Survey and Classification of Character Recognition System

Priya Sharma[#1], Randhir Singh

[#]*Lecturer, Department of ECE*
SAI Polytechnic, Badhani, Punjab, India
Head, Department of ECE
SSCET, Badhani Punjab, India

***Abstract*— Variation in handwriting among different writers occurs since each writer possesses own speed of writing, different styles, sizes or positions for characters or text. Variation in handwriting styles also exists within individual person's handwriting. This variation may take place due to: writing in various situations that may or may not be comfortable to writer; different moods of writer; style of writing same characters with different shapes in different situations or as a part of different words; using different kinds of hardware for handwriting. This paper provides a survey, and classification of various character recognition techniques.**

***Keywords*—Character recognition, Pre-processing, Segmentation, Features.**

I. Introduction

Handwriting recognition is in research for over four decades and has attracted many researchers across the world. Character recognition systems translate such scanned images of printed, typewritten or handwritten documents into machine encoded text. This translated machine encoded text can be easily edited, searched and can be processed in many other ways according to requirements. It also requires tinny size for storage in comparison to scanned documents. Character recognition systems help humans ease and reduce their jobs of manually handling and processing of documents. Computerized processing to recognize individual character is required to convert scanned document into machine encoded form.

Character recognition is mainly of two types online and offline. In online character recognition, data is captured during the writing process with the help of a special pen on electronic surface. In offline recognition, prewritten data generally written on a sheet of paper is scanned.

1) Offline Character Recognition: Generally all printed or type-written characters are classified in offline mode. Off-line handwritten character recognition refers to the process of recognizing characters in a document that have been scanned from a surface such as a sheet of paper and are stored digitally in gray scale format. The storage of scanned documents have to be bulky in size and many processing applications as searching for a content, editing, maintenance are either hard or impossible. Such documents require human beings to process them manually, for example, postman's manual processing for recognition and sorting of postal addresses and zip code. Character recognition systems translate such scanned images of printed, typewritten or handwritten documents into machine encoded text. This translated machine encoded text can be easily edited, searched and can be processed in many other ways according to requirements. It also requires tinny size for storage in comparison to scanned documents.

2) Online Character Recognition: The online mode of recognition is mostly used to recognize only handwritten characters. In this the handwriting is captured and stored in digital form via different means. Usually, a special pen is used in conjunction with an electronic surface. As the pen moves across the surface, the two- dimensional coordinates of successive points are represented as a function of time and are stored in order. Recently, due to increased use of handheld devices online handwritten recognition attracted attention of worldwide researchers. This online handwritten recognition aims to provide natural interface to users to type on screen by handwriting on a pad instead of by typing using keyboard. The online handwriting recognition has great potential to improve user and computer communication.

In online handwriting recognition, it is very natural for the user to detect and correct misrecognized characters on the spot by verifying the recognition results as they appear. The user is encouraged to modify his writing style so as to improve recognition accuracy. Also, a machine can be trained to a particular user's style. Samples of his misrecognized characters are stored to aid subsequent recognition. Thus both writer adaptation and machine adaptation is possible.

II. Methodology

The character recognition system involves many steps to completely recognize and produce machine encoded text. The computer actually recognizes the characters in the document through a revolutionizing technique called Character Recognition. The various phases involved in character recognition are termed as: Pre-processing, Segmentation, Feature extraction and Classification. The block diagram of proposed recognition system is shown in figure 1.

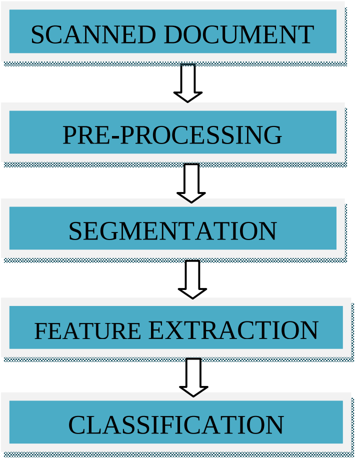

Figure 1: Block diagram of the character recognition system

III. Pre-processing

The image is taken and is converted to gray scale image. The gray scale image is then converted to binary image. This process is called Digitization of image. Practically any scanner is not perfect; the scanned image may have some noise. This noise may be due to some unnecessary details present in the image.

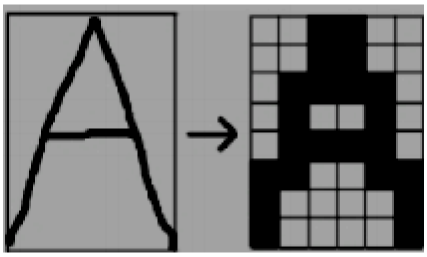

Figure 2: Digitized image

The denoised image thus obtained is saved for further processing. Now, all the templates of the alphabets that are pre-designed are loaded into the system.

IV. Segmentation

In segmentation, the position of the object i.e., the character in the image is found out and the size of the image is cropped to that of the template size. Segmentation can be external and internal. External segmentation is the isolation of various writing units, such as paragraphs, sentences or words. In internal segmentation an image of sequence of characters is decomposed into sub-images of individual character.

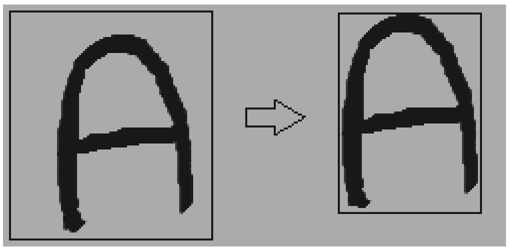

Figure 2: Segmented image

V. Feature Extraction

We have used following listed features for our experiment. First three types of features namely zone density, projection histograms and 8-directional zone density features can be categorized as statistical features while fourth type to tenth type of features can be categorized as geometric features. On the basis of these types of features we have formed different combinations. The characters are classified using each of these feature vectors in neural network classifier.

1) Zone Density Features

In zone density features, the character image is divided into N×M zones where N is the number of rows and M is the number of columns. From each zone features are extracted to form the feature vector. We have created 16 (4×4) zones of 8×8 size each out of our 32×32 normalized samples by horizontal and vertical division.

By dividing the number of foreground pixels in each zone by the number of background pixels in each zone, we obtained the density of each zone. Thus we obtained 16 zone density features.

$$\textbf{Zone density} = \frac{\textbf{no. of 1's in each zone}}{\textbf{no. of 0's in each zone}}$$

2) Projection Histogram Features

Projection histograms count the number of pixels in specified direction. In our approach we have used three directions of horizontal, vertical and diagonal traversing. We have created three types of projection histograms- horizontal, vertical, diagonal-1 (left diagonal) and diagonal-2 (right diagonal). In our approach projection histograms are computed by counting the number of foreground pixels. In vertical histogram, these pixels are counted by column wise. In diagonal-1 histogram the pixels are counted by left diagonal wise and for this purpose the image is rotated by 45 degree. Again this image is rotated by 45 degree to get the horizontal histogram and the pixels are counted by row wise. Again the image is rotated by 135 degree to obtain the right diagonal histogram.

3) Directional Zone Density Features

In 8-directional zone density features, we have created 4 (2×2) zones of 16×16 size each out of our 32×32 normalized samples by horizontal and vertical division. Then each zone is divided into 2 triangular zones, thus forming total 8 triangular zones. The density of each triangular zone is then calculated by dividing the foreground pixels by the total number of pixels in each zone. Then this density of each triangular zone is considered as a feature.

4) Character Area (A) and Perimeter (P)

Area of characters in a binary image is the number of non-zero pixels in a character. Perimeter of characters in a binary image is the length of the smoothest boundary in pixels.

5) AP Feature

The ratio of area of character to perimeter of character is taken as AP feature. It is obtained by dividing the number of non-zero pixels in a character to the length of the smoothest boundary in pixels.

$$\textbf{AP Feature} = \frac{\textbf{Area}}{\textbf{Perimeter}}$$

6) Minor Axis and Major Axis Length

`Minor Axis Length` is the length in pixels of the minor axis of the character that has the same normalized second central moments as the region. `Major Axis Length` is the length in pixels of the major axis of the character that has the same normalized second central moments as the region.

7) Orientation

`Orientation` is the angle in degrees ranging from -90 to 90 degrees between the *x*-axis and the major axis of the character that has the same second-moments as the region.

8) Eccentricity

The eccentricity is the ratio of the minor axis length to the major axis length.

$$\textbf{Eccentricity} = \frac{\textbf{Minor axis length}}{\textbf{Major axis length}}$$

VI. CLASSIFICATION

The classification is done by using the single MLPNN with Gradient descent with momentum and adaptive learning backpropagation algorithm. In hidden layer and output, the sigmoid activation function is used. The features computed are used for classification. In the present study, we have divided the features into three equal halves for training, validation and testing. We have used the extracted features separately for training, validation and testing as well as combined features obtained after appending.

VII. CONCLUSIONS

A number of techniques that are used for character recognition have been discussed. The main research is currently going on in extending Character Recognition to all the popular native languages of India like Punjabi, Telugu, Tamil etc., Template matching method which is easy to implement due to its algorithmic simplicity and higher degree of flexibility to the change of recognition target classes. Its recognition is strongest on monotype and uniform single column pages and it takes shorter time and does not require sample training but one template is only capable of recognizing characters of the same size and rotation is not illumination-invariant. Neural network has ability to recognize characters through abstraction is great for faxed documents and damaged text and ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns

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] Optical character recognition http://en.wikipedia.org/wiki/Optical_character_recognition

[3] 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, 2001, 31(2), pp. 216 – 233.

[4] J. Mantas, "An overview of character recognition methodologies", Pattern Recognition, Vol. 19, pp 425-430 (1986).

[5] V. K. Govindan and A. P. Shivaprasad, "Character recognition – A survey ", Pattern Recognition, Vol. 23, pp 671-683(1990).

[6] B. Al-Badr and S.A. Mahmoud, "Survey and bibliography of Arabic optical text recognition", Signal Processing, Vol.41, pp. 49-77(1995).