Signature Analysis based Character Recognition Approach

Er. Muheet Ahmed Butt  
Scientist, Directorate of Information Technology & Support Systems,  
University of Kashmir, Srinagar, J&K,  
Er. Majid Zaman  
Scientist, Directorate of Information Technology & Support Systems,  
University of Kashmir, Srinagar, J&K, India

Abstract:

The Character Recognition of both typed as well as handwritten characters is still an open area of research. Although significant success has been achieved in type written characters but in handwritten it is still to touch an appreciable level. This research proposes an approach of character recognition which besides producing better results of recognition is extremely less complicated as compared to the prevalent ones. The proposed method is independent of the size of the Character being recognized and is able to accommodate character styles which have slight variations and also does not require thinning and other preprocessing measures as is required in other approaches.

Keywords

Character set, projections, recognition, features, document processing, Histogram, Digitization.

Introduction

The character recognition over the past three decades has been an important area of research. The importance of the character recognition assumed great significance ever since the office automation projects were taken up. Presently the character recognition forms one of the most important activities in document processing. Considering the fact that different languages have different character sets therefore intense research has been going on for automating their recognition during document processing [1][2].

The conventional methods used for the recognition of the characters mostly use a matrix based approach [3] where each character is divided into a predefined number of rows and columns. Then depending upon the character under process a particular set of cells in horizontal and vertical direction is selected. Similarly another approach called connected component [4] traces the character under process from one end to another to find its different parts. Approaches of this nature involve excessive computations and are mostly time consuming. In these conventional approaches some pre-processing like thinning is required before actually taking up the actual character for recognition [5].

This paper presents an efficient procedure for character recognition. The procedure besides being very fast is simple to implement. The method used, transforms the character set into a new one by way of projections and the technique is called as Signature Analysis. These projections are drawn in both horizontal as well as vertical directions. The character recognition decision is made on the basis of the properties of the images obtained as a result of horizontal and vertical projections. The projections can be drawn in any direction and at any angle but we propose that resultant images after projections should be simpler as compared to the original images. Similarly the objective of drawing the projections should be to get as small set of distinct features as possible. A knowledge base is created which stores the features of different characters and is always referred while making the character recognition decision.

Selection of the Feature Set

The characters to be recognized are digitized by using any commonly used scanner. Initially a knowledge base is formed which contains the features of each character of the language. The selection of the features forms one of the most important tasks in the character recognition system. Selection of large number of features unnecessarily results in a complicated recognition procedure whereas an incomplete set of features may result in incorrect recognition system.

Therefore selection of a set of features from projections obtained after horizontal and vertical directions is an important factor in developing an accurate character recognition system. The aim in the selection of the attributes should be to develop an efficient recognition system involving minimum computations. The different potential attributes in horizontal as well vertical projections which can be considered are as under:

a) Long Size Line segments  
b) Medium Size Line Segments  
c) Position of Line Segments  
d) Number of Line Segments  
e) Projection type

On the basis of the above attributes the different characters are classified into different groups. The classification of the characters is based on similarities in attributes of the resultant images obtained after horizontal projections. The figure 1 represents the vertical projections of the characters whereas the figure 2 shows the horizontal projections. The groups so formed on the basis of the attributes of the horizontal projections drawn are as under:
Formation of the Knowledge Base

This is the fundamental step in developing a character recognition system. The knowledge base is developed to store the features of different characters and is always referred during the recognition process. Therefore the development of a comprehensive knowledge base is required for achieving accurate results during recognition. In order to develop the required knowledge base initially the character set of the language is digitized and the attributes as mentioned in section 2 are extracted by drawing projections in both horizontal as well as vertical directions. A two dimensional array is used to store the data of each character after projections are drawn. This array is processed to get the attributes of interest. The attributes so obtained are stored in an array of structures where each record is used to contain the information of one character. In this way the size of the array of structures equals the size of the character set used. The following algorithm is used for the creation of the knowledge base:

1. Start
2. digitize character set
3. store the values of the projections in an array
4. while not end of array
5. count for number of pixels
   a. if number equals the previous count then ignore
   b. else go to step 3
6. find column(s) with maximum number of pixels
7. find column(s) with minimum number of pixels
8. find column(s) with \(<\) maximum and \(>\) minimum number of pixels
9. find the column position of each line segment obtained from step 4 & 6
10. store the values of step 6,7,8 & 9 in the array of structures
11. end

The process of creation of the knowledge base is shown as under.

In the system proposed the characters to be recognized are assumed to be either type written or handwritten. The handwritten characters are assumed to be written either in square boxes or without boxes but not connected with one another. The proposed system does not require any preprocessing in respect of thinning [5] and hence in the first instance the characters are digitized. The digitization of characters using 300 dpi or higher works fine. Every character consists of a set of lines, curves, loops or a combination of these primitives. The projections of the characters result in an image with quite different attributes as compared to the actual character. The vertical projection contains the spatial histogram of the character when scanned in the vertical direction whereas the horizontal projection results in the spatial histogram of the character in the vertical direction. These spatial histograms are stored in a two dimensional array with columns of one row containing the pixel position and the corresponding columns of the other row containing the number of pixels. The number of columns of the array is directly proportional on the width of the character. The array so formed is processed to extract the features required for the recognition. The following algorithm is used for the extraction of the attributes:

1. Start
2. Traverse the Complete Array in a loop
3. count for number of pixels
   a. if number equals the previous count then ignore
   b. else go to step 3
4. find column(s) with maximum number of pixels
5. find column(s) with minimum number of pixels
6. find column(s) with maximum+minimum/2 number of pixels
7. find the column position of each line segment obtained from step 4 & 6
8. end
The features so obtained are compared with the entries forming the knowledge base and the one matching returns the character containing these features.

Experimental Results

The entire OCR development and experimentation has been carried out on an Intel Core 2 Duo. The proposed procedure has been applied on documents containing English characters. The characters were separated by using segmentation procedure. The results were highly satisfactory with a success rate of about 95%. The success rate is dependent on the segmentation as well as on quality of the handwritten characters.

Conclusion

The character recognition is an active area of research and Automatic Recognition of characters of different languages has been going on since decades. Initially the research concentrated on the recognition of typed characters however later recognition of handwritten characters were also taken up. The development of the system involves formation of a knowledge base where initially the features of the character set of a language are stored and later during the recognition process the features of the characters to be recognized are extracted and compared with those of the knowledge base. Whenever a match is found in the knowledge base character recognition is said to be successful. Although OCR’s which show 100% success in recognizing typed characters have been developed but development of such a system for handwritten characters is still a reality.

References