Handwritten Character Recognition Using PCA of State Space Point Distribution

Lajish VL
TCS Innovation Lab - Mumbai,
TCS Yantra Park, Thane, Mumbai, India
Lajish.VL@TCS.Com

Sita G
Department of Electrical Engineering
Indian Institute of Science, Bangalore
Sita.G@TCS.Com

Sunil Kumar Kopparapu
TCS Innovation Lab - Mumbai,
TCS Yantra Park, Thane, Mumbai, India
SunilKumar.Kopparapu@TCS.Com

***Abstract*—In this paper, we investigate the utility of a feature set derived using the principles of non-linear dynamics applied to gray scale images for handwritten character recognition. The feature set is derived using the principles of non-linear dynamics from the state-space-map (SSM) generated from gray-image. The State-Space Point Distribution (SSPD) parameters are then extracted from the SSM. We use principal component analysis (PCA) on the SSPD features in order to decorrelate the features and for dimension reduction. We studied the recognition results using Nearest Neighbor (NN) classifier, with two different distance measures, namely Euclidean and Mahalanobis distances, and a Bayesian classifier. Experimental results obtained using different number of principal components, demonstrate that novel feature set holds promise and is effective for handwritten character recognition.**

I. Introduction

Handwritten Character Recognition (HCR) system can improve human computer interaction and has been an active research area. Promising results are reported in the area of Handwritten Character Recognition (HCR) for languages like English, Chinese, Korean, Japanese, Arabic; and for Devanagari, Bangla, Tamil, Telugu, Kannada in Indian languages. HCR research in Malayalam is still in its infancy and this work experiments with Malayalam character recognition, though the framework is general and can be used for any language. We describe a novel approach for feature extraction from gray-scale images of the handwritten characters and then use it to recognize isolated handwritten characters.

A handwritten character image, denoted by a two-dimensional function $f(x, y)$, is treated as a non-linear dynamical process with the original scalar measurements including the pixel intensity value f and the spatial co-ordinates (x, y). The method of feature extraction exploits the topographic structure in a gray-scale image. A state-space activity map based on the pixel intensity distribution of both the foreground and the background of the gray-scale image is constructed. Parameters obtained from the state-space point distribution (SSPD) are used as the feature set to represent the character. We investigate the performance of these features on Malayalam handwritten character recognition using Nearest Neighbor (NN) with Euclidean and Mahalanobis distances and a Bayes classifier. Principal Component Analysis (PCA) is used to reduce the dimensionality of the SSPD feature set. This reduced dimensional PCA features are used in all classifiers. The paper is organized as follows. Section II describes the theory behind reconstructed state-space and its application to handwritten character images. Section III explains the feature extraction and the classifiers used. Section IV presents the experimental results obtained with various experiments along with a discussion of the results. In the last section (Section V) we present the conclusions.

II. The Reconstructed State-Space for Handwritten Character Images

A number of methods exist in literature for offline handwritten character recognition which are based on gray-scale image based features [1], [2], [3]. We look at this problem from a non-linear dynamical system point of view and investigate the performance of the features obtained by application of principles of non-linear dynamics to gray-scale images for recognition purposes.

In the case of purely deterministic systems, once it's state is fixed, then the states at all future times can be determined as well. Thus by all accounts it is significant to establish a vector space called State-Space for the system such that, specifying a point in this space specifies the state of the system. This helps us study the dynamics of the system by studying the dynamics of the corresponding state-space points. The concept of the state of a system is powerful for non-deterministic systems also. Takens theorem [4] states that under certain assumptions, state-space of a dynamical system can be reconstructed through the use of time delayed (space varying) versions of the original scalar measurements. This new state-space is commonly referred in the literature as a reconstructed state-space. A reconstructed state-space can be treated as a powerful signal-processing domain, specially when the dynamical system of interest is non-linear or even chaotic [5], [6].

A reconstructed state-space for a dynamical system can be produced from a measured state variable, I_n where $n = 1, 2, \cdots, N$, via the method of delays by creating vectors given by

$$I_n = [i_n, i_{n+\tau}, i_{n+2\tau}, \cdots, i_{n+(d-1)\tau}]$$

where d is the embedding dimension and τ is the chosen time or space delay value. The row vector I_n defines the position of a single point in the reconstructed state-space. The row vectors

then can be compiled into a matrix (called a trajectory matrix) to completely define the dynamics of the system and create a reconstructed state-space.

$$I = \begin{bmatrix} i_1 & i_{1+\tau} & i_{1+2\tau} & \cdots & i_{1+(d-1)\tau} \\ i_2 & i_{2+\tau} & i_{2+2\tau} & \cdots & i_{2+(d-1)\tau} \\ i_3 & i_{3+\tau} & i_{3+2\tau} & \cdots & i_{3+(d-1)\tau} \\ \vdots & \vdots & \vdots & \vdots & \\ i_N & i_{N+\tau} & i_{N+2\tau} & \cdots & i_{N+(d-1)\tau} \end{bmatrix}$$

A handwritten character image, denoted by a two-dimensional function $f(x, y)$, can be treated as a dynamical system with the original scalar measurements including the pixel intensity values or amplitude of f and the spatial co-ordinates (x, y). Based on the above theory, a method to model a reconstructed state-space for handwritten character images, through the use of space varying versions of the original scalar measurements has been proposed [7]. The trajectory matrices I_1 with embedding dimension $d = 2$ and space delay or space variation $\tau = 1$ and I_2 with embedding dimension $d = 3$ and space delay $\tau = 1$ are constructed by considering the pixel intensity values i taken from all the spatial co-ordinate points of the image. The matrices I_1 and I_2 thus obtained are given below.

$$I_1 = \begin{bmatrix} i_1 & i_{1+\tau} \\ i_2 & i_{2+\tau} \\ i_3 & i_{3+\tau} \\ \vdots & \vdots \\ i_N & i_{N+\tau} \end{bmatrix} \quad I_2 = \begin{bmatrix} i_1 & i_{1+\tau} & i_{1+2\tau} \\ i_2 & i_{2+\tau} & i_{2+2\tau} \\ i_3 & i_{3+\tau} & i_{3+2\tau} \\ \vdots & \vdots & \vdots \\ i_N & i_{N+\tau} & i_{N+2\tau} \end{bmatrix}$$

A new method of feature extraction based on the state-space representation by trying all possible locations in a gray-scale image was introduced in [7]. A trajectory matrix of embedding dimension $d = 2$ is first formed from this reconstructed state-space of each character image (Figure 1) and then the scatter plot of the row vector of the trajectory matrix, named Sate-Space Map (SSM, Figure 2), is constructed for each character image with one directional space delays. The State-Space Point Distribution (SSPD, Figure 3) parameters are then extracted for each character pattern from the SSM. The SSPD feature for a given gray-scale image is extracted as described in Algorithm 1. For more details refer to [8]. The SSPD features have an additional advantage in terms of being invariant to *small* rotations, mirror flip and translation as seen from I_1 and I_2.

III. Principal Component Analysis based Classification using SSPD features

In this paper, we investigate the performance of the state-space point distribution (SSPD) features (see Algorithm 1) for handwritten character recognition. The SSPD features are initially decorrelated by performing eigen transformation on it using principal component analysis (PCA). The decorrelate feature set is used for building classifier and for recognition. In this paper, we use three classifiers, namely, the NN classifier with Euclidean distance (NN-ED), NN classifier with Mahalanobis distance (NN-MD) and a probabilistic Bayes classifier.

Fig. 1. Character ah

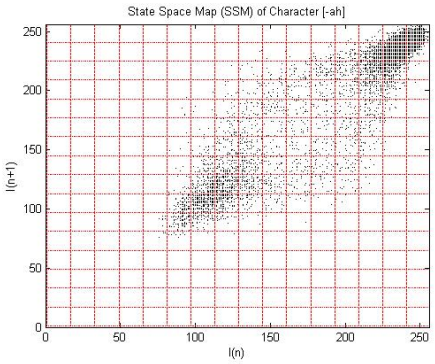

Fig. 2. State-space map for ah

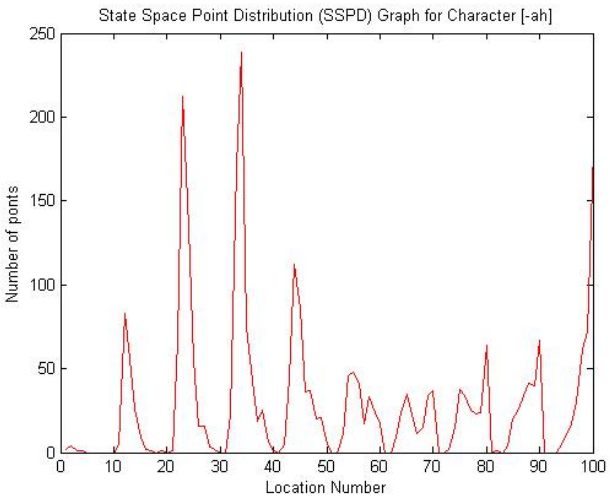

Fig. 3. SSPD feature for ah

A. Eigen transformation of SSPD features using PCA

Assume that there are q character classes denoted by $C_1, C_2, \cdots, C_q$ in a training data-set $\mathcal{S} = [\vec{s}_1; \vec{s}_2; \cdots; \vec{s}_M]$. The data is represented in the form of a matrix $\mathcal{S}$ of dimension $M \times N$ where $\vec{s}_i; 1 \leq i \leq M$ represent the SSPD feature map of the sample i and is represented as a row vector of length N ($N = B * B$ in Algorithm 1). Let $\mathcal{E}_\mathcal{S} = [\vec{e}_1; \vec{e}_2; \cdots; \vec{e}_N]$ be the eigen-vector matrix for the data-set $\mathcal{S}$ computed from the covariance matrix $\mathcal{C}$ as,

$$\mathcal{C} = \sum_{i=1}^{M} (\vec{s}_i - \vec{\mu})(\vec{s}_i - \vec{\mu})^T \tag{1}$$

where $\vec{\mu}$ is the mean SSPD feature for the training data-set $\mathcal{S}$. Any sample $\vec{s}_i \in \mathcal{S}$ may be expressed as a linear combination of the eigen vectors, $\vec{e}_i \in \mathcal{E}_\mathcal{S}$ (also called as principal components). The eigen transformed features, $\vec{z}_{\vec{s}_i}$ for any sample $\vec{s}_i \in \mathcal{S}$ is computed by projecting the sample on to the principal components of $\mathcal{S}$ as

$$\vec{z}_{\vec{s}_i}(n) = \langle \vec{s}_i, \mathcal{E}_\mathcal{S}(n) \rangle \text{ where } n = 1, \cdots, N \tag{2}$$

In other words, any sample in the set $\mathcal{S}$ may be represented as,

Algorithm 1 Pseudo-code to extract SSPD features.

```
Let G be the maximum gray value of an image
Let B = 16; used for controlling the size of SSPD feature vector
im = read image {read the image}
[im_Height im_Width] = size(im);
{calculate the size of the image}
im_size = im_Height * im_Width; {total number of pixels}
for (i = 1; i < im_Height; i++) do
  for (j = 1; j < im_Width; j++) do
    x = im(i, j);
    if (j < im_Width) then
      y = im(i,j+1);
    else if (i < im_Height) then
      y = im(i+1, 1);
    else if ((i == im_Height) & (j == im_Width)) then
      y = im(1, 1);
    end if
    A(x,y) = A(x,y) + 1;
  end for
end for
K=0; L=0
for (k = 1; k < G; k=k+B) do
  K++
  for (l = 1; l < G; l=l+B) do
    L++
    for (m = k; m < k+B; m++) do
      for (n = l; n< l+B; n++) do
        V(K,L) = V(K,L) + A(m,n)
      end for
    end for
  end for
end for
t1 = reshape(V,1,B*B);
t = quantize;
SSPD = t / im_size; {to take care of size normalization}
```

$$\vec{s}_i = \sum_{j=1}^{N} \vec{z}_{\vec{s}_i}(j)\mathcal{E}_{\mathcal{S}}(j) \tag{3}$$

The projection lengths of samples in $\mathcal{S}$ represented in Z are the PCA features for the set in $\mathcal{S}$. By choosing the strong principal components corresponding to larger eigen values in (3), say $K < N$, we can use the reduced number of eigen components to approximately represent the samples in $\mathcal{S}$ thereby leading to dimensionality reduction of the SSPD feature set. One of the greatest advantage of this linear transform is that the features in this domain are highly uncorrelated. We have used the PCA features obtained from the SSPD of the handwritten gray-level images in our experiments for character recognition.

B. Nearest Neighbor (NN) Classification

The centroid of each character set $C_i, i = 1, \cdots, q$ is calculated by projecting all the samples in the training set belonging to the character set C_i onto the K principal components ($\vec{z}_{\vec{s}_i}$) and finding the mean as

$$\vec{\mu}_{C_i} = \frac{1}{\text{No of samples in } C_i} \sum_{\vec{s}_i \in C_i} \vec{z}_{\vec{s}_i}$$

Nearest Neighbor (NN) classification from the centroids of the PCA features of each character class to the test vector ($\vec{t}$), is performed using either Euclidean distance (4)

$$d_{C_i}^{ED} = \|\vec{z}_{\vec{t}} - \vec{\mu}_{C_i}\| \tag{4}$$

or Mahalanobis distance (5).

$$d_{C_i}^{MD} = \frac{\|\vec{z}_{\vec{t}} - \vec{\mu}_{C_i}\|}{\Sigma_{C_i}} \tag{5}$$

In (4) and (5) $\vec{z}_{\vec{t}}$ is the PCA feature of the test vector in the reduced PCA space and Σ_{C_i} in (5) is the variance and is calculated as

$$\Sigma_{C_i} = \frac{1}{\text{no of samples in } C_i} \sum_{\vec{s}_i \in C_i} (\vec{z}_{\vec{s}_i} - \vec{\mu}_{C_i})(\vec{z}_{\vec{s}_i} - \vec{\mu}_{C_i})^T$$

The test sample ($\vec{t}$) is assigned the class label of the most similar centroid (minimum $d_{C_i}^{ED}$ for $i = 1, \cdots, q$ for Euclidean distance and minimum $d_{C_i}^{MD}$ for $i = 1, \cdots, q$ for Mahalanobis distance).

C. Bayesian Classification

We use Bayes decision theory for handwritten character recognition using the PCA features of SSPD of the handwritten characters in $\mathcal{S}$. Bayes' classifier assigns a test vector, $\vec{z}_{\vec{t}} \in Z$ to a class, C_i if

$$P(C_i|\vec{z}_{\vec{t}}) > P(C_j|\vec{z}_{\vec{t}}) \text{ for all } j \neq i$$

where the posterior probability, $P(C_i|\vec{z}_{\vec{t}})$ is defined as,

$$P(C_i|\vec{z}_{\vec{t}}) = \frac{p(\vec{z}_{\vec{t}}|C_i)P(C_i)}{\sum_{j=1}^{q} p(\vec{z}_{\vec{t}}|C_j)P(C_j)}$$

The Bayes classifier essentially depends on the class conditional densities, $p(z|C_i)$ which is computed from the training set. By assuming Gaussian density for the PCA feature set of each character class considered and equally probable classes, the discriminant function reduces to,

$$p(\vec{z}_{\vec{t}}|C_i) > p(\vec{z}_{\vec{t}}|C_j) \text{ for all } j \neq i$$

in other words, assign the test vector, $\vec{z}_{\vec{t}}$ to class C_i where $\arg\max_{C_i} p(\vec{z}_{\vec{t}}|C_i)$. In the next section we give details about the data and the results obtained using the three classifiers.

IV. Experimental Results

We took a database of $q = 5$ classes, namely, Malayalam alphabets ah, yi, eh, oh, ka; a sample set is shown in Figure 4. Each class had 50 samples of varying sizes and written by different people on a white piece of paper. The written data was scanned and stored as an image (BMP format). For all the experiments, we used a set of 30 samples randomly picked from each character set, to train the classifier and the remaining 20 samples to test the classifier. The dimension of

Character	Samples				
	1	2	3	4	5
അ [ah]					
ഇ [yi]					
എ [eh]					
ഒ [oh]					
ക [ka]					

Fig. 4. Typical character set.

Classifier	ah	yi	eh	oh	ka	Acc
NN-ED	36	47	37	30	34	**73.6 %**
NN-MD	46	50	50	50	50	**98.4 %**
Bayes	49	50	50	41	46	**94.4 %**

TABLE I
RECOGNITION WITH TRAIN DATA (ALL 50 SAMPLES) WITH PCA SIZE 11.

the SSPD feature of all characters is represented as a 256 length vector ($B = 16$ in Algorithm 1). We obtain the PCA features from this data matrix as described in Section III-A. It has been observed from the eigen values that the data has large amount of redundancy and the principal components beyond the first 11 eigen directions carry very little information. Hence in our analysis we considered only upto first 11 principal components, this reducing the dimension of the SSPD reduced from 256 to 11.

The performance of the three classifiers on the training set data is checked and tabulated in Table I. A set of 50 samples are used for training and the same 50 samples are used for testing the classifier in this case. One can observe that the classifiers are able to *learn* the training data set sufficiently well in all the three cases. While NN-MD and the Bayes classifiers are able to perform with an accuracy (high 90%+); NN-ED performs pretty poorly at 73.6%. Note that both NN-MD and Bayes use second order statistics in the form of covariance matrix which results in an enhanced performance compared to the NN-ED classifier. Table II shows the average recognition accuracies obtained across the 5 classes for various PCA sizes on the 20 test data. Table III shows the performance of the three classifiers on the 20 test samples each for 5 classes

PCA size	NN-ED	NN-MD	Bayes
5	65%	55%	55%
8	65%	65%	60%
11	70%	90%	70%

TABLE II
COMPARISON OF OVERALL RECOGNITION ACCURACIES WITH VARYING PCA SIZES.

Classifier	ah	yi	eh	oh	ka
NN-ED	65	75	50	40	70
NN-MD	60	100	60	45	90
Bayes	40	75	20	45	90

TABLE III
RECOGNITION (IN %) WITH TEST DATA WITH PCA SIZE 11

for a PCA size of 11. Again, it can be observed that the overall performance of the NN-MD classifier is superior to both the NN-ED classifier and the Bayes classifier [9]. Experiments were also conducted on mirror flip images. In this case the training set was not mirror flipped but the test images were mirror flipped. The mirror flipped character also produced recognition accuracies similar to the results shown in Table III.

V. CONCLUSIONS

A novel feature set has been used to model handwritten characters for recognition which is image size, mirror flip and small image rotation invariant. The features are extracted easily since it does not require any preprocessing (example, binarization, size normalization). Added advantage is that the feature extraction procedure is computationally less complex (it takes about 0.265 seconds to compute the SSPD feature from an image of size 120×100 in a Matlab environment on a laptop) and the dimension of the SSPD feature vector of length 256 after PCA is comparatively small (about 11). In effect the SSPD parameters derived from the gray-scale based SSMs of handwritten character samples can be effectively utilized for high speed HCR applications.

ACKNOWLEDGMENTS

The authors would like to thank the members of the TCS Innovation Lab - Mumbai for their support and encouragement.

REFERENCES

[1] L. Wang and T. Pavlidis, "Direct gray-scale extraction for character recognition," *IEEE Trans. PAMI*, vol. 15, no. 10, pp. 1053–1067, 1993.
[2] T. Peuker and D. Donglas, "Detection of surface-specific points by local parallel processing of discrete terrain elevation data," *Journal of Computer Graphics and Image Processing*, vol. 4, no. 4, pp. 375–387, 1975.
[3] T. P. S. Kahan and H. Baird, "Building a font and size independent character recognition system," *IEEE Trans. on PAMI*, vol. 9, pp. 274–288, 1987.
[4] F. Takens, "Detecting strange attractors in turbulence," in *Lecture notes in Mathematics Volume 898*, R. D.A. and Y. L.S., Eds. Berlin: Springer, 1981.
[5] E. Ott, *Chaos in dynamical systems*. Cambridge University Press, 1993.
[6] G. Baker and J. Gollub, *Chaotic dynamics - An Introduction*. Cambridge University Press, 1996.
[7] V. L. Lajish and N. K. Narayanan, "Handwritten character recognition using gray-scale based state-space parameters and nn classifier," in *ICSIP-2006*, vol. 1, Hubli, India, 2006, pp. 196–201.
[8] V. L. Lajish, "Adaptive neuro-fuzzy inference based pattern recognition studies on handwritten character images," Ph.D. dissertation, University of Calicut, Kerla, India, 2007.
[9] R.O.Duda and P.E.Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973.