Offline Tamil Handwritten Character Recognition using Chain Code and Zone based Features

M. Antony Robert Raj[1], S. Abirami[2]
Department of Information Science and Technology
Anna University, Chennai – 600 025
Email: antorobert@gmail.com, abirami_mr@yahoo.com

Abstract.

Recognition of handwritten characters is one of the important research areas in Pattern Recognition. But recognition is more complicated and tough for the languages/scripts having more curvy structures. This is more applicable to Indian languages since most of the Indian languages have tough structural shapes and of course the recognition is highly challengeable. In this context, this paper is intended to design an offline Tamil handwritten character recognizer which would extract the necessary features to recognize the characters that may produce conflicts. The novelty of this paper lies in extracting the features from the Tamil handwritten characters by applying Zoning over them and capturing their structures through the usage of Chain Code Algorithm. Further, the selected features are analyzed to calculate their area which is an additional feature. Finally, extracted values are given as input to Support Vector Machine to classify the characters successfully. In this research, Tamil vowels and consonants have been taken into consideration and it achieves an average rate of 80% accuracy in recognition.

Keywords: OCR, Zoning, Chain Code and SVM

1. Introduction

Optical character recognition (OCR) is one of the most popularized research fields in the pattern recognition, which helps to interpret the handwritten or printed text image as to machine understandable and modifiable text. Converting character image into recognizable text by machine is a highly complicated work for researchers. The reason behind this is handwriting style of every individual is unique. The style of individual handwriting will vary with respect to mood, age, education, situation and so on.

The OCR process is of two types; offline and online with respect to handwritten recognition. Online handwritten OCR helps to identify the characteristic features from pen tip movement. Offline process has not been researched as extensively as the online one, where contributions are available in many languages including Tamil. Offline procedure limits the recognition of written characters because of its complications. In this paper, we have chosen Offline Mode Recognition procedure for Tamil handwritten character set.

Tamil, a South Indian language has complex structure due to its cursive shapes and loops. Tamil language got great recognition '*Semmozhi*' from Indian government. In character set, Tamil alphabet contains 247 characters which include 18 consonants, 12 vowels, 216 combinational characters and one special character. We have chosen thirty characters (vowels and consonants) from Tamil Character set in this work.

Tamil is one of the recognized languages in many countries. Many of ancient Tamil handwritten invaluable documents can be available to the common man through computerization and they can reap their benefits. Our project is envisaged to accelerate the process of recognizing difficult handwritten documents.

The OCR system consists of five main steps: Pre-processing, Segmentation, Feature Selection, Feature Extraction and Classification.

Section-2 deals with the survey of work done in this area. Section-3 describes the pre-processing steps, feature extraction and classification. Experimental results are presented in section-4. Finally, conclusion and future works are discussed in Section-5.

2. Literature Survey

Jun Cao eta al [3] developed algorithm in neural network to classify the exact handwritten numerals. To extract the features, they used the Directional Chain Code Histogram based procedure and achieved lowest error rate.

Rajashekararadhya et al [4] employed Zoning and projection based handwritten character recognition system. Those two procedures were used for extracting the features. Here again zoning [8] [10] [11] method was employed and extracted zone centroid coordinates as features. Support vector machine (SVM) system and Neural network were chosen for providing better recognition rate.

N. Shanthi and Duraisamy [6] [12] achieved considerable accurate results from SVM based recognition system. Zoning has been implemented and pixel density values are calculated to extract the needed features from character image.

R. Ramanathan et al [7] took Tamil font for their recognition research. SVM algorithm was used for identifying several fonts in Tamil. Gabor filter was used to extract the features and they were trained by SVM to achieve reasonable accuracy for testing samples.

Akshay et al [9] did the experiments by taking ‘Horizontal and Vertical line, their position’, ‘branching and its position’ as features. He also explored zone, movements and number of transitions for verifying their utility. Euclidean Distance was used for classifying the character

3. Zoning based Handwritten OCR system

In the Pre-processing phase, Otsu’s [1] method has been followed to binarize the character image. Binarized image was given as an input to noise removal process, where Median and Gaussian filters were applied to fine-tune the image. Skeletonization was another method to secure thinner image. Thinning algorithm [4] has been added later to peel off the thickness of the image. Normalization has been applied over the images of various sizes to obtain a standard size.

Zoning has been applied over the normalized image to divide the image into equal pieces. Chain Code Procedure has been applied on the character portion of the pieces in each zone to select the required portion of character structure for further process. Area calculation has been done on the next level to extract the necessary features used for Classification. Support Vector Machine (SVM) is a popular classifier used in our work to classify the exact machine recognizable character. The figure 1 shows the overall procedure followed in the forthcoming sections.

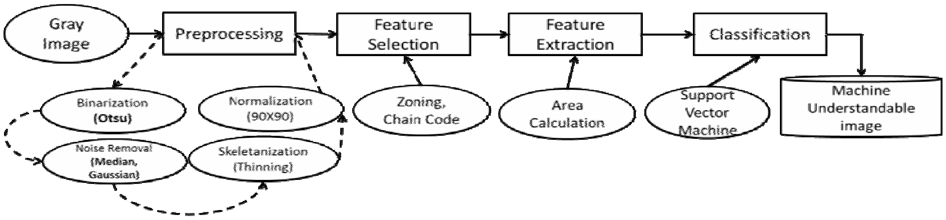


Figure.1 Zoning based Handwritten OCR system

3.1 Pre-processing

Pre-processing is required to obtain the images in a good condition. The collected character samples were converted into a black and white image using Binarization and thinned further to apply feature selection techniques on it.

3.1.1 Binarization

Binarization is a procedure used to convert grey scale image into black and white images. Thresholding is a method which is used in Binarization. Thresholding task is used to distinguish foreground from background. There are two types of Thresholding [1] techniques Local and Global. Global approach has been taken in this work which has only one peak value for the entire image instead of many peak values for the entire image. Otsu's [1][3] Thresholding concepts are implemented in our work where peak values were calculated within two classes of black and white to separate foreground and background.

3.1.2 Skeletonization

It is a process to peel off the unwanted pixels from the character image without affecting the natural shape of the real character. Thinning algorithm has been applied over the binarized image to obtain its skeleton. This reduces the size of the image, retaining single pixel in the contour of the image.

3.2 Feature Selection and Extraction

3.2.1 Feature Selection

To select features from the image, zone–based approach [2] has been proposed here. To extract the features, selection plays a major role in recognition system. In many works, zoning procedure [4] [8] [10] [11] has been used for extracting features but in our work we chose zoning for selecting the necessary features. Zoning not only helped us to extract good features, but also aids to apply Eight Directional Chain Code [3] [5] method over it to extract crucial features from each zone. Without affecting the real shape, the image was resized into a standard size (90×90 pixels) image to apply zoning over the image. In zoning way, normalized image (90×90) was further divided into nine equal zones (10×10). Now each zone contained character portion of image as shown on the figure 2.

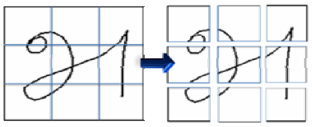


Figure.2 Sample Character in Zone (90X90 → 9(10X10))

Then the Eight Directional Chain Code [3] [5] method has been applied on each zone to extract the features. Chain code was applied on character pixels in an anticlockwise direction as depicted in figure 3.

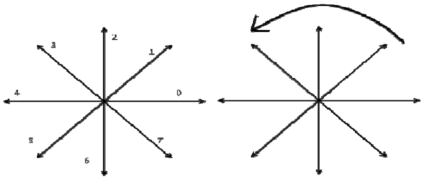


Figure.3 Chain Code Procedure

In chain code procedure, we travelled row wise from first pixel on the sub image to find the first black pixel with one neighbor. If we found it, then the chain code method was applied on the same pixel and its travel started on anticlockwise direction as shown in the figure 3 and checked on eight adjacent pixels for finding neighbor black pixel. The chain code will not traverse reversely. This procedure came to an end with another black pixel which contained only one neighbor or any junction point was found (if there are three neighbors to the current pixel). The chain code extracted the one pixel if there was only one pixel in zone (dots). Visited foreground (black) pixels were selected for further processes. Continued these steps until no further black pixels were available to visit by chain code.

If the chain code algorithm was not able to find the pixel with one neighbor in the whole image, we checked each pixel which contains two neighbors. If we found it, it might have circular shapes. Here also, visited all black pixels and selected the pixel portion for further process. The figure 4 shows selected features from the character ‘அ’

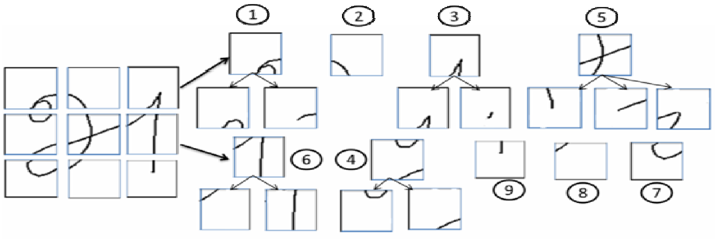


Figure.4 Selected Features from ‘அ’

3.2.2 Feature Extraction

Various parts of the character are segregated using the feature selection algorithm which results in different shapes. Segregated components possess the shape of curves (open or closed curve), points or linear form. Here, curves and closed curves are not considered for experimentation since they are more irregular in shapes. Instead of taking various shapes for our work, we calculated area of the curves to extract the features from the characters.

In this area calculation, we recognized both the ends of every given shape, by travelling along each row of the image from top left corner seeking for the pixels having only one neighbor. If pixels with only one neighbor have been identified, those points were taken as the two end points A1 & A2 ((r1, c1), (r2, c2)) of that particular shape and assumed as line or open curve. If we were unable to find the pixel with one neighbor, we searched for pixel with two neighbors and considered them as closed curves.

From the point A1 & A2, we computed the mid-point of the two end points using the Eq-1 and it is taken as A3 (r3, c3). Later, the mid-point of A1 to A3 (taken as A4 (r4, c4)) and A3 to A2 (taken as A5 (r5, c5)) has been calculated subsequently.

$$\mathbf{Dx} = \left(\frac{(\mathbf{r1} + \mathbf{r2})}{\mathbf{2}}, \frac{(c\mathbf{1} + c\mathbf{2})}{\mathbf{2}}\right) \qquad - \qquad (1)$$

Further, distance between A1 to A4, A4 to A3, A3 to A5 and A5 to A2 has been calculated as shown in figure 5. All the measurements seem to be the same and hence taken as ‘D’.

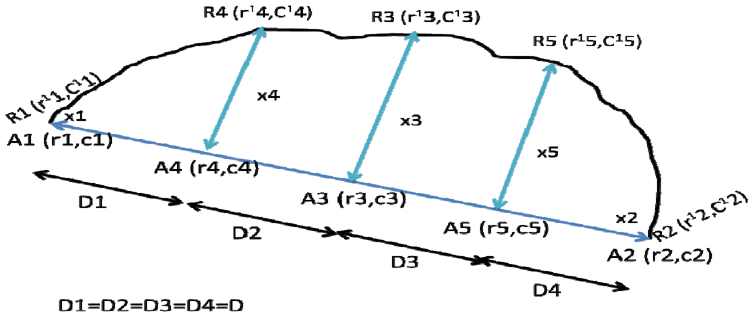


Figure.5 Extracting features from selected sample

We traversed from each mid-point (A1, A2, A3, A4 and A5) to the curve direction till we located the black pixel and marked it as R1, R2, R3, R4 and R5. The distance between R1 to A1, R2 to A2 and etc. were calculated and they are taken as x1, x2, x3, x4 and x5 respectively as shown in equation 2.

$$x = \sqrt{(ri - ri1)^2 + (ci - ci1)^2} \quad - \qquad (2)$$

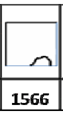
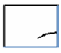
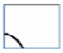
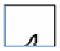
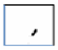
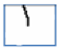
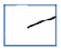
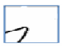
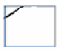
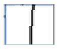
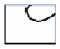
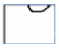
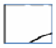
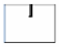
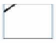
Figure 5 shows the Procedure of extracting features from given samples. The distance between R1 to A1 and R2 to A2 were zero. Finally the area was calculated using the given equation (Eq-3).

$$\text{Area} = D\left(\frac{x1}{2} + x4 + x3 + x5 + \frac{x2}{2}\right) \quad - \qquad (3)$$

If the shape was linear, the area of the line might be zero. If the shape occurs to be a closed curve, then we considered it inside the bounding box, and took the mid-point of the bounding box by drawing the horizontal line through it which in turn divides the closed curve into two curve shapes and also we did the above said procedure for finding the area calculation.

From each zone (9 Sub zones), we extracted needed features from 14 to 16 selected character portions of the Tamil Character 'அ' as shown in the Table 1

Table.1 Extracted Feature from Each Character Portions

Character Portion															
AREA	1566	3819	3916	1003	81	1088	52	4298	2240	50700	7004	3072	3613	23	71

3.3 Classification

Support Vector Machine (SVM) [6] [7] [13][14] is a statistical learning algorithm which achieved excellent results in various applications including pattern recognition. SVM algorithm classifies the samples by using support vectors which were the sub sets of training samples. The feature space was created by SVM classifier using the attributes in the training data collected from various zones of image sets.

Linear separation of hyper plane and complexity are defined using kernel functions. $(w, x) + b = 0$ gave as the hyper plane, were 'w' is the normal vector of the plane (weight factor). The given set of labeled training samples are X_i ,Y_i; where i=1 to n and X_i is the sample sets and Y= 1 or -1. The SVM gave us the solution from the equation [6] Y_i $(W^T \phi (X_i) + b) \geq \xi_I$, $\xi_I \geq 0$. The testing and training samples were collected and fed into classification algorithm. SVM predicted the values targeted by us from the given testing samples. The RBF is one of the most famous kernel type which was implemented in our work to predict our samples.

4. Experimental Results

4.1 Data collection

Thirty characters (vowels and consonants) from Tamil Language were chosen for our experiments. Data samples were collected from HP Lab Dataset, where we have gathered 7500 characters of our experiments. From these data sets 5000 character samples were chosen for Training set and rest of them were chosen for testing set. This data stored in Microsoft excel and fed in Matlab, where the data was loaded in SVM function to get the accuracy rate.

Overall accuracy achieved for testing samples was 80.4%. The experiment result we have achieved for different characters are tabulated in the following table 2. The figure 7 shows the graph representation of accuracy rate for each character set.

Table.2 Data Samples and Accuracy Achieved

S. No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Tamil Vowels	அ	ஆ	இ	ஈ	உ	ஊ	எ	ஐ	ஏ	ஒ	ஓ	ஔ	க	ங	ச
Accuracy Achieved	80.4	75.2	81.6	90.2	78.1	70.3	79.6	78	80.1	81.1	76.8	73.4	80.2	78.1	79.9

S. No	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Tamil Vowels	ஞ	ட	ண	த	ந	ப	ம	ய	ர	ல	வ	ழ	ள	ற	ன
Accuracy Achieved	74.1	79.3	83	85.2	80.2	86.9	84.4	87	81.3	86	83.8	76.2	78.5	84.3	79.1

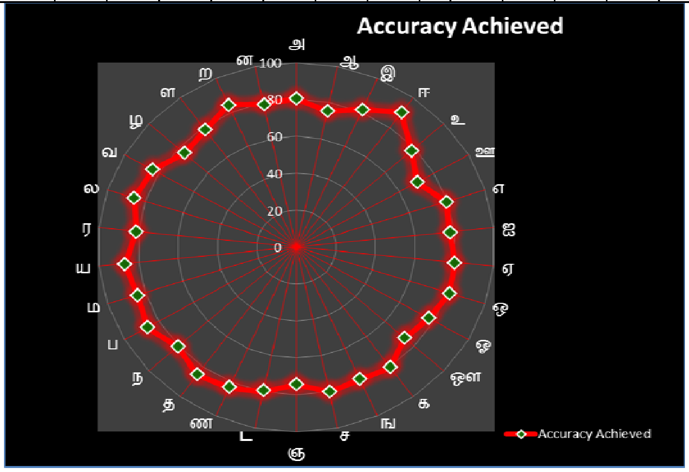


Figure.7 Accuracy rate for each character set

5. Conclusion

In this paper we implemented Zone and Chain Code Algorithm for selecting the correct features which is suitable for area calculation. We have chosen thirty letters from Tamil character set of HP datasets and achieved 80% accuracy rate. These concepts are highly suitable for other Tamil combinational characters also.

Our aim is to consider all Tamil characters for recognition. For that we are striving to fine tuning this chain code and Area calculation procedure. Also we are hoping to develop new algorithms which will complement the procedure which was used by us and which will be applicable to recognize all the characters.

References

[1] Antony Robert Raj. M and S. Abirami, “A Survey on Tamil Handwritten Character Recognition using OCR techniques”, The Second International Conference on Computer Science, Engineering and Applications (CCSEA), 05, pp.115-127, 2012.

[2] Antony Robert Raj. M and S. Abirami, “Analysis of Statistical Feature Extraction Approaches used in Tamil Handwritten OCR”, 13th Tamil Internet Conference- INFITT, pp.144-150, 2013

[3] Jun Cao, M. Ahmadi and M. Shridhar, ”Recognition of Handwritten Numerals with Mutable feature and Multistage Classifier” , Elsevier, Pattern Recognition, Vol 28, No.2, pp: 153 – 160, 1995

[4] S.V Rajashekararadhya, Vanaja Ranjan P, Manhunath Aradhya V N “Isolated Handwritten Kannada and Tamil Numeral Recognition: A Novel Approach”, First IEEE

International Conference on Emerging Trends in engineering and Technology, Page(s): 1192-1195, 2008.

[5] Cheng-Lin Liu, Kazuki Nakashima, Hiroshi Sako, Himromichi Fujisawa, "Handwritten Digit Recognition: Benchmarking of State-of-Art Techniques", Elsevier, Pattern Recognition, Vol 36, pp: 2271 – 2285, 2003

[6] N. Shanthi, K. Duraiswami, "A Novel SVM -based Handwritten Tamil character recognition system", Springer, Pattern Analysis & Application, Vol –13, No.2, 173-180,2010.

[7] Ramanathan R, Ponmathavan S, Thaneshwaran L, Arun.S.Nair, and Valliappan N, "Tamil font Recognition Using Gabor and Support vector machines", International Conference on Advances in Computing, Control, & Telecommunication Technologies, page(s): 613 – 615, 2009.

[8] Rajashekararadhya S.V and Vanaja Ranjan P, "Zone-Based Hybrid Feature Extraction Algorithm for Handwritten Numeral Recognition of two popular Indian Script", World Congress on Nature & Biologically Inspired Computing, page(s): 526 – 530, 2009.

[9] Akshay Apte and Harshad Gado, "Tamil character recognition using structural features" ,2010.

[10] Rajashekararadhya S.V and Vanaja Ranjan P, "Neural Network Based Handwritten Numeral Recognition of Kannada and Telugu Script", IEEE TENCON Conference, pp 1-5, 2008

[11] Rajashekararadhya S.V and Vanaja Ranjan P, "Efficient Zone based Feature Extraction Algorithm for Handwritten Numeral Recognition of Four Popular south Indian Scripts". Int. J. of Theoretical and Applied Information Technology, pages: 1171 – 1181, 2008

[12] Shanthi N and Duraiswami K, "Performance Comparison of Different Image size for Recognizing unconstrained Handwritten Tamil character using SVM", Journal of Computer Science Vol-3 (9): Page(3) 760-764, 2007

[13] Bhattacharya U, Ghosh S.K and Parui S.K, "A Two Stage Recognition Scheme for Handwritten Tamil Characters", Ninth International Conference on Document Analysis and Recognition, Vol: 1 page(s): 511 – 515, 2007

[14] Sukalpa Chanda, Srikanta Pal and UmapadaPal, (2008)," Word-wise Sinhala Tamil and English Script Identification Using Gaussian Kernel SVM",IEEE 2008.