Offline English Hand Written Character Recognition Using Neural Network

Vijay Laxmi Sahu and Babita Kubde

Abstract : Image processing and pattern recognition plays a lead role in handwritten character recognition. The recognition of handwriting can, however, still be considered an open research problem due to its substantial variation in appearance .There are four main steps of handwritten character recognition- Data collection and pre-processing, segmentation feature extraction and classification. The main objective of this research is to find a new solution for handwritten text recognition of different fonts and styles by improving the design structure of the feature extraction. The main aim of this paper is to propose a fast and easy to use feature extraction method that obtains a good performance. This study focuses on isolated characters. Diagonal feature extraction scheme for recognizing off-line handwritten characters is proposed in this work in addition efficient feature such as Eigen value and mean value are also used which improves accuracy to recognize character. Diagonal features are playing an important role in order to achieve higher accuracy of the recognition system. It describes recent achievements, difficulties, successes and challenges in all aspects of handwriting recognition. It also presents a new approach which dramatically improves current handwriting recognition systems.

Keywords : Image Processing, Pattern recognition, Feature eztraction

I. Introduction

Handwriting recognition has been one of the most fascinating and challenging research areas in field of image processing and pattern recognition in the recent years. It contributes immensely to the advancement of automation process and can improve the interface between man and machine in numerous applications. Several research works have been focusing on new techniques and methods that would reduce the processing time while providing higher recognition accuracy.Off-line character recognition is known as Optical Character Recognition (OCR), because the image of writing is converted into bit pattern by an optically digitizing device such as optical scanner or camera. The recognition is done on this bit pattern data for machine-printed or hand-written text. The research and development is well progressed for the recognition of the machine-printed documents. In recent years, the focus of attention is shifted towards the recognition of hand-written script.

Vijay Laxmi Sahui is [1]M tech Student and Babita Kubde is working as Professor, RCET Bhilai, Vijaylaxmibit1987@gmail.com

In general, handwriting recognition is classified into two types as off-line and on-line handwriting recognition methods. In the off-line recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image. But, in the on-line system, the two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available. However, in the off-line systems, the neural networks have been successfully used to yield comparably high recognition accuracy levels. Several applications including mail sorting, bank processing, document reading and postal address recognition require off-line handwriting recognition systems.[1][2].

II. Literature Suryev

The basic idea behind doing a literature survey is to gain knowledge regarding the related work. As was in our case, a lot of research paper were taken into consideration and studied. The basic steps involved in character recognition were pre-processing, character segmentation followed by designing an efficient classifier. In the last four decades, handwriting recognition has been a very active area of research. Writing, which has been the most natural mode of collecting, storing and transmitting the information through the centuries, now serves not only for the communication among humans, but also, serves for the communication of humans and machines. Before we go on to describe the problem at hand and our approach to it, a brief introduction to the various techniques we used is needed. U. Pal et al [3] have proposed zoning and directional chain code features and considered a feature vector of length 100 for handwritten numeral recognition, achieved reasonably high accuracy, but the time complexity of their algorithm is more. In this paper [4], a diagonal feature extraction scheme for the recognizing off-line handwritten characters is proposed. In the feature extraction process, each individual character resized 90x60 pixels and again divided into 54 equal zones which each of size 10x10 pixels. The relevant features are extracted from the pixels of each zone by moving along their diagonals of each zone. These extracted relevant features are used to train a feed forward back propagation neural network for performing classification and recognition tasks. Extensive simulation studies show that the recognition system using diagonal features provides good recognition accuracy & less time for training. Bikash et al. [5] proposed a continuous density HMM to recognize a word image. The histogram of chain-code directions in the image script, scanned from left to right by a sliding window, is used as the feature vector. A handwritten word image is assumed to be a string of several image frame primitives. One HMM is constructed for each word. To classify an unknown word image, its class conditional probability for

each HMM is computed. The class that gives highest probability is finally selected.

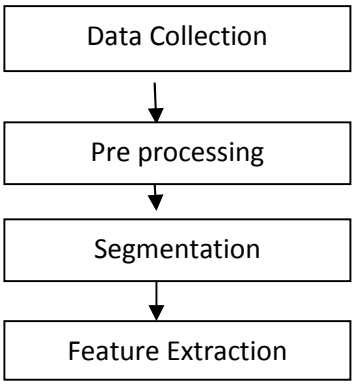
Recently Del Bimbo et al [6] proposed to use deformable templates for character recognition in gray scale images of credit card slips with poor print quality. The templates used were character skeletons .It is not clear how the initial positions in the image were to be tried, then the computational time would be prohibitive. Dinesh Acharya et al [7] have used the 10-segment string, water reservoir, horizontal/vertical strokes, and end point as the potential features for recognition and have reported the recognition accuracy of 90.50%. Drawback of this procedure is that, it is not free from thinning.

T.P. Singh et al [8] presented an effort to compare the performance for pattern recognition with conventional hebbian learning rule and with evolutionary algorithm in Hopfield model of feed forward network. The storing of the object has been performed using hebbian rule and recalling of these stored pattern on presentation of proto-type input pattern has been used by using both convolution hebbian rule and evolutionary algorithm. In this paper [9].

A handwritten Kannada and English Character recognition system based on spatial features is presented. Directional spatial features via stroke length, stroke density and the number of stokes are employed as potential & relevant features to characterize the handwritten Kannada numerals/vowels and English uppercase alphabets. KNN classifier is used to classify the characters based on these features with four fold cross validation. The proposed system achieves the recognition accuracy as 96.2%, 90.1% and 91.04% for handwritten Kannada numerals, vowels and English uppercase alphabets respectively. Dinesh et al [7] have used horizontal/vertical strokes, and end points as the potential features for recognition and reported a recognition accuracy of 90.50% for handwritten Kannada numerals. However, this method uses the thinning process which results in the loss of features .

III. Proposed Methodology

A typical handwriting recognition system consists of pre-processing, segmentation, feature extraction, classification and recognition stages.



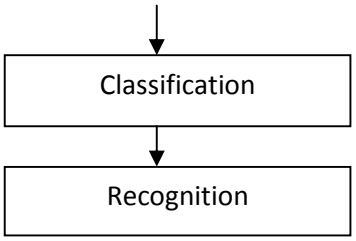


Figure 1. Stages in OCR

A. Data Collection

The database for handwritten compound characters is created by scanning the characters at scanner. The images are stored in bmp file format..To filter the raw images we designed and implemented a program that aligns the characters and separates them from each other. This is needed as the network only classifies one character at a time. To increase the quality of the image we reduced the image noise and increased the contrast. The image is also turned in to a black and white representation and scaled to a fixed size. This is done to make it suitable to feed in to the network. Each data set consists of isolated characters. Preprocessing involves series of operations performed to enhance to make it suitable for segmentation. filtering involves noise removal generated during document generation. Proper filter like mean filter, min-max filter, Gaussian filter, median filter etc may be applied to remove noise from document.

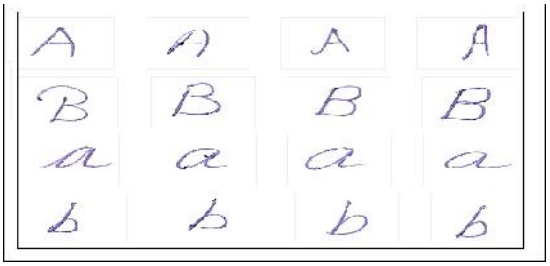


Figure 2. Sample Dataset

B. Pre Processing

Data Pre processing describe any type of processing performed on raw data it for another processing procedure. The main task in pre processing the captured data is to decrease the variation that causes a reduction in recognition rate. The various tasks performed on the image in pre-processing stage are as follows:

Noise Reduction- When the document is scanned, the scanned images might be contaminated by additive noise and these low quality images will affect the next step of document processing. Therefore, a pre-processing step is required to improve the quality of images before sending them to subsequent stages of document processing. Due to the noise there can be the disconnected line segment , large gaps between the lines etc. so it is very essential to remove all of these errors so that's the information can be retrieved in the best way. There are many kinds of noise in images. One additive noise called "Salt and Pepper Noise", the black points and white points sprinkled all over an image, typically looks like salt and pepper, which can be found in almost all documents. Noise reduction techniques can be categorized in two major groups as filtering, morphological operations.

Binarization - Binarization of gray-scale character images is a crucial step in offline character recognition. Good binarization facilitates segmentation and recognition of characters. Binarization process converts a gray scale image into a binary image.

Edge Detection-Edges characterize object boundaries and are therefore useful for segmentation, registration, and identification of objects. Edge detecting an image significantly reduces the amount of data and filters out useless information, while preserving the important structural properties in an image.

C. Segmentation

Image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain visual characteristics. In Character Recognition techniques, the Segmentation is the most important process. Segmentation is done to make the separation between the individual characters of an image. The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries in images. Script segmentation is done by executing the following operations: Line segmentation, Word segmentation and character segmentation[10]

D. Feature Extraction

Feature extraction is the process to retrieve the most important data from the raw data. The major goal of feature extraction is to extract a set of features, which maximizes the recognition rate with the least amount of elements. In feature extraction stage each character is represented as a feature vector, which becomes its identity. Diagonal feature extraction scheme for recognizing off-line handwritten characters is proposed in this work. Every character image of size 100x 100 pixels is divided into 100 equal zones, each of size 10x10 pixels (Fig.3). The features are extracted from each zone pixels by moving along the diagonals of its respective 10x10 pixels. Each zone has 19 diagonal lines and the foreground pixels present along each diagonal line is summed to get a single sub-feature. Thus 19 sub features are obtained from the each zone. These 19 sub features values are averaged to form a single feature value and placed in the corresponding zone. This procedure is sequentially repeated for the all the zones. There could be some zones whose diagonals are empty of foreground pixels. The feature values corresponding to these zones are zero. Finally, 100 features are extracted for each character by averaged value. In addition, We calculate mean value of all the rows in each zone. We are getting 10 mean values from 10 rows for each selected zone. After that we are selecting the minimum of the mean value. Similarly we are calculating the same for the each zone. From this we are getting 300 features(100 feature from average ,100 feature from mean and 100 feature from eigen value) . Now we are assigning the prime number for the above features. For average as a feature, we are assigning prime no. from 1,3,5,7,11,13… for 100 feature, for the non zero average values , we are adding the assigned prime number and again getting a single average value. Again we are assigning the prime number for the Eigen value as a feature. For Eigen value as a feature, we are again assigning prime no. from 1,3,5,7,11,13… for 100 feature, and for the non zero Eigen values , we are adding the assigned prime number and again getting a single Eigen value. Similarly the same step will be followed to find a single feature for mean value. These extracted features are used to train a feed forward back propagation neural network for performing classification and recognition tasks.

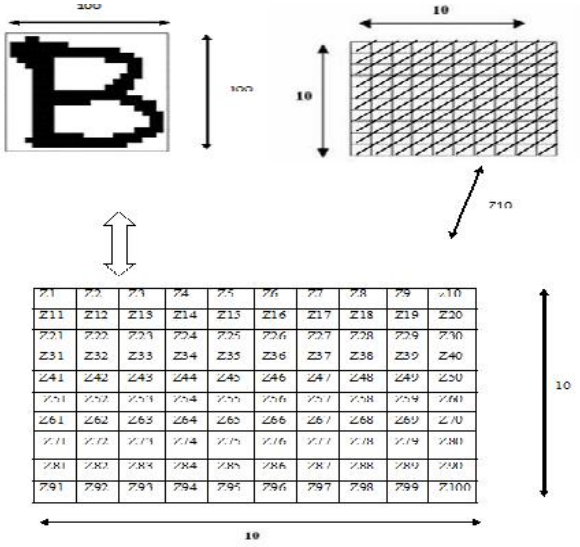


Figure 3 Diagonal Feature extraction

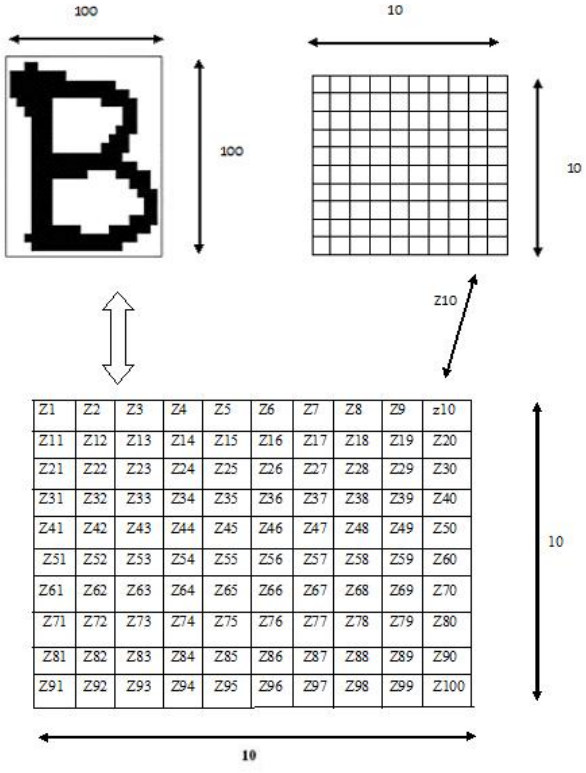


Figure 4 Row based Feature extraction Eigen and mean Value

Classification

A feed forward back propagation neural network having two hidden layers with architecture of 69-100-100-26 is used to

perform the classification. The hidden layers use log sigmoid activation function, and the output layer is a competitive layer, as one of the characters is to be identified. The feature vector is denoted as p where p = (x1,x2,x3,.....xd) where x denotes features and d is the number of zones into which each character is divided. The number of input neurons is determined by length of the feature vector d. The total numbers of characters t determines the number of neurons in the output layer [11].
The network training parameters are: Input nodes: 69 Hidden nodes: 100 each Output nodes: 26 (alphabets) Training Algorithm: Gradient Descent with Momentum Training and Adaptive Learning Perform function: Mean Square Error Training Goal Achieved: 0.000001 Training Epochs: 100000 Training Momentum Constant: 0.9 Training Learning Rate: 0.01

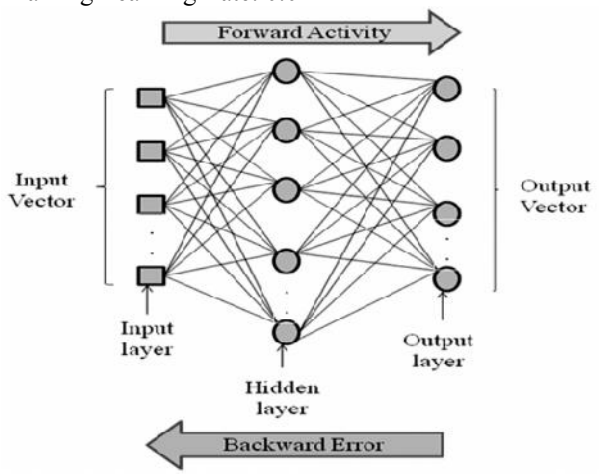


Figure 5: Back-propagation Neural Network

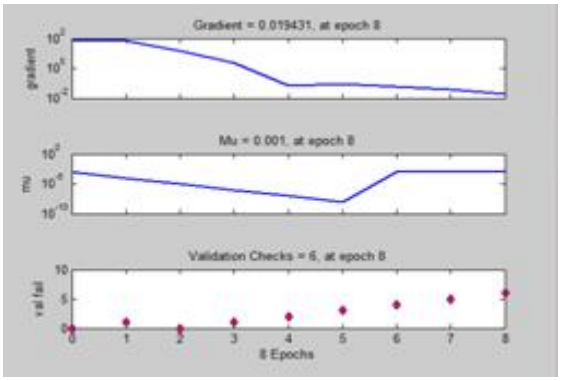


Figure 6:Training States

4. Result and Conclusion

In this paper an OCR system is proposed for English characters using artificial neural network. Recognition system has been implemented using Mat lab. A improved type of feature extraction, namely, an efficient feature extraction method is proposed which can give high recognition accuracy while requiring less time for training and classification both as compare to the diagonal based feature extraction method which gives upto 98.8% accuracy requiring less time for training which is better than previous one. So, using the Eigen value and mean value which is good in categorizing the alphabets into different groups with zone based feature extraction method will be effective process increasing the accuracy and speed.

Fig. 7 Handwritten character recognition system in graphical user interface

Table 1 Recognition Accuracy for different Character

Character	Recognition Accuracy	Character	Recognition Accuracy
A	98%	**N**	98.60%
B	97.50%	**O**	97.80%
C	98.50%	**P**	98.60%
D	98%	**Q**	98.40%
E	99%	**R**	98.50%
F	98.50%	**S**	98.30%
G	98.30%	**T**	98.40%
H	98.40%	**U**	98.50%
I	98.30%	**V**	97.30%
J	98.80%	**W**	98.50%
K	97.90%	**X**	98.50%.
L	96.90%	**Y**	99.30%
M	97.30%	**Z**	98.30%

10. References

[1] Anita Jindal, Renu Dhir, Rajneesh Rani "Diagonal Features and SVM Classifier for Handwritten Gurumukhi Character Recognition," Volume 2, Issue 5, May 2012

ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[2] N. Arica and F. Yarman-Vural, ―An Overview of Character Recognition Focused on Off-line Handwriting", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol.31 no.2, pp. 216 - 233. 2001.

[3] N. Sharma, U. Pal, F. Kimura, "Recognition of Handwritten Kannada Numerals", 9th International Conference on Information Technology (ICIT'06), ICIT, pp. 133-136

[4] J.Pradeep, E.Srinivasan, S.Himavathi "Diagonal Based Feature Extraction For Handwritten Character Recognition System Using Neural Network".

[5] Bikash Shaw, Swapan Kumar Parui, Malayappan Sridhar, 2008, "Offline handwritten Devanagari word recognition: A holistic approach based on directional chain code feature and HMM" IEEE.

[6] A.D. Bimbo, S. Santin, and J. Sanz, "OCR from poor quality images by deformation of elastic templates," in proceedings of 12th IAPR Int. Conf. pattern Recognition, vol.2, pp.433-435, 1994

[7] Dinesh Acharya U, N V Subba Reddy and Krishnamurthy, "Isolated handwritten Kannada numeral recognition using structural feature and K-means cluster", IISN-2007, pp-125 -129

[8] T. Kohonen, The self-organizing map, *Proc. IEEE*, 78(9): 1464-1480, 1990

[9] Velappa Ganapathy, and Kok Leong Liew ,Handwritten Character Recognition Using Multiscale Neural Network Training Technique, World Academy of Science, Engineering and Technology 39 2008

[10] Arif Billah Al-Mahmud Abdullah and Mumit Khan "a survey on script segmentation for bangla ocr" Working Papers 2004-2007

[11] U. Pal, T. Wakabayashi and F. Kimura, ―Handwritten numeral recognition of six popular scripts,‖ Ninth International conference on Document Analysis and Recognition ICDAR 07, Vol.2, pp.749-753, 2007