FEATURE EXTRACTION TECHNIQUE BASED CHARACTER RECOGNITION USING ARTIFICIAL NEURAL NETWORK FOR SINHALA CHARACTERS

JM.Harshana Madusanka Jayamaha[1] and HMM. Naleer[2]

[1,2]Department of mathematical Sciences, South Eastern university of Sri Lanka, Sri Lanka
jmhmjayamaha@gmail.com, hmmnaleer@gmail.com

Abstract:

This paper describes a basic approach, taken for Sinhala handwritten character recognition. The research was performed with the idea of identifying most efficient, effective and accurate method, based on character geometry based feature extraction technique for Sinhala handwritten character recognition. Data acquisition, digitalization, preprocessing, feature extraction was done using the image processing techniques. The classification was measured using an ANN based classifier on a common testing and training data sets. The classification performance was measured for 34 Sinhala characters using this research. Finally, recognized Sinhala character will be printed on a text document.

Keywords:Artificial Neural Network; Handwritten Recognition; Image Processing; Feature Extraction

Introduction

Handwritten character recognition has become one of the most exciting and challenging research area in computer vision and pattern recognition in the recent years. Handwritten character identification is one of artificial intelligence tasks that belongs to the scientific discipline called pattern recognition. Character recognition is a complicated task that uses complicated logics and theories with it and, it requires much more effort to improve the accuracy and performance of the system.

According to the literature handwritten character identification is a thirty-year-old researching discipline and the applications related to the printed character recognition and handwritten character recognition is the oldest and the first applications related to the field of pattern recognition [1].

The process of identification handwritten characters can be considered as a very useful activity. Because it is easy to make modifications on a digital document rather than editing the content of a document written on a paper. Therefore a number of various classification methods are used for online and offline character recognition.

The artificial neural network can be considered as a most appropriate and most efficient method for activities like character pattern identification, which does not undergo any of mathematical algorithms. Currently, the number of classificationmethods and feature extraction methods are used for Sinhala character recognition.

Geometry based feature extraction is one method that can be used to collect character features from the Sinhala characters. Since that from this research, it was tested that the usage of geometry based feature extraction for Sinhala character identification.

Methodology

The proposed system uses feature extraction to detect and classify the handwritten text and, some technique is used for removing the background noise.The methodology consists of four phases.

1. Pre-processing
2. Segmentation
3. Feature extraction
4. Classification and recognition.

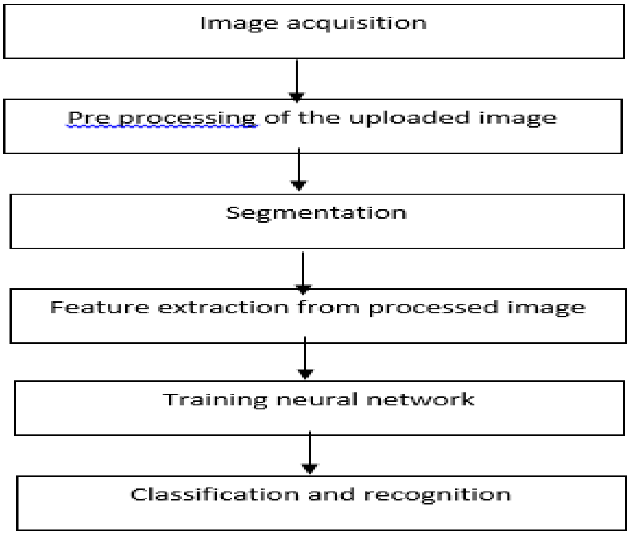


Image acquisition
Pre processing of the uploaded image
Segmentation
Feature extraction from processed image
Training neural network
Classification and recognition

Figure 1 Block diagram of proposed method

Image acquisition

Is the first process. Note that acquisition could be as simple as being given an image that is already in adigital form generally, the image acquisition stage involves preprocessing, such as scaling.

Pre-processing

The pre-processing is a series of operations performed on thescanned input image. It essentially enhances the image rendering it suitable for segmentation. The role of pre-processing is to segment the interesting pattern from the background. Generally, noise filtering, smoothing, and normalization should be done in this step. The pre-processing also defines a compact representation of the pattern. Binarization process converts a gray scale image into a binary image. Dilation of edges in the binarized image is done using Sobel technique.

Segmentation

Procedures partition an image into its constituent parts or objects. In general, autonomous segmentation is one of the most difficult tasks in digital image processing. A rugged segmentation procedure brings the process a long way toward asuccessful solution of imaging problems that require objects to be identified individually. On the other hand, weak or erratic segmentation algorithms almost always guarantee eventual failure. In general, the more accurate the segmentation, the more likely recognition is to succeed.

Feature extraction

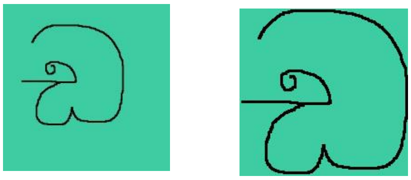
It extracts different line types that form a particular character. It also concentrates on the positional features of the same. The feature extraction technique explained was tested using a Neural Network which was trained with the feature vectors obtained from the system proposed.

Feature extraction based on character geometry

It extracts different line types that form a particular character. It also concentrates on the positional features of the same. The feature extraction technique explained was tested using a Neural Network which was trained with the feature vectors obtained from the system proposed.

Universe of discourse

The universe of discourse is defined as the shortest matrix that fits the entire character skeleton. The Universe of discourse is selected because the features extracted from the character image include the positions of different line segments in the character image. So every character image should be independent of its Image size.



(a) Original image (b)Universe of discourse

Zoning

The image is divided into windows of equal size, and the feature is done on individual windows. The image was zoned into 9 equal-sized windows. Feature extraction was applied to individual zones, rather than the whole image. This gives more information about fine details of character skeleton. Also, positions of different line segments in a character skeleton become a feature if zoning is used. This is because a particular line segment of a character occurs in a particular zone in almost cases.

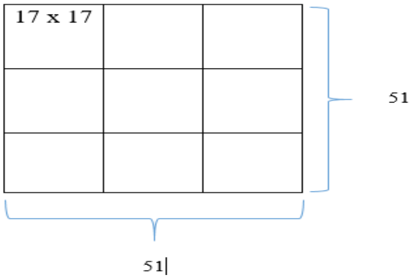


Figure 2. Divided windows of equal size

To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton were defined as starters, intersections, and minor starters.

Starters

Starters are those pixels with one neighbor in the character skeleton. Before character traversal starts, all the starters in the particular zone are found and are populated in a list.

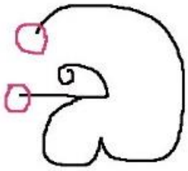


Figure 3. Starters are rounded

Intersections

The necessary but insufficient criterion for a pixel to be an intersection is that it should have more than one neighbor. A new property called true neighbors is defined for each pixel. Based on the number of true neighbors for a particular pixel, it is classified as an intersection or not. For this, neighboring pixels are classified into two categories, Direct pixels, and diagonal pixels.

Direct pixels are all those pixels in the neighborhood of the pixel under consideration in the horizontal and vertical directions. Diagonal pixels are the remaining pixels in the neighborhood which are in a diagonal direction to the pixel under consideration.

Now for finding a number of true neighbors for the pixel under consideration, it has to be classified further based on the number of neighbors it has in the character skeleton. Pixels under consideration are classified as those with 3 neighbors: If any one of the direct pixels is adjacent to anyone of the diagonal pixels, then the pixel under consideration cannot be an intersection, else if none of the neighboring pixels are adjacent to each other than it's an intersection. 4 neighbors:

If each and every direct pixel has an adjacent diagonal pixel or vice-versa, then the pixel under consideration cannot be considered as an intersection. 5 or neighbors: If the pixel under consideration has five or more neighbors, then it is always considered as an intersection once all the intersections are identified in the image, then they are populated in a list.

After the line type of each segment is determined, thefeature vector is formed based on this information. Every zone has a feature vector corresponding to it. Under the algorithm proposed, every zone has a feature vector with a length of 8.

The contents of each zone feature vector are:

1) A number of horizontal lines.
2) A number of vertical lines.
3) A number of Right diagonal lines.
4) A number of Left diagonal lines.

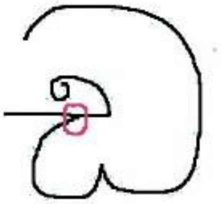


Figure4. Intersections

5) Normalized Length of all vertical lines.
6) Normalized Length of all horizontal lines.
7) Normalized Length of all right diagonal lines.
8) Normalized Length of all left diagonal lines.
9) Normalized Area of the Skeleton.

The number of any particular line type is normalized using the following method,
Value = 1 - ((number of lines/10) x 2).

Normalized length of any particular line type is found using the following method,
Length = (Total Pixels in that line type)/ (Total zone pixels).

The feature vector explained here is extracted individually for each zone. So if there are N zones, there will be 9N elements in feature vector for each zone. For the system proposed, the original image was first zoned into 9 zones by dividing the image matrix. The features were then extracted for each zone. Again the original image was divided into 3 zones by dividing in the horizontal direction. Then features were extracted for each such zone.After zonal feature extraction, certain features were extracted for the entire image based on the regional propertiesnamely Euler Number:

It is defined as the difference of number of objects and number of holes in the image. For instance, a perfectly drawn ’A’ would have Euler number as zero, since number of objects is 1 and number of holes are 2, whereas ‘B’ would have Euler number as -1, since it has two holes Regional Area: It is defined as the ratio of the number of the pixels in the skeleton to the total number of pixels in the image. Eccentricity: It is defined as the eccentricity of the smallest ellipse that fits the skeleton of the image.

Training neural network
Design for the artificial neural network.

The proposed neural network for the research can be represented in graphical form as below figure 5. The training features from the characters are extracted using the feature extraction technique as mentioned in the above section. The ANN is provided 108 feature values from the character features.

The artificial neural network used for recognizing the handwritten and printed font characters is contained in three layers.The Experimentally finalized parameters for the artificial neural network for a training set of 850 characters of handwritten characters is as follows

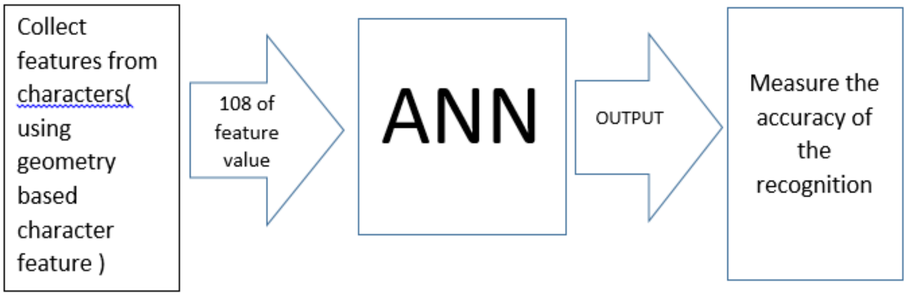

Collect features from characters(using geometry based character feature)
108 of feature value
ANN
OUTPUT
Measure the accuracy of the recognition

Figure 5. Overall Architecture of ANN

Implementation of the input layer.

The input layer for the neural network is contained 108 nodesitself, the layers of the ANN were represented through thetwo-dimensional matrix (108 x 850).

Table 1 Parameters Used for the ANN

Number of layers	Node of layers	
3	Input	108
	Hidden	78
	Output	34

Implementation of hidden layers.

The hidden layers also represented through theone-dimensionalarray. The size of the array is depended on the number of nodes used for the hidden layer. For the implementation of the neural network, it was used 71 nodes for the hidden layer. The outputs calculation associated with the hidden nodes are based on the tangent sigmoid function.

Implementation of the output layer.

The output layer of the neural network is represented using a one-dimensional double type array with 34 indexes. The array may store the result values of hidden layer after applying the activation function on them.

Result and Discussion

In order to test the ANN for the character identification, the neural network was trained using character patterns. The convergence of the ANN can be monitored using the graph drawn between mean square error and the number of iterations. With the time the error of the neural network has to be reduced.

The following graph represented at thefollowingfigure shows the minimization of the error with the iterations.Using the PC with Intel core i5 – 6200u @ 2.30 GHz processor and 8GB RAM with Windows 10 premium environment thefollowing result were obtained.

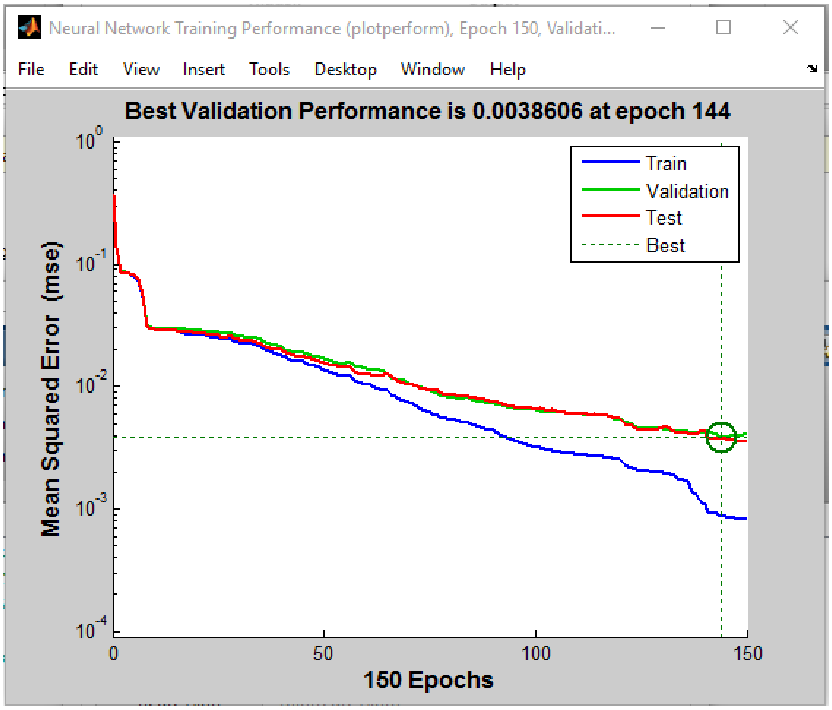


Figure 6 Iterations Vs Mean squared error

Since it is clear the feature extraction method can be used for Sinhala character identification. For the training set, ANN based system shows better results than existing other methods. But currently, it has some problems with classification and has to be improved. Being ANN is an intelligent methodology, since much more training for the ANN is required to obtain better results than obtained.

Table 2. Final Results

Technique Used	**Total Character in database**	**No: of Training characters**	**No: of Testing characters**	**Performance**
Artificial Neural Network	850	680	170	82.1%

Conclusion

The proposed neural network architecture has an ability to classify the character patterns in some degree. But it shows difficulties during the classification of unknown samples. Since as a future enhancement, it is expected to improve the current architecture of the ANN by increasing the number of nodes and layers. Since all of the performance in the character pattern identification is based on feature extraction methodology, it is important to make error free (noise free) character features. Since as the next step, it is expected to insert high-level image processing techniques for the feature extraction process.Since the system is not able to identify the touching characters,water reservoir concept has to be used in future.

References

[1]. Hewavitharana, S., Fernando, H.C., Kodikara, N.D.(2002) Offline Sinhala Handwriting Recognition Using Hidden Markov Models, Indian Conference On Computer Vision , Graphics & Image Processing (Icvgip), Ahmedabad, India.

[2]. Dinesh Deleep, A Feature Extraction Technique Based On Character Geometry For Character Recognition.

[3]. Sandhya Arora,Debotosh Bhattacharjee,Mita Nasipuri, L.Malik,M.Kundu, D.K.Basu,(July 2010) Performance Comparison Of Svm And Ann For Handwritten Devanagari Character Recognition, International Journal Of Computer Science Issues (Ijcsi), Vol. 7 Issue 4, P18.

[4]. Juan R Rabunal &Julian Dorado.(2006)Artificial Neural Networks In Real-Life Applications, Idea Group Publications,

[5]. Gunaratna, D.A.K.S., Kodikara, N.D., Premarathne, H L.(2008) An Intelligent Recognition System For Sri Lankan Currency Notes,Sri Lanka.

[6]. Nilupa Liyanage , Asoka S. Karunananda, (2008) Using Neural Networks For Recognition Of Handwritten Mathematical Documents , Proceedings Of The 5th Annual Sessions Of Sri Lanka Association For Artificial Intelligence, Sri Lanka, Pp 21-25.

[7]. Rohana K. Rajapakse, A. Ruvanweerasinghe And E. Kevin Seneviratne,(1995) A Neural Network Based Character Recognition System For Sinhalascript,

[8]. Karunanayaka, M.L.M., Kodikara, N.D., Wimalaratne, G.D.S.P.(29 Nov- 01 Dec 2004) Off-Line Sinhala Handwriting Recognition With An Application For Postal City Name Recognition Conference Proceedings - 6th International Information Technology Conference On From Research To Reality, Infotel Lanka Society Colombo, Sri Lanka, Pp. 23-29, Isbn 955-8974-01-3.

[9]. Line Eikvil, (1993) Optical Character Recognition.

[10]. Ranpreet Karu,Baljith Singh,(2011) A Hybrid Neural Approach For Character Recognition System,(Ijcsit) International Journal Of Computer Science And Information Technologies, Vol. 2 (2) , 721-726.

[11]. Nor Amizam Jusoh, Jasni Mohamad Zain,(November 2009)Application Of Freeman Chain Codes: An Alternative Recognition Technique For Malaysian Car Plates, International Journal Of Computer Science And Network Security, Vol. 9 No. 11 Pp. 222-227.

[12]. M.Fatih Amasyali,Nilgun Erdem, Hakan Haberdar,Filiz Koyuncu,"Neuro-Chain: A Handwritten Character Recognition System".

[13]. James A. Freeman, David M. Skapura,(1991)Fundamentals Of Neural Networks Architectures, Algorithms And Applications.Addison-Wesley Publishing Company,

[14]. Ben Krose, Patrick Van Der Smagt. (1996) An Introduction To Neural Networks.

[15]. Tom M Mitchell.(1997)Machine Learning. International Ed.Singapore: Mcgraw-Hill,

[16]. Hewavitharana, S&Kodikara, N.D. A Statistical Approach To Sinhala Character Recognition.

[17]. Rafael C. Gonzalez, Richard E. Woods, Steven L.Eddins. Digital Image Processing Using Matlab , Second Edition.

[18]. R.Gonzalez, R.Woods. Digital Image Processing 3rd Edition.