Performance Measure of Scanned Tamil Land Documents using Neural Network Approach

M. Malathi
Department of Computer Science
Government Arts and Science College (Autonomous)
Salem-636007, India

M. Srividya
Department of Computer Science
Government Arts and Science College (Autonomous)
Salem-636007, India

Abstract- Recognition is process to find out the noisy or distorted image to make an accurate image. Classification and recognition technique which implemented on scanned Tamil land document. In the pre-processing stage, the given dataset is filtered by using median filter. After that, segmentation process is applied for every word image is splitted into character. Then, feature extraction is done by Gabor wavelet. For post processing stage, classification is process to check out the dataset which using neural network technique like supervised learning method or unsupervised learning method to find out the correct and Incorrect classification measure using confusion matrix. Hence, finally implemented Gabor wavelet technique to find the feature extraction and selection and then classification is done new pattern recognition technique using MATLAB. We also find out performance of plotting function like Training state, Regression, Gradient and validation.

Keywords-*Neural Network;feature selection ;pattern recognition;plotting function*

*****

I. INTRODUCTION

Character Recognition is a successful technique in the field of patter recognition. Both Handwritten and printed or scanned documents may be recognized which depends on the quality of the input documents. We can recognize the inconsistent character like style and curvature of shape, size etc., It has two types of character recognition such an online character recognition and offline character recognition. Online character recognition is only used for handwritten documents to recognize the inconsistent character. By using digital pen, the users write on a screen which encourage improving the writing style them. For that, writing style can vary by person to person. It is the natural way to correct misrecognized character by using online character recognition. Offline character recognition is used for printed or scanned documents, encoded text and storage was tiny size only.

II. LITERATURE SURVEY

Recognition of Distorted character by using Edge Detection Algorithm who proposed by S.K.Thilagavathy and Dr. R.Indra Gandhi [2] were the authors deals about the distorted character recognize on Tamil, Telegu and Malayalam languages. It is a complicated process to recognize the distorted character from the unique languages by using Edge detector method.

Real Time Recovery of Text based on FPCA was proposed by Baby Sathya .S and Rajesh Kumar T [3] who deals about the distorted or old historical documents. Their aim is to make the digitization of all those documents and detecting the horizontal and vertical line of text from the image.

Offline Recognition of Image for content based Retrieval was proposed by Thakur D and S.Sikchi [4] whom discussed about the OCR images from the Devanagiri Script. By using, Neural Network who can recognize the character from distorted text from database

Statistical Textures Feature based on Handwritten and Printed Text Document classification in South Indian Documents was proposed by Mallikarjun Hangargea, K.C. Santoshb, Srikanth Dodd mania, Raj Mohan Pardeshia, [5] those authors explains about the texture feature like mean, standard deviation and entropy to classify a word by using K nearest neighbor classifier. It demonstrates by using Roman Script databases by publicly to get average rate.

A Literature Review on Handwritten character Recognition was proposed by Mansi Shah and Gordhan B Jethava [6] whom discussed on digitizing the scanned images by converting ASCII code .It deeply explains about the many classifiers with multilingual languages.

Database Development of Historical documents: skew correction and detection was experimented by S P Sachin, Banumathi K L and Vanitha R [7] who suggested on recognizes text by using Hough Transform Method and Base line method. It compared with existing algorithm to perform the accurate result.

Neural Network based Approach for recognition the text image who introduced by Gaurav Kumar and Pradeep Kumar Bhatia [8] whom explains about the multiple handwritings of different people with low resolution image to recognize by using Multiple Layer Feed Forward Network. It deeply

 ISSN: 2321-8169
3409 - 3412

concentrated on recognize character from text file by using neural network.

Performance of English Character Recognition with and without noise [9] and Survey and classification of Character Recognition system was proposed by Priya Sharma and Randir Singh [10] who deeply discussed on variation of handwritten which using different hardware to recognize the character by using single MPLNN with Gradient descent.

A system for offline Recognition of Handwritten recognition by using Malayalam script was proposed by Jomy John, Kannan Balakrishnan, and Pramod K. V [11] who has experiment by using Support Vector Machine. It concentrates on combined feature of curvature vector and gradient descent feature and to get accuracy about 97.6% on both online and offline handwritten documents.

Handwritten Sanskrit Character Recognition with pre-processing steps was proposed by R. Dinesh Kumar and Prof. Dr. J Suganthi [12] who explains about the degraded documents and handwritten character to recognize the feature extraction technique were applied by using Top Hat Transforms method.

A Novel Hybrid model for Tamil Handwritten Character Segmentation was proposed by Dr. S Pannirselvam and S Ponmani [13] who discussed about the noise removal technique with the help of median filters. We can segment each character as Horizontal and Vertical Profile of them with experiment of different database from the surroundings.

III. METHODOLOGY

Our aim is to classify a character wise classification in scanned Tamil Land documents. First, words are cropped from document. And then words are segmented into character. Statistical texture features are computed by Gabor Wavelet Technique. These features are integrated to form a single feature vectors which are then used for handwritten and printed text separation via Neural Network technique

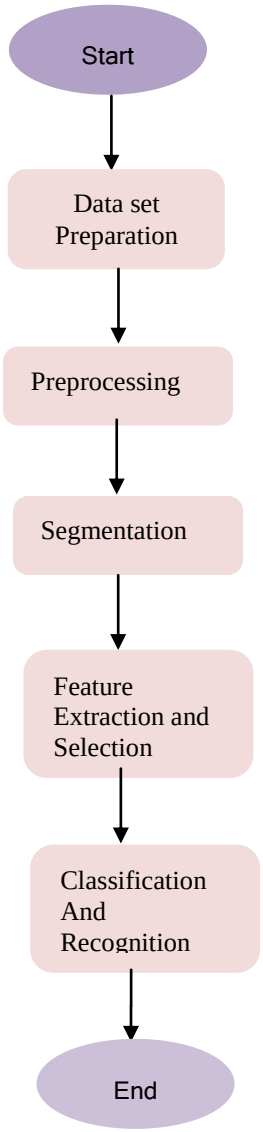
A. Dataset Preparation

We collect original land document Xerox from neighbor and friends nearly 50 document along with land property details. And, we scanned the entire property document which is very confidential one. Using cropping tool, we can cropped the all word from 20 documents. We omitted the repeated words and taken particular words were tested.

B. Pre-processing

First, the image is loaded as the input image. Filtering operations takes place for the input image. The median filter is used for the removal of noise and smoothen the image. The median filter is a nonlinear digital filtering technique, often used to remove noise. Such noise reduction is a typical pre-processing step to improve the results of later processing (for example, edge detection on an image). Median filtering is very widely used in digital image processing because, under certain conditions, it preserves edges while removing noise. Median filtering smoothen the image and is thus useful in reducing noise. Unlike low pass filtering, median filtering can preserve discontinuities in a step function and can smooth a few pixels whose values differ significantly from their surroundings without affecting the other pixels .It is also useful in preserving edges in an image while reducing random noise. Impulsive or salt-and pepper noise can occur due to a random bit error in a communication channel.

The following Fig 1. Shows the Systematic flow of classification



C. *Segmentation*

Segmentation is the main process for the character recognition. It divides an image into sub images. Like that, a line can be divided into word and word into character after preprocessing technique it made easily to get the required output. For each image, region of interest is more important for word and character and it is a crucial one for segmentation process.

D. *Feature selection and extraction*

It is the heart of a pattern recognition application. Feature extraction techniques like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Gradient based Feature Zoning and Euclidean distance .Histogram might be applied to extract the feature of individual characters. This stage is represented as a feature vector which it is identity one. The main goal is extract a set of features from recognized character. It improves the recognition rate and reduces the misclassification. Here, let's study and implement a Gabor wavelet feature technique to select and extract the feature of character individually. Gabor wavelets is a namely the interest point detection. There are several approaches to the interest point detection using Gabor functions or wavelets. More specifically, the two most common approaches involve the edge detection from the feature image or the corner detection using a combination of responses to several filters with a different orientation. It is the utilization of multi scale partial differential operator.

E. *classification*

Classification is a technique to sort out misclassified and classified values from the known data set .Hence; we implemented the classification and recognition technique using MATLAB. using command line function 'npr' which is a new pattern recognition method in a neural network method.we just passing Input and target values to the command function along with 'nntrain tool'. Back propagation is a neural network learning method. It is very usefulness for classification and prediction approach in data mining. It is well suited for continuous values to classify a data and high tolerance to find out the noisy data very easily.

IV. EXPERIMENTAL RESULT AND CONCLUSION

An approach to represent and compare word images, both on document and on natural domains. We show how an attributes-based approach based on a pyramidal histogram of characters can be used to learn how to embed the word images and their textual transcriptions into a shared, more discriminative space, where the similarity between words is independent of the writing and font style, illumination, capture angle, etc. This attributes representation leads to a unified representation of word images and strings, resulting in a method that allows one to perform query-by-example or query-by-string searches, as well as image transcription, in a unified framework. Here, preprocessing as noise removal and feature extraction and post processing like classification using neural network approach. We implement a neural network technique to find the correct classification as 92.8% and misclassification value as 8.2% using confusion matrix.

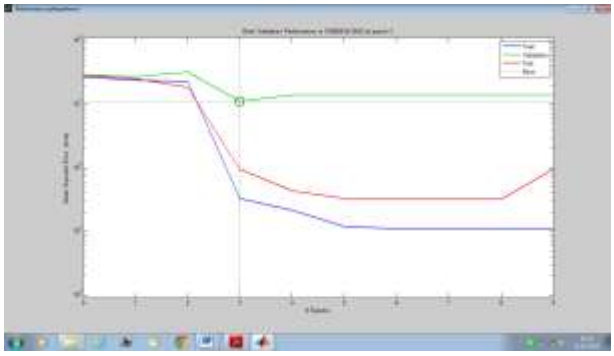


figure 2 Best Validation Performance

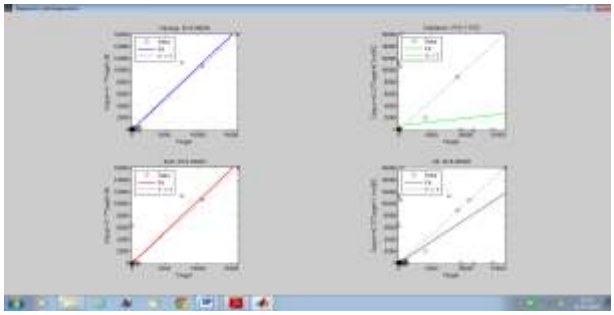


figure 3 Regression stage

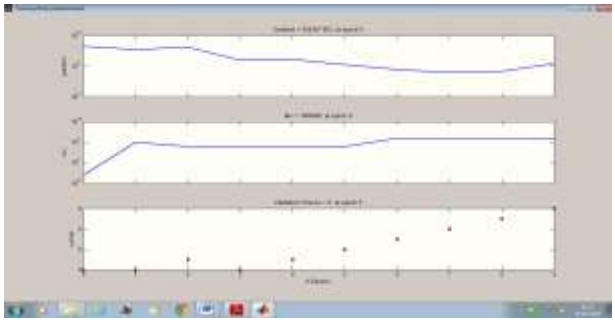


figure 4 Training stage for validation

figure 5 Confusion matrix

REFERENCES

[1] www.mathwork.com

[2] SK.Thilagavathy and Dr. R.Indhra Gandhi, "Recognition of Distorted character by using Edge Detection Algorithm", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 1, Issue 4, June 2013.

[3] Baby Sathya S and Rajesh Kumar T, "Real Time Recovery of Text based on FPGA", International Journal of Emerging Trends in Electrical and Electronics, Vol. 3, Issue. 3, May-2013, ISSN: 2320-9569.

[4] Thakur D and Prof S.Sikchi "Offline Recognition of Image for content Based Retrieval" International Journal Of Computer Science And Applications Vol. 6, No.2, Apr 2013, ISSN: 0974-1011 (Open Access).

[5] Mallikarjun Hangargea, K.C.Santoshb, Srikanth Doddamania,Rajmohan Pardeshia,"Statistical Texture Features based Handwritten and Printed Text Classification in South Indian Documents" Appeared in proceedings of ICECIT-2012 published by Elsevier Ltd. PP:215-221.

[6] Mansi Shah and Gordhan B Jethava, "A literature review on hand written character recognition", Indian Streams Research Journal, Vol -3 , Issue –2, March.2013, ISSN:-2230-7850.

[7] S P Sachin, Banumathi K L and Vanitha R,"Database development of historical documents: skew detection and correction",International Journal of Advanced Technology in Engineering and Science, Volume No.02, Issue No. 07, July 2014 ,ISSN (online): 2348 – 7550.

[8] Gaurav Kumar and Pradeep Kumar Bhatia,"Neural Network based Approach for Recognition of Text Images", International Journal of Computer Applications ,Volume-62, Issue No.14, January 2013,ISSN-0975 – 8887.

[9] Priya Sharma, Randhir Singh, "Performance of English Character Recognition with and without Noise", International Journal of Computer Trends and Technology,vol.4,Issue no.3,2013.

[10] Priya Sharma, Randhir Singh, "Survey and Classification of Character Recognition System", International Journal of Computer Trends and Technology- vol.4,Issue3- 2013.

[11] Jomy John, Kannan Balakrishnan, Pramod K. V, "A System for Offline Recognition of Handwritten Characters in Malayalam Script" , *I.J. Image, Graphics and Signal Processing, April* 2013, vol.4, pp.53-59 Published Online April 2013 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijigsp.2013.04.07.

[12] R. Dinesh Kumar , Prof . Dr. J. Suganthi , "Handwritten Sanskrit character Recognition with Preprocessing stages" International Journal of Modern Computer Science and Applications (IJMCSA) Vol-1, Issue No.-5, November, 2013, ISSN: 2321-2632 (Online).

[13] Dr.S.Pannirselvam, and S Ponmani, "A Novel Hybrid Model for Tamil Handwritten Character Segmentation", International Journal of Scientific & Engineering Research, Vol. 5, Issue 11, November-2014 , ISSN 2229-5518.