A COMPARATIVE ANALYSIS OF FEATURE EXTRACTION TECHNIQUES FOR HANDWRITTEN CHARACTER RECOGNITION

Rajbala Tokas1,Aruna Bhadu2
1 M.Tech*(CS),Swami Keshwanand Institute of Technology,Jaipur,Rajasthan,India,2 M.Tech*(SE) Govt. Engineering College Bikaner,Rajasthan,India.
[1]Rajtokas26@gmail.com,
[2]aruna7bhadu@gmail.com

ABSTRACT

Image processing and pattern recognition plays a lead role in handwritten character recognition. There are three main steps of handwritten character recognition- Data collection and pre-processing, feature extraction and classification. In this paper, we have presented different feature extraction methods to classify the 26 handwritten capital alphabets written by 25 different writers with their advantage& disadvantage &compression to each other. Analysis of these feature extraction methods with Back propagation neural network classifier has been done. Neural network is the classifier which we are using for classification with most of the feature vector types.

1. INTRODUCTION

There are verities of writing style of handwritten character that vary from writer to another writes. Due to this as compared to the humans it becomes very hard task for the computer to perform handwritten character recognition. The main reason that accounts for the absence of an accurate system for the handwritten character recognition the difference in style of writing. The accuracy of the recognition of the hand written character depends upon the database used for the recognition. It will become very difficult to recognize a hand written character due to the unwanted slants, skews and curves. Looking at the past we can see that a let of methods have been developed for the handwritten character recognition .the neural network is used in most of the handwriting recognition methods due to its classification efficiency and ease of use.

1.1. Handwritten character recognition techniques

As far as most of the methods are considered, these are three steps in the handwritten character recognition i.e. pre-processing, feature extraction & classification. A clean character image is formed by the help of the pre-processing that can be used directly or indirectly by the feature extraction step. In tries to remove the repetition of values in data. And finally in the classification stage estimating an output class is done. It is clear that even human too will some time make mistakes when come to pattern recognition. Pattern distortion, presence of unwanted object or disoriented pattern will affect the accuracy take when it comes to the computer to recognize the handwritten characters. It will be more difficult. The research of recognition for handwritten character is still on the exploration stage and the rate of recognition is very low. So we need to find new ways or to improve the existed methods to solve the problem. Handwritten Character recognition is a form of pattern recognition process. In reality It is very difficult to achieve 100% accuracy even human too will make mistake when come to pattern recognition, Pattern distortion, presence of unwanted object or disoriented pattern will affected the percentage accuracy.

1.1.1Feature extraction

The main aim of feature extraction is to make improvement in the accuracy and speed of the classifies for the pattern recognition. The extraction of the features of the characters is done is such a way that the complete portion of binary image covered and there is a distinct property associated with the each position. So we can say that feature extraction is a precise way in which a pattern can be described. It is which a pattern can be described. It is one of the most important parts of any system using pattern recognition. In this paper a complete flowchart of hand written English character recognition is given .

(1) Capture the scan character

(2) Perform the normalization process

(3) Perform binarization

(4) Apply feature extraction technique (boundary tracing technique).

(5) Implement the neural n/w classifier

(6) Get the recognition characters.

1.1.2Recognition & classification

It is the intellectual power of a machine to accept the handwritten character & recognition it no matter what the

source is like the touch screen, papers, photograph or any other devices. It is a difficult task to recognize the hand written characters due to the different size boldness, rotation, format and resolution recognition could become very simple if the classifier being used does not better about the best data.

The result from the process of classification helps compare the rate of accuracy, training, testing time and classification time of new feature extraction technique with some of the existing techniques. The observation of the result indicates that the new approach presents better results of classification as compared to other techniques. If measured in the terms of recognition rate. The classifications that are popular for the handwritten character are back propagation neural network template matching, associative memory, support vector machine, KNN neighbours, Bayesian classification etc. back propagation neural network is used in this Paper. For the dataset having less number of classes of characters template matching. Artificial immune system and associative memory are used.

2 .CONCEPT OF FEATURE EXTRACTION

Feature extraction methods falls among these categories. Any method can be a combination of these categories also:

1. Statistical

2. Structural

3. Global transformations and moments

2.1 Statistical

Statistical methods based on a planning of how data collected and selected, which helps to make a hypothesis about the type of data. It is based on the probability theory and hypothesis. Statistical distribution of pixels of an image takes care of variations in writing styles. The main feature extraction methods under this category are:
1. Partitioning into regular or irregular regions
2. Profiles and Projections
3. Distances and Crossings

2.1.1Partitioning into regular or rregular regions.

The character image is divided into partitions called zones. Zone can be of regular or irregular size, overlapped or non-overlapped. From each zone such features are extracted which make distinct identity for each class. The features obtained from zoning are restricted to a particular region instead of whole region of image. Partitioning of an image is shown in figure 2.1.

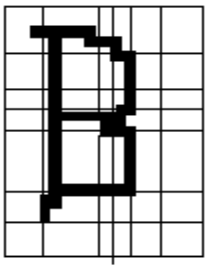

Fig. 2.1: Partitioning an Image into Zones

1. Zoning Density Features: The number of normalized or not normalized foreground pixels, in each zone is a feature.
2. Topology Features: The based on the shape of foreground pixel, features are calculated. Shape is found such that it uniquely identifies the character class.
3. Zoning Direction Features: Based on 3x3 matrix of neighbourhood pixels in each zone directional histogram is obtained. It depends on the outline of the image. For each zone the contour is followed and a directional histogram is obtained by analyzing the adjacent pixels in a 3x3 neighbourhood. Figure 2.2 given below shows the 8 direction of a pixel in an image.

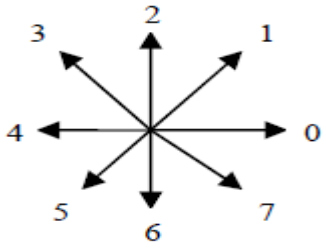

Figure 2.2: Directions of a Pixel.

2.1.2Profiles and projection histograms

Two-dimensional character images can be converted to the one-dimensional vector by making projections of the character images. Projection features are independent of noise, but dependent on rotation of image. Number of pixels in vertical and horizontal direction of a character image is counted to make projection histograms. Projection histograms used to make difference between p" and"q" and "m" and "n" type of characters. Distance between bounding box and the boundary of the character image in terms of pixels is profiles of image.

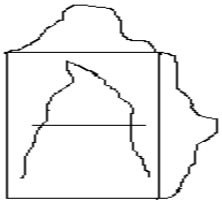

Figure 2.3: Projection of an Image

2.1.3Crossings and distances

Number of transitions from background to foreground pixels throughout the image along vertical line is vertical crossing and Number of transitions from background to foreground pixels throughout the image along horizontal line is horizontal

crossing Vertical distances are the distances between the first pixel from the top and bottom boundaries of the image. Horizontal distances are the distances between the first pixel from the left and right boundaries of the image. Figure 2.4 given below shows the crossings and distances in an image.

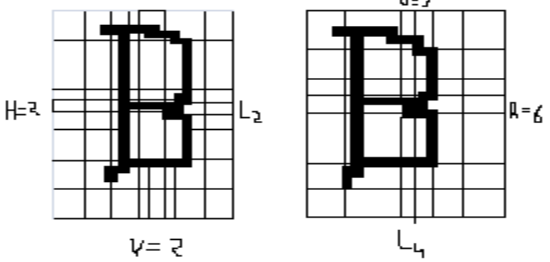

Figure 2.4: Crossings in an Image

2.2 Structural features

Structural feature space is extracted such that each value contains some information about structure of the image. Feature values are calculated from the structural and geometrical properties of the character. Examples of structural features are number of horizontal lines or vertical lines, aspect ratio, number of cross points, number of loops, number of branch points, number of strokes, horizontal curves at top or bottom, etc.

2.3 Global trans-formations moments

The Fourier Transformation of the outline of the image is extracted. First 'n' coefficients of the Fourier Transform can be used to reconstruct the outline of the image, so these n coefficient scan be taken as a feature vector of that particular character. Moments are used to recognize scale, translation, and rotation invariant of a character image. The original structure of the image can be rebuilt from the moment coefficients.

3.COMPARATIVE ANALYSIS OF FEATURE EXTRACTION METHOD

The main objective of a feature extraction technique is to accurately retrieve features from the character image. In this chapter we describe some methods to retrieve feature space. There are some feature extraction methods which we have implemented.

1. Creating a character matrix of binary values of image.
2. Region and pixel frequency based feature extraction methods-13-region, 15-region, 16-region, 25-region feature extraction methods.
3. Diagonal, Horizontal and Vertical line feature extraction methods.
4. Image centroid and zone centroid based distance metric feature extraction system.
5. Topology based feature extraction method.
6. Direction Based Features.
7. Cross-corners features
8. Distance and crossing based features
9. Hybrid methods

Table 3.1: Prons & Cons of different Feature Extraction Methods

S.no	Method	Merits	Demerits
1	Binary	Easy to implement and good for low resolution images	Training is slow for large size images, Redundant values in feature space
2	13-region	Feature space is small	Feature space is small
3	15-region	Information loss in feature Space	Attributes of different classes have little difference
4	16-region	Good accuracy	Zoning need complete understanding of character
5	25-region	less information lost	Redundancy in feature space
6	Diagonal	This gives more exact information	This is more specific to characters build by strainght lines
7	Vertical	More simple than diagonal	Range of values is large
8	Centroid	It can be merged with others to give good accuracy	It work good for only characters having maximum curves
9	Direction	Zoning does not required	Feature values are redundant
10	Topology	More accurate features	Large feature space
11	Cross-Corner	Substantial increase in accuracy	Feature space is large
12	Distance & Crossings	Accuracy is increased	Features are not related to each other
13	Hybrid	Testing is confirmed by Boosting	Selection of features is critical

4. CLASSIFIER AND OTHER TECHNIQUES

After getting feature space form the binary character image, an efficient classifier is used to classify the class of a character. The most traditional classifier used for handwritten character Recognition is Neural Network. Other than neural network there are a lot of classifiers for classification problem, for

example-Bayesian theory, support vector machine, K-nearest neighbor and artificial Immune system etc. Here I explaining following mostly used techniques for handwritten character classification:-

1. Back propagation Neural Network
2. Kohonen Network
3. Artificial Immune System

4.1 Back propagation neural network

Application of neural networks for handwritten character recognition is very common and important. Handwritten character recognition can be implemented by using a back propagation neural network that has been trained according to train dataset. During training, the network is trained to associate output characters with input characters. Trained network is used to identify the associated output character of input characters [2][5]. The critical phases for solving a problem using neural network are:

Phase I: The collection, preparation and analysis of the training data. Working of this phase is discussed in above.

Phase II: The design, training and testing of the neural network. This includes which Architecture and training method is used. Testing is to find the class of the input Character.

4.1.1 The architecture of back propagation neutral network:

During design of back rogation network shown in figure 4.1, we should keep in mind these questions:- what should be the numbers of layers and neurons, the activating functions, the initial weights, the choice of training algorithm and the training samples, the learning rate, momentum etc. The given points help us to find answers of these questions [6][1][7]. The design of input and output layers: The number of input layer units is the dimension of input samples. Dimension of input samples depends on the number of features of handwritten character sample. The number of output layer neurons depends on the number of possible classes in training dataset. So, number of neurons in output layer are 26 (26 uppercase English letters). The design of hidden layer: Number of hidden layers depends on application. In case of one hidden layer, the number of hidden layers' neurons can be calculated by the Following formula, but this is not a hard and fast formula:

$$\text{Hidden} = \sqrt{\text{Input2} + \text{Output2}}$$

Suppose our feature vector size is 25 and there are 26 alphabets of English to classify, Therefore we construct 25×26×26 back propagation network model. The choice of initialized weights: Before training, back propagation network must be Initialized because initial values are very significant to how soon the training achieve Local minimum. Weight of each neutron can adjust the biggest change between desired Output and actual output and their activation function. Initialized weight is generally get The random between -1 and 1. We choose 0.5 as initialized weight.

The choice of error function and activation function: During each iteration of training Process, the threshold error values is compared with the network may cause for degradation of performance in network, retaining some network capabilities as it is.

5.EXPERIMENTAL RESULTS AND ANALYSIS

In this chapter we discuss the recognition ability of different feature extraction methods on using Back propagation Neural Network. Environment used for simulating these methods is: Intel (R) core(TM) duo CPU, 2GB RAM.

5.1 Datasets

Two types of data sets are used: Dataset1 is taken from MATLAB dataset of Dillep Gaurav for handwritten character recognition in which first 15 sets are less noisy than remaining 10 characters. Dataset2 is build by writing characters by different writers on computer screen which have randomly noisy sets.

1 Description of dataset1: in this data set total numbers of samples are 650 BMP images of 8 bit pixel depth written by 25 different writers of each capital English alphabet (A-Z). 15 first set are less noisy compared to next 10 characters sets. The Dataset is partitioned into four group of training and test sets as mentioned in table 5.1.

2. Description of dataset2: in this Total numbers of samples are 650 also. 25 BMP images written by 25 different writers of each capital English alphabet (A-Z) are taken in this dataset which have noisy characters randomly. The Dataset is partitioned into four group of training and test sets as mentioned in table 5.1. TTD1, TTD2, TTD3 and TTD4 stands for training and test set1, training and test set2, training and test set and training and test set4 respectively.

Table 5.1: Train set and Test set Distribution

Distribution	Training	Testing
TTD1	1-15	16-25
TTD2	11-25	1-10
TTD3	1-20	21-25
TTD4	6-25	1-5

Table 5.2: Recognition rate using dataset1 and dataset2

Feature Extraction Method	Dataset1				Dataset2			
	TTD1	TTD2	TTD3	TTD4	TTD1	TTD2	TTD3	TTD4
Binary Feature	21	43	57	23	28	36	26	30
13-Region	16	37	20	21	66	65	59	50
15-Region	56	68	63	70	62	58	61	71
16-Region	16	32	26	42	62	60	57	78
25-Region	70	81	76	92	45	58	66	53
Diagonal	68	90	76	90	75	83	81	86
Vertical 25	60	76	76	80	48	60	79	55
Ext-Diagonal	53	41	67	87	75	76	76	58
Ext-Vertical	67	80	66	84	75	76	76	58
Geometry Based	56	70	63	65	45	59	76	50
Centroid	70	81	76	94	50	64	81	70
Direction	60	90	76	94	50	70	65	36
Cross-Corner	70	93	83	95	76	80	82	93
Distance & Crossings	69	90	80	94	77	78	86	90
Hybrid	71	72	80	94	72	74	86	33

5.2 Analysis:

Experiment's results on dataset1 and dataset2 using back propagation are mentioned in table 5.2. Cross-corner, diagonal, direction methods are most accurate methods according to the experimental results . Other methods can be combined with each other for making hybrid method to increase accuracy rate.

6. CONCLUSION & FUTURE WORK

Feature extraction is the most crucial & important part of handwritten character recognition. With feature extraction we also implemented steps of preprocessing to normalize an image of a character. We selected Back propagation neural network for classification purpose. Comparative analysis of different feature extraction methods in terms of accuracy is done in our work. Cross-corner, diagonal, direction methods are most accurate methods according to the confusion matrices & 13-region have least recognition rate. Combining different feature vector into a single feature vector for hybrid method & result shows that hybrid method have higher recognition rate compared to its individual feature extraction method in case of accuracy. In the future work we are going to introduce a hybrid approach for the feature extraction techniques with using the pros and cons of different feature extraction technique.

REFERENCES

[1] Velappa Ganapathy and Kok leong Liew. Handwritten character recognition using mutiscale neural network training technique. World Academy of Science, Engineering and Technology, 39:32–37, March 2008.

[2] Anita Pal & Dayashankar Singh. Handwritten english character recognition using neural network. *International Journal of Computer Science & Communications*, 1:141–144, July-December 2010.

[3] S.V. Rajashekararadhya and Dr. P. Vanaja Ranjan. Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south indian scripts. *Journal of Theoretical and Applied Information Technology*, pages 1171–1180, 2005-2008.

[4] J.Pradeep, E.Srinivasan, and S.Himavathi. Diagonal feature extraction based on handwritten character using neural network. *International Journal of Computer Applications*, 8:17–21, October 2010.

[5] Dong Xiao Ni. Application of neural network to character recognition. *Proceeding of Student/Faculty Research Day,CSIS, Pace University*, pages C4.1–C4.6, May 2007.

[6] Jamal Fathi Abu Hasana and Fakhraddinmamedov. Character recognition using neural netwok. *World Acedamy of Science, Engineering and Technology, Turkey*, August 2006.

[7] Xin Wang, Ting-Lei Huang, and Xiao yu Liu. Handwritten character recognition based on bp neural network. *Third International Conference on Genetic and Evolutionary Computing*, pages 520–524, 2009.

[8] Yuefeng Chen, Chunlin Liang, Lingxi Peng, and Xiuyu Zhong. Handwritten character recognition based on artificial immune systems. *International Conference on Computer Application and System Modeling*, 12:273–276, December 2010.

[9] S. Rajasekaran and G.A. Vijayalakshmi Pai. Associative memory. pages 87–116, 2008.