



Classification of Handwritten Ancient Tamil characters using Complex Extreme Learning Machine

N.Sridevi¹, Dr.P.Subashini²

¹Research Scholar, ²Professor

Department of Computer Science

Avinashilingam University for Women, Coimbatore, Tamil Nadu, India.

Abstract: Classification is the problem of identifying, to which set of categories a new observation belongs, on the basis of a training set whose category membership is known. The process of handwritten script classification involves extraction of some defined characteristics called features to classify an unknown handwritten character into one of the known classes. Zernike moments and regional features are extracted from the Tamil characters and they are formed as feature vectors. Complex Extreme Learning Machine is used to classify the handwritten ancient Tamil characters. Complex Extreme Learning Machine is trained with feature vectors. From the experimental result it is observed that the classifier when trained by combining Zernike moments with regional features gives a highest classification accuracy of 82.63%.

Keywords: Zernike moments, Regional features, Complex Extreme Learning Machine, Classification, Handwritten Characters.

I. Introduction

Tamil is one of the oldest languages known in India dating back over to two millennia. Inscriptions were written in olden days which are rich in knowledge related to different fields like literature, astrology, medicine, history and so on. Getting information from these inscriptions is essential. To do this, the common man must know to read the ancient Tamil scripts, whose writing styles and variations are different. This lays the main reason to carry out this research work in classification of ancient Tamil scripts. Classification is the problem of identifying to which set of categories a particular character belongs on the basis of training set. The process of classification involves extraction of some definable information called features to classify an unknown handwritten character into one of the known classes.

From the literature, it has been observed that the previous researchers have used traditional algorithms like SVM, SOM, HMM and two layer feed forward networks [1– 3]. These algorithms are far slower because the parameters must be tuned iteratively which increase the time taken for classification. Hence, to overcome this drawback of the traditional algorithms, this research work intends to use Complex Extreme Learning Machine (CELM) for classification of handwritten ancient Tamil scripts because in CELM the parameters are generated randomly and this saves the manual tuning of the parameters. The organization of the paper is as follows: In section 2, features extracted from the characters are explained in brief. Classification using CELM is given briefly in section 3. Results and discussion are discussed in section 4 and finally section 5 concludes this paper.

II. Feature Extraction

Feature extraction is defined as the process of extracting information from the raw data which is useful for classifying the unknown type into known class. Features are classified into two groups, they are structural features like strokes, end points, etc., and statistical features which are derived from the statistical distribution of points like zoning, moments, etc., [4]. Here statistical feature (Zernike moments) along with regional features are taken for classification of handwritten ancient Tamil scripts. In feature extraction each character is represented as a feature vector, which becomes its identity [5]. Therefore, in this paper features are extracted and formed as different feature vectors.

A. Zernike Moments

Zernike moments are a class of orthogonal moments. They are rotation invariant. The Zernike polynomials are a set of complex orthogonal polynomials defined over the interior of a unit circle [6]. Zernike moments are the projections of the image function onto these orthogonal basis functions. The Zernike moment of order n with repetition m for an image is given by

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y B(x, y) [v_{nm}(\rho, \theta)]^* \quad (1)$$

where, * is the complex conjugate operator

Table.1 shows the features obtained by using Zernike moments for the sample ancient handwritten Tamil scripts

Table 1: Zernike Moments of order 6

Characters	Moment 0	Moment 1	Moment 2	Moment 3	Moment 4	Moment 5	Moment 6
அ	3.97	2.72	19.65	3.33	17.53	5.49	-0.54
ஈ	3.51	1.10	6.67	0.38	-8.11	-0.40	-12.03
ஊ	3.53	0.65	1.91	5.58	0.81	4.05	3.48
உ	4.60	1.53	86.56	28.39	-26.19	34.52	-43.34
ஊ	3.59	0.02	7.31	1.12	23.14	-0.06	29.79
ஊ	3.73	6.43	2.04	2.88	3.58	4.88	-4.72
ஊ	5.05	7.22	6.68	0.16	5.07	0.43	-3.27
ஊ	5.18	7.68	5.31	1.62	-1.38	3.60	4.34
ஊ	3.00	0.78	1.89	0.24	1.16	0.19	2.87

B. Regional features

Here the features like Centroid, Orientation, Eccentricity, Extent, Mean and Standard deviation of Tamil characters are calculated

- Centroid – Specifies the center of mass of the region.
- Eccentricity – the eccentricity of the ellipse that has the same second-moments as the region.
- Orientation – This is the measure of angle in degrees between the x-axis and the major axis.
- Extent – It is defined as the ratio of pixels in the region to pixels in the total bounding box.
- Mean – Mean value of a region.
- Standard Deviation – It is defined as the measure of the dispersion of a set of data from its mean.

Table 2 shows the regional features extracted from the sample ancient handwritten Tamil scripts

Table 2: Regional Features

Characters	Centroid	Eccentricity	Orientation	Extent	Mean	Standard deviation
அ	828.00	43.89	23.73	0.82	0.22	472.09
ஈ	655.00	40.52	19.35	0.90	0.20	365.23
ஊ	700.00	28.05	23.37	0.47	0.29	409.30
உ	664.00	22.00	23.86	26.34	5.40	0.70
ஊ	1758.00	3.00	1.00	1.00	18.55	8.00
ஊ	471.00	16.49	27.02	0.89	0.29	272.65
ஊ	671.00	19.50	25.86	0.73	0.39	248.55
ஊ	524.00	22.68	18.28	0.59	0.22	293.58
ஊ	346.00	39.00	16.10	20.22	3.64	0.84

III. Classification

Tamil is one of the oldest languages known in India. Tamil has 247 characters in it out of which 12 are vowels, 18 are consonants, and 216 are composite characters (formed by combination of vowels and consonants) and one special letter known as Aydhham. Tamil scripts are classified according to these groups. Target of the classification is to reduce the number of possible characters for an unknown character, from the known one [7]. Here, complex extreme learning machine is used for classification of handwritten ancient Tamil scripts.

A. Complex Extreme Learning Machine

Given a series of training samples (z_i, y_i) , where $i = 1, 2 \dots N$, $z_i \in C^n$ and $y_i \in C^m$, the outputs of the single hidden layer feed forward network with complex activation function for these N training data is given by [8]

$$\sum_{k=1}^N \beta_k g_c(w_k \cdot z_i + b_k) = o_i, \quad i = 1, \dots, N \quad (2)$$

Where, $w_k \in C^n$ is the complex input weight vector connecting the input layer neurons to the hidden neuron, $\beta_k = [\beta_{k1}, \beta_{k2}, \dots, \beta_{km}]^T \in C^m$ is the complex output weight vector connecting the hidden neuron and the output neurons and $b_k \in C$ is the complex bias of the k^{th} hidden neuron. g_c is a fully complex activation function. The above N equations can be written as

$$H\beta = O \quad (3)$$

and the number of hidden neurons is usually less than the number N of training samples.

$$H(w_1, \dots, w_N, z_1, \dots, z_N, b_1, \dots, b_N) \quad (4)$$

$$= \begin{bmatrix} g_c(w_1 \cdot z_1 + b_1) & \dots & g_c(w_N \cdot z_1 + b_N) \\ \vdots & \dots & \vdots \\ g_c(w_1 \cdot z_N + b_1) & \dots & g_c(w_N \cdot z_N + b_N) \end{bmatrix}_{N \times N} \quad (5)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_N^T \end{bmatrix}_{N \times m}, O = \begin{bmatrix} o_1 \\ \vdots \\ o_N \end{bmatrix}_{N \times m} \text{ and } Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times m} \quad (6)$$

Here, the complex matrix H is called the hidden layer output matrix. For fixed input weights and hidden layer biases, least squares solution of the linear system with minimum norm of output weight can be obtained [9]. The resulting least square solution is given by

$$\hat{\beta} = H^\dagger Y \quad (7)$$

where, $H^\dagger$ is the Moore-Penrose generalized inverse of complex matrix H.

B. Activation Function Type

The complex activation functions can be used in complex extreme learning machine. These include circular functions, inverse circular functions, hyperbolic functions and inverse hyperbolic functions [10]. In this research work, “arcsinh” is used as the activation function.

$$\arcsin h(z) = \int_0^z dt / ((1+t^2)^{1/2}), \text{ where } z \in C \quad (8)$$

C. Moore-Penrose Generalized Inverse Matrix

Matrix A is the Moore Penrose generalized inverse of complex matrix B, if $ABA = A$, $BAB = B$, $(AB)^* = AB$, $(BA)^* = BA$

Singular value decomposition (SVD) is used to calculate the Moore-Penrose generalized inverse of H. SVD is a factorization of a real or complex matrix which is used in many signal processing and statistics applications. The singular value decomposition of an $m \times n$ complex matrix M is a factorization of the form

$$M = U \Sigma V^* \quad (9)$$

where, U is a $m \times m$ complex matrix, Σ is an $m \times n$ diagonal matrix with non negative real numbers on the diagonal and V^* is an $n \times n$ complex matrix

D. The general of algorithm for CELM

Given a training set N, complex activation function $g_c(z)$ and hidden neuron number $\tilde{N}$

Step 1: Randomly choose the complex input weight w_k and the complex bias b_k , where $k = 1, \dots, \tilde{N}$.

Step 2: Calculate the complex hidden layer output matrix H

Step 3: Calculate the complex output weight β using eq 7

The following procedure describes how the complex extreme learning machine is used in classification of handwritten ancient Tamil scripts. Training data and testing data are given as input the CELM.

Step 1: From training and testing data sets, the class labels are extracted and it is saved as Target vector.

Step 2: Complex random numbers are generated for the input weight of size (number of Hidden neurons X number of input neurons).

Step 3: Bias of hidden neurons are randomly generated from the complex numbers.

Step 4: Hidden layer output matrix H is calculated using the “arcsinh” activation function for the training data.

Step 5: Moore Penrose inverse matrix $H^\dagger$ is calculated

Step 6: Output weight is calculated using Eq 7.

Step 7: To find the actual output of the training data, the output weight is multiplied with $H^\dagger$.

Step 8: Repeat steps 5 to 8 to calculate the output of testing input.

Step 9: Classification accuracy for training and testing data is calculated.

IV. Results and Discussion

To classify the handwritten ancient Tamil characters, a sample of 500 characters is taken and their features are extracted. In order to train the Complex extreme learning machine, three different feature vectors are formed from the Zernike moments (FV1), Regional features (FV2) and by combining Zernike moments with regional features (FV3). Out of these 500 characters, 300 characters are used for training the complex extreme learning machine and remaining 200 characters are used as testing data. Number of hidden neurons is fixed as 50 neurons. The classification accuracy obtained by complex extreme learning machine using different feature vectors is tabulated.

Table 3: Classification accuracy obtained by using different feature vectors

Feature Vector	Training Time	Testing Accuracy
FV1	0.0313	78.84
FV2	0.0123	80.79
FV3	0.0569	82.63

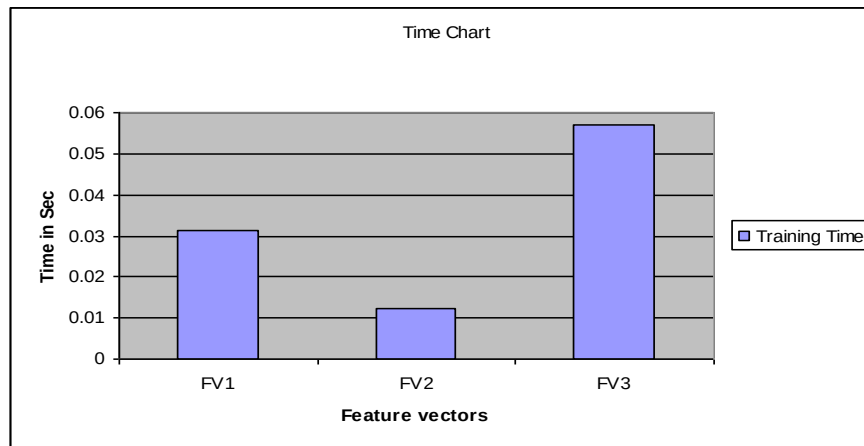


Figure 2: Time taken by CELM when trained using different vectors

Fig 2 shows that training time taken by CELM when trained with different feature vectors. It is observed that, for training the vector FV2 the CELM takes less time when compared to other feature vectors. The training time taken for FV3 is higher than other two vectors because number of attributes in this vector is more (i.e. 12). From the figure 3, it is found that when CELM trained using FV3, gives highest accuracy rate of 82.63 compared to others even though the time taken for training is more.

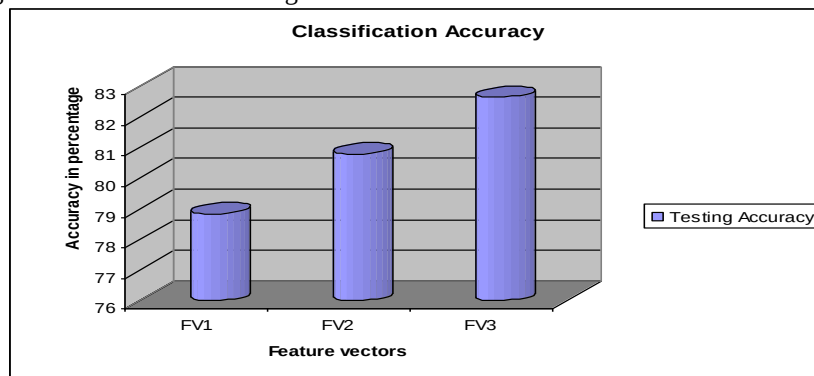


Figure 3: Classification accuracy obtained using various feature vectors

V. Conclusion

Complex extreme learning machine is used for classification of handwritten ancient Tamil scripts. Zernike moments and regional features are extracted from the Tamil characters and they are formed as feature vectors. These feature vectors are trained using CELM. It is found from the experimental results, that when Zernike moments combined with regional features and used as a feature vector for training, a highest classification accuracy rate of 82.63% is obtained. The main advantage of using CELM is that, the parameters need not be tuned manually and iteratively and more over the solution obtained is unique least norm square solution. Different combinations of CELM can be used which helps to increase the classification accuracy further.

References

- [1] Venkatesh J and Sureshkumar C, "Tamil Handwritten Character Recognition using Kohonon's Self Organizing Map", International Journal of Computer Science and Network Security, Vol.9, No.12, pp.no:156-161, 2009.
- [2] Jagadeesh Kannan R et al, "Off-Line Cursive Handwritten Tamil Character Recognition", International Conference on Security Technology, IEEE Computer Society, pp.no:159-164, 2008.
- [3] Rajakumar S and Subbiah Bharathi V, "7th Century Ancient Tamil Character Recognition from Stone Inscriptions", Indian Journal of Computer Science and Engineering, Vol.3, No.5, pp.no:673-677, 2012.
- [4] Heutte L et al, "A structural/statistical features based vector for handwritten character recognition", Pattern Recognition Letters 19, pp.no: 629 – 641, 1998.
- [5] Sridevi .N and Subashini .P, "Moment Based Feature Extraction for Classification of Handwritten Ancient Tamil Scripts", International Journal of Emerging trends in Engineering and Development, Vol.7, No.2, pp.no: 106-115, 2012.
- [6] Teague .M.R, "Image Analysis via the General Theory of Moments", Journal of the Optical Society of America, Vol.70, No.8, pp.no:920-930, 1979.
- [7] Dr.N.Sengottaiyan et al, "Handwritten South Indian Language Recognition using Artificial Neural Network", International Journal of Advanced Research in Technology, Vol. 1, No.1, pp.no: 87 – 90, September 2011.
- [8] Ming-Bin Li et al, "Fully complex extreme learning machine", Neurocomputing 68, pp.no: 306 – 314, 2005.
- [9] Guang-Bin Huang et al, "Extreme learning machine: Theory and applications", Neurocomputing 70, pp.no: 489 – 501, 2006.
- [10] Kim .T and Adali .T, "Approximation by fully complex multilayer perceptrons", Neural Computing 15 (7), pp.no: 1641 – 1666, 2003.