American Journal of Intelligent Systems 2017, 7(3): 64-67
DOI: 10.5923/j.ajis.20170703.05

Recognition of Formatted Text using Machine Learning Technique

Rakshana J. Shetty*, **Nithin Kumar Heraje**

Department of Computer Science & Engineering, St Joseph Engineering College, Mangaluru, India

Abstract Character recognition plays an important role in extracting the required text from a document. It is vital in many areas like banking and health services where the extraction of some of the details of the customers and patients saves the time, like extracting the bank details or the medical condition of the patients etc. Keeping this vital role of in mind this system is implemented. The main aim is to recognize the printed characters in a given input image and extracting it. It is the process in which the characters are detected and recognized from an image. Optical character recognition for the formatted English text is done. The Machine Learning technique is used where the system is initially trained for all the alphabets and numbers of the English language along with the desired output. Finally the accuracy of the system is plotted according to the output obtained.

Keywords Machine Learning Technique, Character Recognition, Artificial Intelligence

1. Introduction

Formatted Character Recognition (FCR) has various practical applications and is considered to be one of the most fascinating areas of pattern recognition. The interface between human and machine can be improved and it can have an immense contribution in the advancement of an automation process. The mechanism consists of converting machine printed document into text format which can be edited if required.

The main purpose here is to take formatted English characters as input, process it, train the system, to recognize the pattern and produce the output which is then transferred to a text file or a doc file. The produced output can then be modified if required. The characters of English language are only recognized here but it can be further developed to recognize the characters of different languages as well.

The system implemented has four different steps. Initially pre-processing is done which amplifies the image in advance to processing. Next step is segmentation which helps in locating each of the individual characters and its boundaries.

Here line segmentation, word segmentation and character segmentation is performed so that individual segmentation of characters is achieved. The third step is to identify the features of the individual characters. This step is trivial as it upgrades the identification of the characters. The final step is the classification. The template matching technique is used here so that the characters can be matched accordingly so that the errors can be reduced.

2. Literature Review

Some of the methods implemented for character recognition are discussed below.

Chirag I Patel et al. implemented a method where the main objective is recognizing the characters in a given scanned document and studying the effects of changing the models of ANN. Today Neural Networks are widely preferred for Pattern Recognition chores. Different behaviours of various models of Neural Network used in OCR are discussed. Several parameters namely number of Hidden Layers, size of these Hidden Layers and epochs are considered. Multilayer Feed Forward network along with Back propagation is deployed. In the Pre processing stage some simple algorithms for segmentation of characters, normalizing of characters and De-skewing is implemented [1].

Anshul Gupta et al. adapted segmentation based approach for recognition of cursive word. Initially the segmentation of cursive words into individual characters is performed. Later these words are compared with the words in the dictionary so that the meaningful word can be obtained [2].

Rafael M. O. Cruz et al. performed a method where recognition is performed for each individual cursive characters. Nine different features have been extracted from the characters and the drawbacks of these features has been explained. The two features proposed are edge map and multiple zoning and these two features are further modified. Here a nine layered multilayered perceptron is used to which each individual feature is given as input. The output which is obtained from the classifier is merged with each other with the help of various rules like mean rule, sum rule etc. The

* Corresponding author:
rakshana123shetty@gmail.com (Rakshana J. Shetty)

Published online at http://journal.sapub.org/ajis

accurate result is obtained is by using the edge map feature [3].

M. Blumenstein et al. used techniques for segmented recognition of characters which are neural network based. Two unique features of feature extraction are explored along with two different neural networks. Back-Propagation (BP) and Radial Basis Function (RBF) network classifiers are used for comparison of directional and transition features. These are the two different features used [4].

Yong Haw Tay et al. implemented the recognition based segmentation method for the identification of cursive words. A comparative study is made between two methods of recognition. For recognition the first system makes use of the amalgamation of Neural Network and Hidden Markov Model (HMM). Discrete HMM is made use in second method [5].

Radmilo M. Bozinovic et al. implemented Holistic method for recognition of cursive word. Here representation of a word is done with the help of different phases of variation like points, letter, features etc. Based on the statistical dependencies among letter and feature, generation of a vector is achieved [6].

H. Bunke et al. deployed Holistic method for implementing the recognition of cursive word. Features are extracted from the skeleton of word. From the words edge information the feature vector is generated. The features include, location of edge with respect to four relative reference line [7].

3. Proposed Methodology

The block diagram of the formatted character recognition system is shown in figure 1.

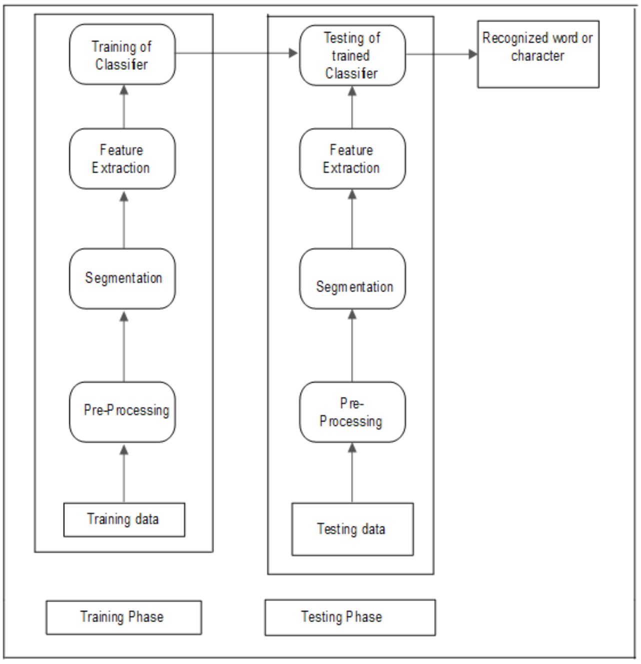


Figure 1. Block diagram of text recognition system

The gathered data is divided into training data and testing data. To train the system the data required is called training data and to test the system the data required is called testing data.

A. *Pre-processing*

In this stage a series of operations are performed on the scanned input image. The image rendering feature is enhanced which is suitable for segmentation. Here the major operation which takes place is the segmentation of interesting pattern from the background. Grey threshold is done to convert the powerful intense image to a two discrete valued image i.e. a binary image. Here Otsu method is used for this process. Generally, noise filtering, smoothing and normalization is achieved in this step. The pre-processing also defines a tightly packed representation of the pattern. Binarization is the process of converting a pixel image into a binary image. The edges of the binarized image are made to appear wider. And also the function bwareaopen is used to filter unwanted pixels in the image.

B. *Segmentation*

In this stage, an image which is usually a sequence of characters is made to break down into sub-images of solitary characters. The input image obtained in the pre-processed stage is segmented into solitary characters by designating a number to each individual character using a labeling procedure. Labeling delivers information about the number of characters the image contains. Each character is unvaryingly rescaled into pixels. Normalization: After extracting the character the size of the characters needs to be normalized. Clip data is used so that the image displayed is clipped at appropriate edges.

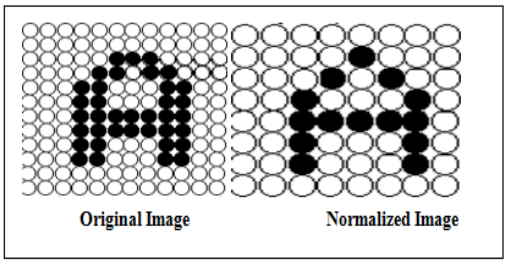


Figure 2. Normalization example

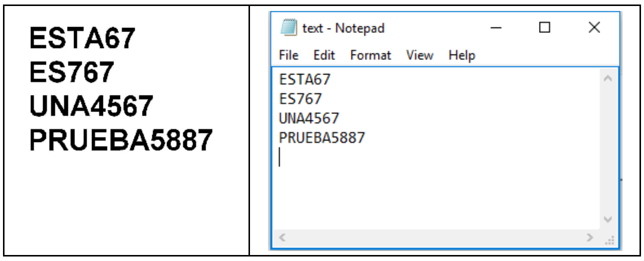
Segmentation includes: line segmentation which is segregation of the lines from the paragraph, Word segmentation which is the segregation of word from the lines and Character segmentation which is segregation of characters from the words.

C. *Feature Extraction*

Here, the trivial features of the individual characters which are required for classifying them at the identification stage are extracted. This is a significant stage as its victorious operation upgrades the recognition rate and declines the rate of misclassification. Feature vector is generated by different features like extracting directional features, binary features etc.

D. *Classification*

This is the decision making stage of the recognition system and it employs the features extracted from the preceding stage. The feature vector is symbolized as Y where Y = (f1, f2,....., fn) where f denotes features and n is the no. features extracted from character. Depending on the comparison of the above specified feature vector, characters are efficiently classified into suitable class and recognized.



ESTA67
ES767
UNA4567
PRUEBA5887

```
ESTA67
ES767
UNA4567
PRUEBA5887
```

Figure 3. Sample Output

In machine learning technique the training data with correct details of the class is applied to train the model. This model is made use to test the data for correct classification. Training data comprises of both, the input and the expected results. The model subjected to a learning process and depending on the learning it classifies the test data. Corr2 function is used and the computation of the correlation of the input image and the template is done.

Here a folder is created with the images of the alphabets of the English language along with the numerals. This data provided is used to train the model so that the characters are recognized.

Once the model is trained, the match pattern is obtained to generate the associated character. Output will be the editable version of the uploaded image and will be saved in a .doc or in .txt file. A sample output obtained is shown in the Figure 3.

4. Evaluation

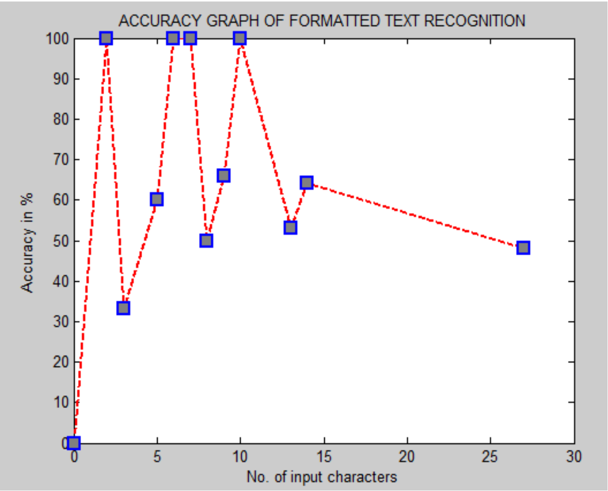


Figure 4. Accuracy Graph

The evaluation is done by plotting the accuracy graph which is represented in the Figure 4. The accuracy graph is obtained by plotting the number of input characters against the percentage of accuracy attained. Here the total number of properly recognized character in the output image is divided by the total number of input characters. And this fraction is multiplied with 100 to get the percentage accuracy. The graph obtained varies depending on the output obtained for different number of characters in the input image. The graph represented below depicts the output obtained for the set of images given.

5. Conclusions

Character recognition is a fascinating field due to its vivid applications in different sectors like banking, healthcare etc. It has drastically reduced the human involvement. The identification of handwritten characters is lot more difficult than the formatted characters due to the variations in the human writing styles. Different methods are being deployed for this purpose but none of the methods assure to give 100% accuracy. Further improvements can be done by translating the characters of the uploaded image to characters of different languages and recognition of cursive characters and so on.

REFERENCES

[1] Chirag I Patel, Ripal Patel and Palak Patel, “Handwritten Character Recognition Using Neural Networks”, International Journal of Scientific & Engineering Research vol.2, Issue 5, May-2011.

[2] Anshul Mehta, Manisha Srivastava and Chitralekha Mahanta “Offline handwritten character recognition using neural network”, in Proceedings of 2011 IEEE International conference on computer applications and Industrial Electronics.

[3] Rafael M. O. Cruz, George D. C Cavalcanti and Tsang Ing Ren, “An Ensemble classifier for offline Cursive character recognition using multiple features Extraction technique”, in Proceedings of 2010 IEEE International conference.

[4] M. Blumenstein, B. Verma and H. Basli, “A Novel Feature Extraction Technique for the Recognition of Segmented Handwritten Characters”, in Proceedings of the 2003 Seventh International Conference on Document Analysis and Recognition (ICDAR’03).

[5] Yong Haw Tay, Pierre-Michel Lallican, Marzuki Khalid, Christian Viard Gaudin and Stefan Knerr, “An Offline Cursive Handwritten Word Recognition System”, in Proceedings of 2001 IEEE International conference.

[6] Radmilo M. Bozinovic and Sargur N. Srihari, “Off-Cognitive line Cursive Script word recognition”, in Proceedings of IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol.11. No. 1, January 1989.

[7] H. Bunke, M. Roth and E.G. Schukat-Talamazzini, "Offline Cursive Handwriting Recognition Using Hidden Markov Models", Pattern Recognition, Vol. 28, No. 9, Elsevier 1995.