ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

HANDWRITTEN ENGLISH CHARACTER RECOGNITION USING LVQ AND KNN

Rasika R. Janrao [*], Mr. D. D. Dighe
[*] Department of Electronics and telecommunication, Matoshri Collage of Engineering, Nashik, India

DOI: 10.5281/zenodo.60830

ABSTRACT

A Handwritten character recognition (HCR) is an important task of detecting and recognizing in characters from the input digital image and convert it to other equivalent machine editable form. It gives high growth in image processing and pattern recognition. It has big challenges in data interpretation from language identification, bank cheques and conversion of any handwritten document into structural text form. Handwritten character recognition system uses a soft computing method like neural network, having area of research for long time with multiple theories and developed algorithm. Feature Extraction done in character recognition by introducing a new approach, diagonal based feature extraction. We used two Dataset, first one is own database of 26 alphabets, 10 numbers and 5 special characters written by various people and second is standard CEDAR database. The character recognition is carried out by supervised KNN classifier and LVQ. The results show that KNN has better results than LVQ.
.

KEYWORDS: Handwritten character recognition, KNN, LVQ

INTRODUCTION

Handwritten character recognition (HCR) is the process of conversion of scanned handwritten documents into the text document so that it becomes editable and researchable. It contains conversion of text image into letter codes which are useful in computer and text processing applications. All collecting databases from all different handwritten style samples is considered as static images of handwriting. So that the recognition is slightly problematic as different individuals have a dissimilar style of writing.

Sometimes a document written in the past used for recognition, then scanned image will need to be extracted for getting individual characters. Various tools exist for extraction, but may occur common imperfection in it. Commonly, when characters are connected, then both characters considered as single sub-image. Therefore there is a recognition problem but Yet there are many algorithms are available to minimize the problem related to conncted characters. After the extraction of individual characters occurs, a recognition system is used to find the equivalent computer character.

The handwritten character recognition is globally divided into two type i.e. Offline and Online character recognition.

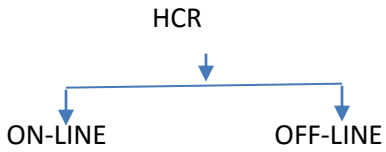

Figure 1:Classification of HCR methods

OFF-LINE: The writing is typically taken optically by a scanner and the completed writing is accessible as an image.

ON-LINE: It has the two dimensional organizes of sequential points are denoted as a function of time and the order of knocks prepared by the writers are also accessible.

In the research area, handwritten character recognition system is exploring with new techniques and improving performance accuracy. Recognition system used in application like document reading, mail sorting, postal address recognition and bank processing.

In HCR, (a) Pre-Processing, (b) Segmentation and (c) Feature Extraction these are the main stages. The pre-processing provides shape to input image in the form which is appropriate for segmentation. Segmentation process resized every segmented individual character in m×n pixel for training purpose. The feature extraction method selection is directly affect on the recognition accuracy. The various methods used for feature extraction, such as Fourier descriptor, zoning, gradient feature and Gabor feature.

In this paper, using a diagonal based feature extraction method in which all individual characters resized into 90×60 pixel and divided in 54 equal zone of 10×10 pixel. Feature extraction is done diagonally means from pixel of each zone feature moving with their diagonal.

LITERATURE SURVEY

Youssef Es Saady et. al. [1] proposed, the method to detect Amazigh handwriting recognition method. The method is based on horizontal and vertical centerline of the character. The characters are segmented into horizontal and vertical lines and position of the character is obtained according to those lines. The features are calculated using sliding window techniques. The characters are classified using MLP. The correction rate obtained was 99.28 % for the 19437 Amazigh printed characters and 96.32% for the 20150 Amazigh handwritten characters.

Dileep Kumar Patel et. al. [2] proposed the DWT based handwritten character recognition system along with Euclidean distance metrics. The classification of the testing vector is depend upon the Euclidean distance. Minimum disatace decide the class of the test vector. This system gives good accuracy of 90%.

Prasad P. Chaudhari et. al. [3], proposed a grid approach for recognition of an offline handwritten character using grid approach is proposed. Extracted features are train by neural network as classifier of the character in classification stage. The recognition system of experimental results shows that this technique is effective and reliable. The overall procedure results in recognition rate are 96.9%.

Reetika Verma et. al. [4], proposed a neural network alog with surf feature approach to solve complex character recognition issue. This approach has been evaluated using noise parameter. The evalution is performed by PSNR and MSE. The classification is done by using back propogation neural network. The success rate of this systm is 98.77%

Apash Roy et. al. [5], proposed a system with feed forward neural network having ability of a machine to interpret handwritten characters from sources like paper document, photograph etc. to digital computerized form is the aim of HCR systems. The neurons of output layer have a feedback connection from their output line. Experimental result shows that an efficient recognition. However, the proposed system is not a complete one. Some other techniques may be combined with this approach to increase the efficiency of the system. The work can be extended to recognize characters or numerals of some other languages also.

G V S S K R Naganjaneyulu et. al. [6] proposed a new algorithm to recognize the characters and alphabets from low resolution images. It uses low pass filter in L2 space which slightly improve the performance. The results show that the performance of OCR on low pass filtered images in WSS is far better than the other two cases as images are low resolution. The result shows that OCR giving better performance on the images which are low pass filtered in Weighted Sobolov space.

ISSN: 2277-9655
Impact Factor: 4.116

Hassan El Bahi et. al. [7] Present an offline English HCR system based on phone camera capture images. Adaptive thresholding method is used to get meaningful information. Further the features were extracted using combination of GLCM, Zernike Moments, Gabor Filters, Zoning, Projection Histogram and Distance Profile. Finally classification is performed by three classifiers i.e. SVM, Naive Bayes and multilayer perceptron (MLP). The database was created containing 3380 samples collected from different writers.

J. Pradeep et. al. [8], proposed HCR system based on neural network. The feature were extracted by using diagonal lines. Total 54 features were extracted by resizing image into 9x6 pixel. The system is tested on 570 different samples of 26 alphabates. This system gives better accuracy than horizontal and vertical method of feature extraction.

PROPOSED METHOD

The proposed handwritten character recognition (HCR) system has different steps. The detailed blocks are explained below. The details of each step are described below. Left side of the block diagram is a training process while right hand side is the testing procedure of the proposed system.

1. Database

Collection of Handwritten character image is the first task for HCR system. In proposed approach, the alphabate, numeric and special characters of English language are collected from two database i.e. one from standard CEDAR database [9] and other from our own created database. The database consist of 26 alphabates, 10 numeral and five special characters (@, γ, #, α, &). The sample of database image is shown in below fig.3

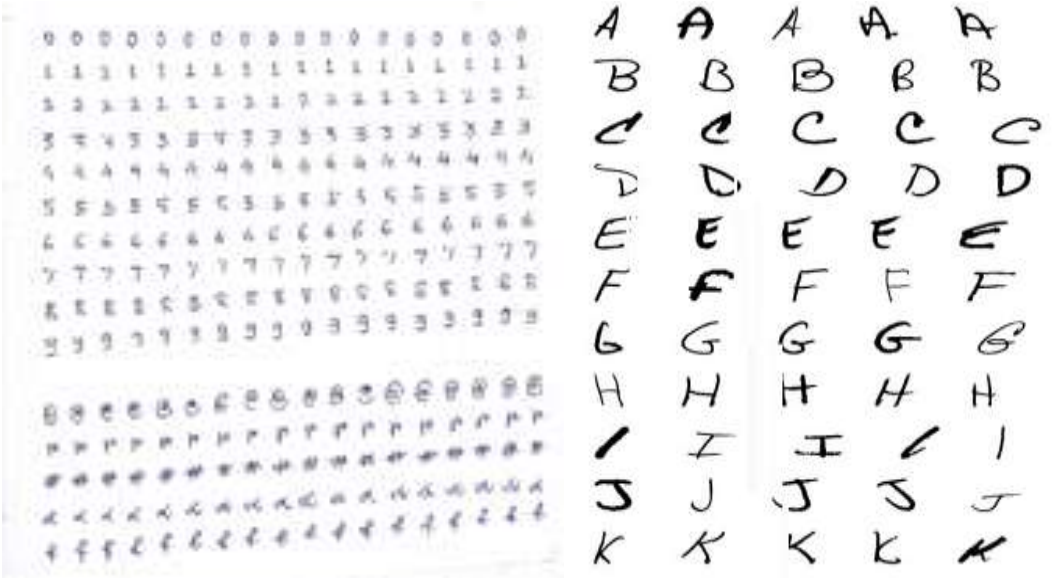

Figure 2: Sample image from (a) Own database (b) CEDAR databasse

2. Preprocessing

The pre-processing is a sequence of processes completed on the scanned input image. It basically improves the image interpretation which is suitable for segmentation. The various tasks performed on the image in pre-processing stage are shown in Fig.2. The input image is in RGB, covert into grayscale. The noise from the image is reduced by using median filter[10].

ISSN: 2277-9655
Impact Factor: 4.116

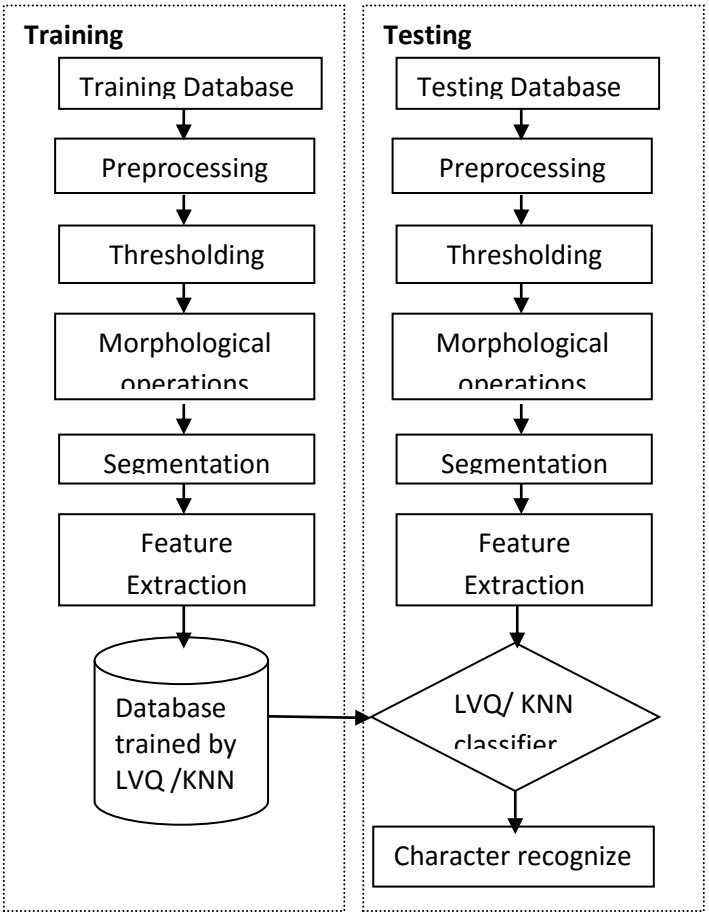

Figure 3: Block Diagram of proposed system

Conversion of an image i.e. from Gray scale into a Binary is called Binarization. The local thresholding method is applied on the grayscale image as

$$g_{(x,y)} = \begin{cases} 1 & f(x,y) > Th \\ 0 & \text{otherwise} \end{cases} \quad \ldots\ldots\ldots (1)$$

Where, g(x,y) is binary image pixel, f(x,y) is gray image pixel, Th is the local threshold value

The binary image is enhance by using morphological filters like erosion and dilation. Erosion is a morphological operation which removes growing part while dilation fills the gaps within the object.

3. Segmentation

An image breaks down into the sub-images of individual character called segmentation process. In our proposed system, the pre-processed input image is divided into sequestered characters by transferring a number to each character using a marking procedure. This labeling provides information about the number of characters in the image. Each individual character is equivalently resized into 90X60 pixels for cataloguing and recognition phase.

ISSN: 2277-9655
Impact Factor: 4.116

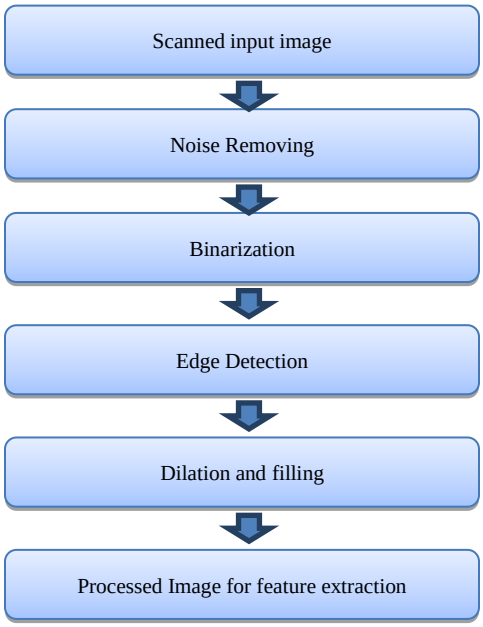

Figure 4:Pre-processing of handwritten character

Our purpose is to take handwritten English characters as input, process the character, train the NN system, to detect the shape and modify the character to a beautified form of the input.

The propose algorithm is expected at emerging software which will be supportive in identifying characters of verbal English language. As well as, it is also helpful in recognizing special characters. The one of the drawbacks is that it is restricted to only English characters and numerals. In a further development to distinguish the characters of dissimilar verbal languages. It can overcomes the concept of NN (neural network).

NNs are mainly useful for resolving difficulties that cannot be stated as a sequence of phases, such as identifying patterns, classifying them into sets, sequence prediction and data mining. The NN is presented with a target vector and also a vector which contains the pattern information, this could be an image and handwritten data. The neural network then tries to regulate if the input records contests a pattern that the neural network has remembered. ANN trained for classification is planned to take input models and categorize them into sets. These sets may be fuzzy, without clearly defined boundaries. This methods concerns detecting free handwritten characters.

4. Feature extraction

In this stage, the features of the characters that are critical for organizing them at recognition stage are removed. It is a key step such as its recent functioning improves the recognition rate and decreases the misclassification [11] . We proposed diagonal feature extraction structure for recognizing off-line HCR. Each and every character image of size 90x60 pixels is separated into 54 equal zones, each of size 10x10 pixels Fig.3. the feature are calculated from each 10x10 zone. Each zone has total 19 diagonal lines so from one zone we get 19 features. These 19 features are average and stored into each zone as shown in fig. 5. This process will continue till the 54 zones. As a result each handwritten character can be represent by 54 features.

ISSN: 2277-9655
Impact Factor: 4.116

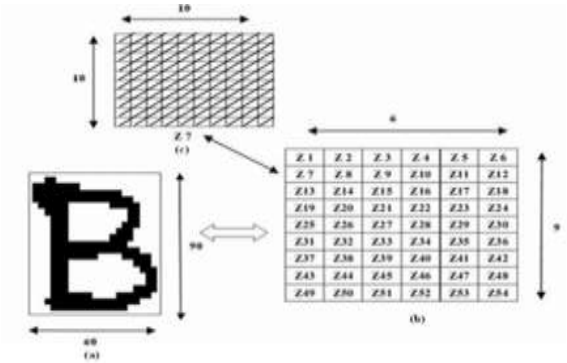

Figure 5:Procedure for extracting feature from the characters

CLASSIFICATION AND RECOGNITION

extracted features were trained with two type of classifier i.e. LVQ and KNN. Both the classifiers are explained below

A. LVQ

Linear vector quantization (LVQ) is the artificial neural network algorithm. The main aim of LVQ is to convert large set of input data into smaaler set of code book vectors that provide better approximation to the original iunput space. LVQ has adavanatage of dimewntionality reduction over other. The error of LVQ is given by

$$D = \sum_x |x - w_{I(x)}|^2 \quad \ldots\ldots\ldots(2)$$

Where, x is the input vector and there reperesentative is $w_{I(x)}$

B. KNN

Knn is a simplest supervised learning algorithm used to classify the data. The k-nearest neighbor algorithm (KNN) is a method for classifying object by majority of votes. The value of k decide the search space. If k=1 means the object is simply assign the class which is nearest. The advantage of the Knn is its classification accuracy is better when feature space are small but which feature space is large then its accuracy get decrease.

RESULTS

The system has been implemented using MATLAB 2013b. We made a database of handwritten character of different people. The scanned image is taken as input from database. The input images are in RGB format. First we convert that image into gray scale as shown below. To detect character, edge is the important factor so we used canny edge detection algorithm. Edge detection followed by morphological operation such as dilation, erosion and media filtering to minimize the noise and extract the proper character. The results of proposed preprocessing step is show below-

The results of the proposed system is analyse on the basis of two methods viz. qualitative and quantitative analysis

A. Qualitative analysis

Qualitative data analysis is non-statistical; its methodological approach. It is a shown in pictorial form. The input images from the own database and standard cedar database are shown below

ISSN: 2277-9655
Impact Factor: 4.116

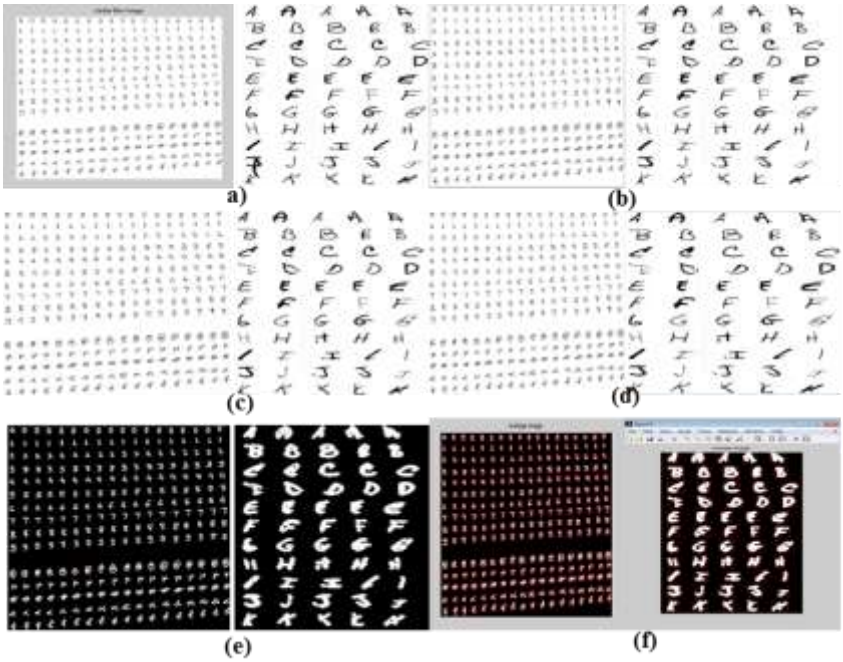

Figure 6:Qualitative analysis : (a) Input Image (b) Gray image (c) Median filter (d) binary image (e) Enhance binary Image (f) extracted characters

The extracted features were fed to the Neural network for classification. In the proposed algorithm, we used linear vector quantization and K-nearest neighbor classifier to classify the data.

B. Quantitative analysis

Quantitative research is the systematic empirical investigation of observable phenomena via statistical, mathematical or computational techniques. In this proposed approach we do quantitative analysis by using accuracy of character detection.

The quantitative analysis of proposed system is done by using accuracy parameter. The accuracy of the proposed system using LVQ and KNN on own and CEDAR database is analysed below.

Table1. Accuracy analysis

classifier	CEDAR	OWN
LVQ	**77.80%** (net = 100 & epoch 200)	**88.29%** (net 100, epoch 150)
KNN	**100%** (k= 2)	**93.65%** (k =2)

ISSN: 2277-9655
Impact Factor: 4.116

The Graphical representation of the proposed system is shown below. The blue color bar show the classification accuracy of the linear vector quantization while Red color bar shows accuracy of k-nearest neighbor classifier.

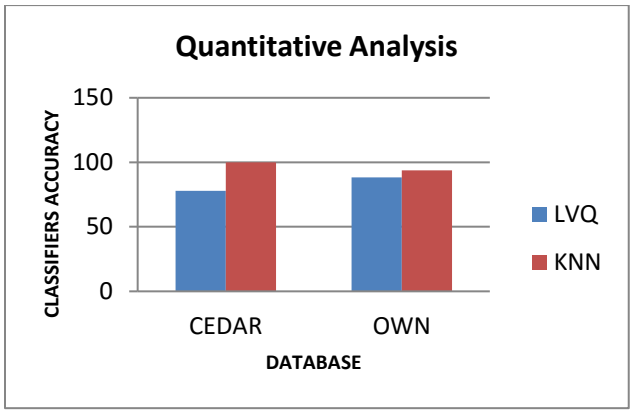

Figure 7:Graphical analysis of LVQ and Knn.

The graphical analysis shows that KNN classifier shows good accuracy on both the database.

CONCLUSION

The new approach of diagonal feature extraction is proposed. Total 54 features were extracted by diagonal based feature extraction approach. For classification purpose neural netwok based Lvq and Knn classifier are used. From proposed approach Knn gives appropriate accuracy for Standard database (CEDAR) and self generated handwritten database.

Table 2. Accuracy Evaluation result of CEDAR database

Authors	Feature extraction	Classifier	% Accuracy
Singh & Hewitt(2009)	Hough Transform	LDA	67.3
		Nearest neighbor	63.5
Varnvakar et al.(2009)	Structural feature	SVM	63.5
Gauri Katiyar and Shabana Mehfuz(2016)	Hybrid feature	MLP neural network	94.65
Proposed method	**Diagonal feature extraction**	**LVQ**	**77.80**
		KNN	**100**

Table shows accuracy evaluation result for CEDAR database. The proposed system performs diagonal based feature extraction using LVQ and KNN with 100% accuracy.

REFERENCES

[1] Youssef Es Saady, Ali Rachidi, Mostafa El Yassa, Driss Mammass, "Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character", International Journal of Advanced Science and Technology Vol. 33, August, 2011, pp 33-50.

[2] Dileep Kumar Patel, Tanmoy Som, Sushil Kumar Yadav, Manoj Kumar Singh, "Handwritten Character Recognition Using Multiresolution Technique and Euclidean Distance Metric", Journal of Signal and Information Processing, 2012, 3, 208-214

[3] Prasad. P. Chaudhari, K. R. Sarode, "Offline Handwritten Character Recognition by using Grid approach", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 3, Issue 4, April 2014 ISSN 2319 – 4847

[4] Reetika Verma, Rupinder Kaur, "An Efficient Technique for character recognition Using Neural Network & Surf Feature Extraction", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1995-1997, ISSN 0975 – 9646

[5] Apash Roy and N. R. Manna, "Handwritten Character Recognition with Feedback Neural Network", International Journal of Computer Science & Engineering Technology (IJCSET), ISSN: 2229-3345 Vol. 5 No. 01 Jan 2014.

[6] G V S S K R Naganjaneyulu, A.V. Narasimhadhan, K. Venkatesh, "Performance evaluation of OCR on poor resolution text document images using different preprocessing steps", TENCON 2014, 2014 IEEE Region 10th Conference, 2159-3442, pp 1- 4.

[7] Hassan El Bahi, Zouhir Mahani and Abdelkarim Zatni, "An offline handwritten character recognition system for image obtained by camera phone", Mathematical Applications in Modern Science, ISBN: 978-1-61804-258-3, pp 180-189.

[8] J. Pradeep, E. Srinivasan and S. Himavathi, "Diagonal Based Feature Extraction for Handwritten Alphabets Recognition System Using Neural Network", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, Feb 2011

[9] S. Singh and M. Hewitt, "Cursive digit and character recognition in CEDAR database," Pattern Recognition, 2000. Proceedings. 15th International Conference on, Barcelona, 2000, pp. 569-572 vol.2. doi: 10.1109/ICPR.2000.906138

[10] Rafael C. Gonzalez, Richard E. woods and Steven L.Eddins, "Digital Image Processing using MATLAB", Pearson Education, Dorling Kindersley, South Asia, 2004.

[11] S.V. Rajashekararadhya, and P.VanajaRanjan, "Efficient zone based feature extraction algorithm for handwritten numeral recognition of four popular south-Indian scripts," Journal of Theoretical and Applied Information Technology, JATIT vol.4, no.12, pp.1171-1181, 2008.