sections 3.1 to 3.7, in which we discuss the various local features that have been used for classification. Section 4 describes the DTW algorithm and the different combinations of features used in the recognition. In Section 5 we present our results on Tamil and Telugu datasets and also discuss a 2 stage recognition method. In Section 6 we give concluding remarks about our work.

## 2. Data Collection

### 2.1. Telugu Symbol Set

Telugu is one of the major languages in India. The script contains syllables that are composed of vowels, consonants and their combinations. In a consonant–vowel combination, the vowels are orthographically indicated by signs called "*matras*". The consonants, vowels, matras and consonant modifiers are included in the symbol set. Also included are some consonant-vowel combinations which cannot be segmented. A symbol set containing 141 symbols, sufficient to cover the entire telugu script has been chosen for data collection. All the complete symbols that are a subset of the entire symbol set are shown in Figure.1.

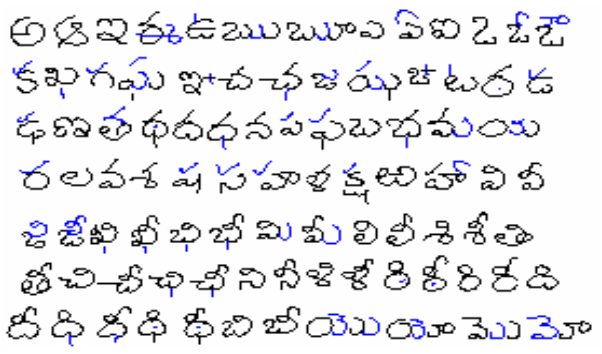

**Figure 1: A subset of the collected Telugu character set**

### 2.2. Data Collection Tool

We have used the Acecad Digital Notepad [7] for the purpose of collecting online handwritten Telugu symbols. Using a pen – paper based device such as the digital notepad, has made data collection more natural for the writers to give their input rather than using digitizers such as Tablet-PCs, where writers are expected to write on different surfaces that make the handwriting unnatural. The writers were provided with forms containing boxes printed on A5 sheets and were asked to write telugu symbols, one in each box, using the digital pen and pad. The ink from each of the boxes is extracted and stored in the UNIPEN format using a data collection tool that we developed for collecting data using the Digital Pad. Telugu data covering 141 unique symbols has been collected from 143 writers with 2 trials per symbol by each writer.

## 3. Local Features

Local features are the features extracted at each point of the stroke. During preprocessing, the characters are uniformly re-sampled to 60 points, followed by normalization of both $x$ and $y$ coordinates to a range of [0-10]. Below, we define 8 different local features that have been used.

### 3.1. Preprocessed x-y Feature

The raw $x$ and $y$ coordinates are preprocessed as mentioned above, and the preprocessed $x_i$ and $y_i$ coordinates are used as the features at the point $p_i$.

### 3.2. Shape Context (SC) Feature

The Shape Context feature has been proposed in [3] as a feature that is calculated at each point of the stroke. At every point the *log* of the distance of the point from rest of the points ($r$) and the slope angle of the line joining the point to each of the points ($\theta$) are calculated. To make the feature scale invariant, the distances $r$ are normalized by the mean distance. Finally at every point $p_i$ a histogram $\hat{h}_i$, built by dividing the log-polar space to bins with respect to $r$ and $\theta$, is given by

$$\hat{h}_i^k = \#\{ q \neq p_i \ : k = r_q * \theta_q \} \qquad (1)$$

where q is a point in the stroke, $k$ is the bin number, $r_q$ and $\theta_q$ are the $r$ and the $\theta$ bins of $q$ respectively. The number of bins to be chosen is determined empirically. SC is a semi local feature which describes the relative position of the entire character with respect to the point of consideration. Since shape contexts are histograms, $\chi^2$ distance is used for finding the distance between two points using shape context feature. Figure.2 gives a pictorial view of the shape context feature with 40 bins at a point.

### 3.3. Tangent Angle (TA) Feature

At the point $p_i$ the slope of the tangent, $\theta_i$, is calculated. The dissimilarity measure of the local tangent angles at two points $p_i$ and $p_j$ is calculated as

$$0.5( 1-\cos ( \theta_i - \theta_j ) )$$

where $\theta_i$ *and* $\theta_j$ are the tangent angles at $p_i$ and $p_j$ respectively.

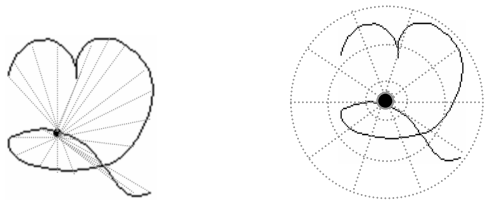

**Figure 2: (Left) Distances from a point (highlighted) to rest of the points. (Right) Shape Context feature with 4 *r*-bins and 10 $\theta$-bins at the point considered.**

### 3.4. Generalized Shape Context (GSC) Feature

The Generalized Shape Context (GSC) feature has been proposed in [8] as an extension to the SC feature discussed above. In addition to the shape context feature, at every point $p_i$, the unit length tangent that gives the direction of the edge at the point is calculated. The x and y components of this vector can be calculated by finding $\cos(\theta_i)$ and $\sin(\theta_i)$, where $\theta_i$ is the tangent angle at $p_i$. The tangent vectors of all the points falling in a bin are added to the bin. The $d$-bin histogram is then converted to a $2d$-dimensional vector $v_i$,

$$v_i = \langle \hat{h}_i^{1,x}, \hat{h}_i^{1,y}, \hat{h}_i^{2,x}, \hat{h}_i^{2,y}, \ldots, \hat{h}_i^{d,x}, \hat{h}_i^{d,y} \rangle \quad (2)$$

where $\hat{h}_i^{k,x}$ and $\hat{h}_i^{k,y}$ are the $x$ and $y$ components of $\hat{h}_i^{k}$. Shape Context is a special case of GSC where the tangent angles are zero. $L^2$ norm has been used as a distance measure for GSC feature.

### 3.5. Normalized Derivative Features

Using the method given in [9] the normalized first derivatives with respect to $x$ and $y$ are calculated at each point $p_i$ as given below.

$$x_i' = \frac{\sum_{t=1}^{r} t.(x_{i+t} - x_{i-t})}{2.\sum_{t=1}^{r} t^2} \qquad y_i' = \frac{\sum_{t=1}^{r} t.(y_{i+t} - y_{i-t})}{2.\sum_{t=1}^{r} t^2} \quad (3)$$

where $r$ defines the number of neighboring points involved in the computation. We have taken $r$ as 2. The normalized derivatives are given by

$$\hat{x}_i' = \frac{x_i'}{\sqrt{x_i'^2 + y_i'^2}} \quad , \quad \hat{y}_i' = \frac{y_i'}{\sqrt{x_i'^2 + y_i'^2}} \quad (4)$$

Similarly, the normalized second derivatives are calculated by substituting $x_i$ and $y_i$ with $\hat{x}_i'$ and $\hat{y}_i'$ in the above equations. We have used both normalized first and second derivatives for classification.

### 3.6. Curvature Feature

For a plane curve, the curvature at a point $p_i$ has the magnitude equal to the reciprocal of the radius of an osculating circle (a circle that closely touches the curve at the given point). The curvature $\kappa_i$ at point $p_i$ is calculated as

$$\kappa_i = \frac{x'.y'' - x''.y'}{\left(x'^2 + y'^2\right)^{3/2}} \quad (5)$$

where $x'$, $y'$ and $x''$ and $y''$ are the normalized first and second derivatives.

## 4. Elastic Matching with Local Features

### 4.1. Dynamic Time Warping

For matching online handwritten symbols the famous elastic matching algorithm, DTW has been used. It is a technique that finds optimal alignment between two time series if one time series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between 2 time series can then be used to find the similarity between them. Let $X$ and $Y$ be two time series of lengths $|X|$ and $|Y|$ given by

$$X = x_1, x_2, \ldots\ldots, x_i, \ldots., x_{|X|}$$
$$Y = y_1, y_2, \ldots\ldots, y_i, \ldots.. y_{|Y|}.$$

A two dimensional cost matrix $D$ of size $|X|$ by $|Y|$ is constructed where the value at $D(i,j)$ is given by

$$D(i,j) = min \begin{cases} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{cases} + d(x_i, y_i). \quad (6)$$

where the particular choice of the recurrence relation varies with the application and the distance function, $d$, depends on the features used. The DTW warp path is calculated starting at $D(|X|,|Y|)$ and by backtracking the minimum cost index pairs. We have used the Sakoe-Chiba band constraint, with 40 as the width of the band, to make DTW faster. For the purpose of classification, DTW is used with varied sets of features that are given below.

### 4.2. Different Sets of Features

We have used 4 sets of features for experimentation. The first set consists of the x-y features alone. The next set consists of the SC feature combined with the TA feature. The third set contains

the GSC feature alone and the final set contains 7 features *viz.* the preprocessed *x-y* features, the normalized first and second derivatives and the curvature. From now on we will call the set of 7 local features as the L7 features where L stands for Local. Euclidean distance is used as the distance measure for the L7 feature set.

## 5. Results and Discussion

### 5.1. Tamil Data Set

The Online handwritten Tamil Data "*hpl-tamil-iwfhr06-train-online* and *"hpl-tamil-iwfhr06-test-online*" [7], that covers 156 Tamil symbols, has been used for recognition. We call it the Competition Data set. The training and testing sets contain 50683 and 26926 samples respectively. We have used the training samples alone for our experimentation, out of which, 40586 samples have been used for training and 4680 for testing. We call this set as the Validation set. We have used the Nearest Neighborhood (NN) classifier that uses the DTW distance found between the test patterns and templates using different feature sets.

### 5.2. Selection of Optimal Number of Bins

We first consider SC and TA features for finding the optimal number of bins to be used. A combination of their respective distance measures, with weights 0.9 and 0.1 respectively, is used as the cost measure. The same parameters have been used by Belongie.et.al[3] for the recognition of the MNIST data set of hand-written digits for the calculation of the cost matrix.

**Table 1: Optimal Number of Bins for SC Feature**

| Bins : $r$<br>$\theta$ | 3<br>6 | 3<br>8 | 3<br>12 | 4<br>8 | 4<br>10 | 5<br>12 |
|---|---|---|---|---|---|---|
| Accuracy | 78.93 | 79 | 79.33 | 79 | 80.27 | 78.26 |

For this purpose, 10 prototypes have been generated from the validation set and have been tested on the first 10 symbols of the validation set. 40 bins have given optimal results as shown in Table1.

### 5.3. Classification Results on Tamil Data

In the case of GSC features also we have used 40 bins. The DTW distance has been found using each of the x-y, SC+TA, GSC and L7 features. The results are reported on all the 156 symbols from the validation set. 50 prototypes have been used for training. Table 2 shows the classification results using the 4 feature sets. The x-y features stand last in the recognition rate where as the L7 features have given best accuracy. From the point of recognition speed SC+TA are the slowest, while GSC are better than SC+TA in terms of error rate and speed. x-y and L7 features are faster than the former 2 sets.

**Table 2: Recognition results on the Validation set with 50 prototypes with each of the feature sets.**

| Feature Set | Accuracies ( % ) | | Time (secs/symbol) |
|---|---|---|---|
| | Top1 | Top2 | |
| X-Y | 77.65 | 91.13 | 0.5 |
| SC + TA | 79.08 | 92.98 | 240 |
| GSC | 79.37 | 91 | 6 |
| L7 | 82.37 | 93.65 | 1 |

The selection of prototypes also affects the recognition rate. Better results were obtained when tested with the prototypes selected using the same feature as those used for testing instead of the x-y features. In Table 2 we observe that the accuracy of L7 feature set using normal prototyping method is 82.37%. However, when trained with prototypes generated using the L7 features, the accuracy increased to 83.45%.

Finally, using 50 prototypes, DTW with L7 features has given an accuracy of 82.34% on the Competition Data set.

### 5.4 Classification Results on Telugu Data Set

As mentioned in Section 2.2, out of the 38389 (after cleaning the totally collected 40326 samples) samples of Telugu symbols that have been collected, 29174 have been used for training and 9215 for testing. Using the L7 features on this data set, with 50 prototypes, we have achieved an accuracy of 87.22%. We have also tested the telugu data set using DTW with both x-y and L7 features without prototyping. Figure.3 shows the top 5 accuracies using x-y and L7 features without prototypes. 90.6 % is the top1 accuracy using L7 feature and it has taken 6.024 seconds for recognizing one symbol.

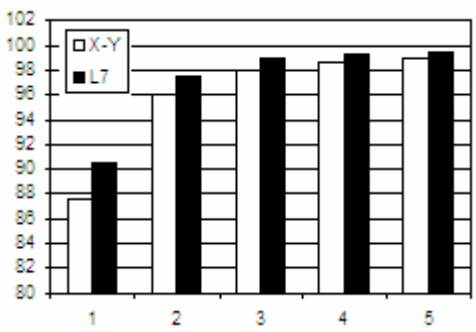


**Figure 3:Top 5 accuracies of DTW without prototypes using x-y and L7 features on Telugu data.**

We also have considered recognition in two stages, where in the first stage we have selected only the best 100 samples from the entire training set using NN classifier with Euclidean distance and in the second stage we have used NN classifier with DTW distance. In the second stage, only the 100 samples selected in the first stage have been used for matching. We have implemented the 2-stage recognition using x-y features and L7 features respectively. The accuracies over the entire 141 symbols are given in Table 3.

**Table 3: Results of Two-Stage recognition on Telugu Data Set.**

| Feature Set | Accuracy (%) | Time (secs/symbol) |
|---|---|---|
| X-Y | 87.06 | 0.1778 |
| L7 | 89.77 | 0.2514 |

## 6. Conclusions and Discussion

We have compared the results of DTW with four different sets of features. We have used the Tamil data for experimentation. The SC and TA features are found to be slow since we have used $\chi^2$ distance as a cost measure. GSC feature is better than SC+TA in terms of accuracy and is 35 times faster than the latter. The L7 were the best features in terms of speed and accuracy. On Telugu data we obtained an accuracy of 90.6% using DTW and the L7 features. We have used two-stage recognition for Telugu symbols using NN classifier with Euclidean distance in the first stage and the NN classifier with DTW distance in the second stage. It has given an accuracy of 89.77% and a speed of 4 symbols/sec using L7 features. In the case of Telugu characters, even the L7 features were unable to distinguish several confusing pairs. The pairs ద,డ ధ,ధ స,స ష,ష ఏ,ష and have been misclassified since we have not considered the pen up information. Few more examples of the misclassified pairs are ఛ,భ ము,యు ర,ద that are structurally almost alike. This misclassification can also be avoided by using discriminative features, calculated at each point, that take into account the structure of the character around the point also.

## 10. References

[1] R.Niels and L.Vuurpijl, "Using Dynamic Time Warping for Intuitive Handwriting Recognition", *Proc. of 12th Conference of the International Graphonomics Society*, Salerno, 2005, pp. 217-221.

[2] Niranjan Joshi, G.Sita, A.G.Ramakrishnan, and M.Sriganesh , "Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition", *Proc. of IWFHR-9*, Tokyo, 2004, pp. 444-449.

[3] S.Belongie, J.Malik, and J.Puzicha, "Shape Matching and Object Recognition Using Shape Contexts", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 24, No. 4, April 2002, pp. 509-522.

[4] D.Zhang and Goujun Lu, "Review of shape representation and description techniques", *Pattern Recognition* , Vol. 37, No. 1, January 2004, pp. 1-19.

[5] Y.Tonouchi and A.Kawamura, "An On-Line Character Recognition Method Using Length-Based Stroke Correspondence Algorithm", *Proc.of ICDAR'97*, Germany, 1997, pp. 633-636.

[6] HP Labs Isolated Handwritten Tamil Character Dataset http://www.hpl.hp.com/india/research/penhw-interfaces-1linguistics.html

[7]Acecad Digimemo Model A502 *http://www.acecad.com.tw/dma502.htm*.

[8] G.Mori, S.Belongie, and J.Malik, "Efficient Shape Matching Using Shape Contexts", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 27, No.11, November 2005, pp. 1832-1837.

[9] M.Pastor, A. Toselli, and E.Vidal, "Writing Speed Normalization for On-Line Handwritten Text Recognition", *Proc. of ICDAR'05,* Seoul, 2005, pp. 1131-1136.