Handwriting Recognition System using Diagonal Feature Extraction

Saiyam Jain[1] Latika Kharb[2]
[1]Student [2]Professor
[1,2]Department of Masters in Computer Applications
[1,2]Jagan Institute of Management Studies, Delhi, India

Abstract— Handwriting recognition system plays a very important role in today's world. Handwriting recognition is very popular and computationally expensive work. At present time it is very difficult to find correct meaning of handwritten documents. There are many areas where we need to recognize the words, alphabets and digit. There are many application postal addresses, bank cheque where we need to recognize handwriting. There are basically two different types of handwriting recognition system online and offline handwriting recognition. There are many approaches are present for offline handwriting recognition system. Handwriting recognition system can be used to solve many complex problems and can make human's work easy.

Keywords: Handwriting Recognition System, Diagonal Feature Extraction

I. What is Handwriting Recognition System?

Handwriting recognition is an ability and technique of the system which receive the input from touch screen, electronic pen, scanner, images and paper documents. Offline handwriting recognition system is an art of identifying the word from images. As we know every person has their different writing style, so it is very difficult to recognize the correct handwritten characters and digits. Handwriting recognition system is developed to achieve the accuracy and reliable performance. So handwriting recognition is most challenging area if image and pattern recognition. Handwriting recognition is very useful in real world. There are many practical problems where handwriting recognition system is very useful like documentation analysis, mailing address interpretation, bank check processing, signature verification, postal addresses. Some recognition system identify strokes, other apply recognition on single character or entire words. So handwriting recognition system is work as a communication medium between human and machines.

II. History

Optical character recognition system has been studied in last many decades. In 1914 Emanuel Goldberg developed a system that read handwritten character and digits and converted then into a telegraph code. At the same time Edmund Fournier d'Albe developed the Optophon, a handheld scanner that scan the printed page and produced the output. Goldberg continued to develop Handwriting recognition system for data entry.

A. Hand-Printed English Character Recognition based on Fuzzy Theory

This paper (Puttipong Mahasukhon, Hossein Mousavinezhad, Jeong-Young Song, 2012) recognize the handwritten English character using fuzzy theory. This technique has two functions feature extraction and pattern recognition. Every character contains different deviations for e.g. position, size and shape due to different writers. In thin paper recognition is tested on 26 lowercase handwritten English characters and each character is stored in binary bitmap image.

B. Pre Processing

In this step the document scanning are performed. There are number of tools are present to perform this step.

C. Feature Extraction

The recognition rate mainly depends on the feature of the character. Feature can be structural, topological and geometrical. Geometrical features like angle and distance are explored in order to achieve high recognition rate. The properties that need to determine in fuzzy theory are number of circles, number of strokes and numbers of dots. International Journal of Computer Applications (0975 – 8887) Volume 114 – No. 19, March 2015

D. Character Recognition

After feature extraction a recognition system is used to identify the corresponding character. Here we used fuzzy theory to recognize the character. In this method a variation is deal with the membership function which is composed by x, y coordinates and length of the segment of the character. And then degree to compare segments of input to segment in database is calculated. There four steps are present for handwritten English character. In first step is read an input data from the handwritten documents. These characters are saved in binary bitmap format. The input image resolution should be 100x100 pixels to make sure that the character is not too big and small. In second step extraction of features are present. There are six segments are present, which consist dot(D), circle(C), rightincline line (RI), left-incline line (LI), up-down line(UD), leftright line (LR) as feature parameters. In third step matching the target image with the training images is performed and calculates the degree of similarity is calculated using Min-Max operation. For calculating the degree of similarity we have to fuzzify each input segment. In the last step system represent the result in a printed character.

E. Neural Network based Handwritten Character Recognition system without feature extraction:

A neural based offline handwriting recognition system without feature extraction is developed. In this method each character is resized into 30x20 pixels and after that using feed forward back propagation neural network is used to train the pixels.

F. Image Acquisition

The recognition system accepts a scanned image as an input. The images can be in JPEG, BMT format.

G. Pre-processing

In pre-processing stage various operation are performed like on image like binarization, noise removing, and edge detection.

H. Segmentation

In segmentation stage a sequence of character is segmented into sub-image of individual character. Each character is resized into 30x20 pixels.

I. Classification and Recognition

This stage is the decision making stage of the recognition system. The classifier contains of two hidden layers are present. The hidden layers used log sigmoid activation function to train the data. The number of neurons in the output layer is 26.

J. Post- processing

Post processing is last stage of recognition system. It prints the actual output after recognition after calculating equivalent ASCII value.

III. PROPOSED FEATURE EXTRACTION METHOD

In this stage, the features of the characters that are crucial for classifying them at recognition stage are extracted. This is an important stage as its effective functioning improves the recognition rate and reduces the misclassification. Diagonal feature extraction scheme for recognizing off-line handwritten characters is proposed in this work. Every character image of size 90x 60 pixels is divided into 54 equal zones, each of size 10x10 pixels. The features are extracted from each zone pixels by moving along the diagonals of its respective 10X10 pixels. Each zone has19 diagonal lines and the foreground pixels present long 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, 54 features are extracted for each character. In addition, 9 and 6 features are obtained by averaging the values placed in zones row wise and column wise, respectively. As result, every character is represented by 69, that is, 54 +15 features.

A. Combining Multiple Classifiers

The results of the different classifiers may be combined to obtain better classification accuracy. The results can be combined at different stages in the classification process. We have used a confidence level fusion technique where each classifier generates a confidence score for each of the six scripts. The confidence score is a number in the range ½0; 1, where 0 indicates that the test pattern is least likely to be of the script associated with the score, while a confidence score of 1 indicates that the test pattern is most likely to be the corresponding script. The confidence scores generated by the individual classifiers are summed and normalized to the range ½0; 1 to generate the final confidence score. The script which has the highest score is selected as the true class. In our experiments, we combined the results obtained from SVM, KNN (k = 5), and neural network classifiers. For SVM classifier, the confidence was generated from the output of the individual (two class) classifiers. The confidence score for the KNN classifier was computed as the proportion of neighbors which belong to the decided class. The output value of the node corresponding to the decided class gives the confidence value for the neural net-based classifier. The combined classifier could attain an accuracy of 87.1 percent on 5-fold cross-validation. The standard deviation of error over the cross-validation runs was 0.3 percent. Gives the confusion matrix for the combined script classifier which discriminates individual text lines

IV. FEATURE SELECTION

An interesting question to ask in any classification system is, how good are the available features, for the purpose of classification. A very effective, although suboptimal, way to determine the best subset of features for classification, given a particular classifier, is the sequential floating search method (SFSM) [23], [24]. The features described in Section 3 are ordered according to their contribution to classification accuracy. The plot in the increase in performance as each feature is added by the SFSM. None of the features were eliminated during the selection process even though removing the features 10 and 11 does not considerably degrade the performance of the classifier. In addition to classification using the features described in Section 3, experiments with dimensionality reduction using PCA and LDA techniques were conducted. The application of PCA reduced the classification accuracy to 78 percent. This is probably due to the difference in the scales of various features. Applying LDA resulted in approximately the same classification accuracy. However, LDA was able to attain this using only eight extracted features instead of the 11 input features.

V. CLASSIFICATION OF CONTIGUOUS WORDS AND TEXT LINES

In many practical applications, contiguous words belonging to the same script are available for classification. We expect that the script recognition accuracy will improve as the number of consecutive words of text in a test sample increases. In the case of online script recognition, this boils down to the number of words of text that is required to make an accurate prediction. The plot in Fig. 11 shows the increase in accuracy of the combined classifier as a function of the number of words in a test sample. A set of words was considered as a single pattern for classification in this case. We notice that with five words, we can make a highly accurate (95 percent) classification of the script of the text. The script classification accuracy improves to 95.5 percent when we use an

A. Neural Network System for Continues Handwritten Words Recognition

A new Method of continuous hand written word recognition is derived. This Method performs segmentation of the word onto triplets and triplets contain 3 letters. And two subsequent of triplets have 2 common letters. Such overlapping gives high recognition rate. The main problem with recognition

system is performing the operation on continuous word. In this each word is subdivide into triplet and each triplet contains three letters. Two neighbour triplets always contain two common letters which represent the overlapping between letters. This overlapping is used to give high recognition rate.

VI. CONCLUSION AND FUTURE SCOPE

This article represent the different technique are available for recognize the hand written documents. This article also focuses on that in today's world hand writing reorganization is very difficult but very important. There are many applications where we need hand writing recognition system like bank cheque, postal addresses, and form documents. In all the techniques main stage is feature extraction. This Paper represents the comparison between all the techniques. This algorithm can be used for recognize Hindi, Punjabi, Urdu and many more languages. We can add fuzzification with Back propagation algorithm to improve the efficiency and correctness of the algorithm. This algorithm can be used to recognize the word and paragraph also.

REFERENCES

[1] S. Mori, C.Y. Suen and K. Kamamoto, "Historical review of OCR research and development," Proc. of IEEE, vol. 80, pp. 1029-1058, July 1992.
[2] V.K. Govindan and A.P. Shivaprasad, "Character Recognition – A review," Pattern Recognition, vol. 23, no.7, pp. 671- 683, 1990.
[3] R. Plamondon and S. N. Srihari, "On-line and off- line handwritten character recognition: A comprehensive survey,"IEEE. Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 63-84, 2000
[4] U. Bhattacharya, and B. B. Chaudhuri, "Handwritten numeral databases of Indian scripts and multistage recognition of mixed numerals," IEEE Transaction on Pattern analysis and machine intelligence, vol.31, No.3, pp.444-457, 2009.
[5] Kharb, L., & Singh, R. (2008). Assessment of component criticality with proposed metrics. INDIACom-2008: Computing for Nation Development, by AICTE, IETE, and CSI, 453-455.
[6] Kharb, L. (2014). Proposing a Comprehensive Software Metrics for Process Efficiency. International Journal of Scientific and Engineering Research (IJSER), 5(6), 78-80.
[7] 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.

