A Brief Review and Survey of Segmentation for Character Recognition

Rohini B. Kharate *, Dr.S.M.Jagade 2, Sushilkumar N. Holambe 3
1 P.G. Student (ME-E&TC), TPCT’s College of Engineering, Osmanabad (M.S.) India.
2 Principal, TPCT’s College of Engineering, Osmanabad (M.S.) India.
3 T.P.C.T’s college of Engineering Osmanabad- India.

ARTICLE INFO

Keywords:
Image Segmentation
Euclidean Distance Metric
Pattern Feature Extraction Recognition

General Terms:
Character recognition
gray-scale character recognition
multistage graph search
classification-based segmentation
Training
Learning rule
Distorted patterns

ABSTRACT

In the proposed methodology, the character segmentation regions are determined by using projection profiles and topographic features extracted from the gray-scale images. Then a nonlinear character segmentation path in each character segmentation region is found by using multi-stage graph search algorithm. Finally, in order to confirm the nonlinear character segmentation paths and recognition results, recognition-based segmentation method is adopted. Through the experiments with various kinds of printed documents, it is convinced that the proposed methodology is very effective for the segmentation and recognition of touched and overlapped characters. This paper is an approach to develop a method to get the optimized results using the easily available resources. Segmentation helps to extract out some common features among distinct handwriting styles of different people.

1. Introduction

Character recognition, is known as OCR (Optical Character Recognition) is an area within the pattern recognition. Optical Character Recognition deals with automatic recognition of different characters in a document image leading to clear and unambiguous recognition, analysis and understanding of the document content. The task of recognition can be broadly separated into two categories: machine printed data and the handwritten data. Machine printed characters are consisting of font used by user. They are unique and uniform. While handwritten characters are non-uniform; there size, shape depends on writer and the pen used by the writer. Handwriting of same writer may vary depending on the situation in which he is writing. But it is very difficult to design software, which is capable of identifying the characters with great accuracy. Basically, identification of the characters is one type of pattern recognition process. Variation in handwriting leads to great difficulty in identifying the character patterns. Different writing styles lead to the distortion in patterns from the standard patterns used to train the network, giving false results. A strong generalization method is required to identify the distorted patterns. Multiscale Training Technique (MST) is used in many places to solve the generalization problem [1, 2, and 3]. Results of MST depend largely on resolution of the character images. Image resolution and the training speed have to be optimized to achieve the highest percentage of accuracy. Work has been done on identification of the characters in Devnagri script by combining multiple feature extraction techniques like intersection, shadow feature, chain code histogram and straight line fitting, [4]. Another approach towards feature extraction technique is to calculate only twelve directional feature inputs depending upon the gradients, where the features of the hand written characters are the directions of the pixels with respect to their neighboring pixels, [5]. Hybrid methods are also applied to recognize the handwritten characters. One such method is a prototype learning/matching method that can be combined with support vector machines (SVM) in pattern recognition, [6]. K-nearest neighbor methods can be used to recognize the patterns, [7]. In K-nearest neighbor method the pattern is obtained by looking into k number of nearest patterns having the least Euclidean distance with that of the pattern. Artificial Neural Nets (perceptron learning) are used to train the nets and later using the nets identifying the characters, [8, 9, and 10]. But obtaining 100% accuracy is still a challenge for many such nets. In this paper a concept of recognizing hand written character pattern has been developed and implemented called Row-wise segmentation technique. RST helps in minimizing errors in pattern recognition due to different handwriting styles to great extent. In this method input pattern matrix is segmented row-wise into different groups. Target pattern is also grouped where each group is the numeric equivalent of the chronological position of each English alphabet. Each input segment is fully interconnected with each target group. Number of target groups is equal to the number of rows in the input matrix. In general, the overall program has been divided into two parts, training and testing. Training requires the net to read segmented input patterns and testing requires the net to read any test character pattern, to read the produced target samples and to count the majority of samples and to find out the numeric equivalent of the sample to identify the character.

* Corresponding author.
Email address: kharateborah@gmail.com
2. Character Recognition & Research

Document image understanding and analysis means that transforms the information of a document in the from of paper into an electronic format i.e. text without manual keyboard entry. It is still challenging and interesting task to design a system which gives high recognition accuracy, without considering quality of the input document and different character font style variation. Optical Character Recognition (OCR) is the process of translating images of handwritten, printed text into a format understood by machines. The worlds information of literature, history, and other information is in hard-copy documents. OCR systems convert this information by converting the text on paper into electronic form. OCR system based on segmentation can be divided as preprocessing of given input, segmentation, feature extraction and Classification. For document analysis we have to use OCR system based on segmentation. We can avoid preprocessing if the document is noise free but its not possible because when we scan the noise appear in the document image. Segmentation is the decomposition of an image into sub images. Segmentation is dependent on local decisions with regards to shape similarity, as well as global decisions with regards to surrounding context. In the late 1960s and 1970s researchers observed that segmentation cause more errors in reading characters, whether hand or machine-printed. In the 1980’s researchers give new dimension to OCR to less constrained documents [1]. Some authors have surveyed segmentation [2] [3] [4] [5] [6] [7], or document analysis [8] [9] more details can be found in [10].Many of techniques have been developed for Latin but for Devanagari Veena Bansal and R.M.K. Sinha [11][12], U. Gaurin, B. B. Chaudhuri [13], had done but still we are lagging. More research in noise free, error free and distortion free segmentation technique is required for high accuracy and recognition for Devanagari document processing and OCR. The extensive applications of Handwritten Character Recognition (HCR) in recognizing the characters in bank checks and car plates etc. have caused the development of various new HCR systems such as optical character recognition (OCR) system. There are so many techniques of pattern recognition such as template matching, neural networks, syntactical analysis, wavelet theory, hidden Markov models, Bayesian theory and minimum distance classifiers etc. These techniques have been explored to develop robust OCR systems for different languages such as English (Numeral) [1-3], Farsi [4], Chinese (Kanji) [5,6], Hangul (Korean) scripts [7], Arabic script [8] and also for some Indian languages like Devnagari [9], Bengali [10], Telugu [11-13] and Gujarati [14]. OCR is an area of pattern recognition process that has been the subject of considerable research during the last few decades. Machine simulation of human functions has been a very challenging research field since the advent of digital computers [15]. The ultimate objective of any OCR system is to simulate the human reading capabilities so that the computer can read, understand, edit and do similar activities as human do with the text. Mostly, English language is used all over the world for the communication purpose, also in many Indian offices such as railways, passport, income tax, defense and public sector undertakings such as bank, insurance, court, economic centers, and educational institutions etc. A lot of works of handwritten English character recognition have been published but still minimum training time and high recognition accuracy of handwritten English character recognition is an open problem. Therefore, it is of great importance to develop automatic handwritten character recognition system for English language. In this paper, efforts have been made to develop automatic handwritten character recognition system for English language with high recognition accuracy and minimum classification time. Handwritten character recognition is a challenging problem in pattern recognition area.

One work is due to Kumar & Singh [13] and they proposed Zernike moments based approach for English hand written character recognition . The other work on English hand written character recognition is proposed [14] and 64 dimensional chain code features have been used in that work, but still no any standard OCR is available for the same. An excellent survey of the area is given in [15]. For recognition of handwritten numerals, Ramakrishnan et al. [16] used independent component analysis technique for feature extraction from numeral images. Considered a strategy combining decisions of multiple classifiers. In all these three studies, very small sets of samples were considered. In an attempt to develop a bilingual handwritten numeral recognition system, Lehal and Bhatt [17] used a set of global and local features derived from the right and left projection profiles of the numeral images for recognition of handwritten numerals of English and Roman scripts.

3. The OCR Data Set Creation

We have collected handwritten characters from different peoples of different age group (i.e. 03 to 75), i.e. of 7000 people of different age groups, the datasheet used for collection is shown in figure 1. We have visited schools, High school, colleges, Government offices, Adult education schools for the Data collection.

While preprocessing we have consider the distortion in image because of users pen and writing quality. We have performed Preprocessing operations for rectification of distorted images, improving the quality of images for ensuring better quality edges in the subsequent edge determination step.

In order to remove noise and diminish spurious points, Which are introduced by uneven writing surface, we have used filtering operation. we have performed smoothing, sharpening, thresholding and contrast adjustment by using filtering operation [22][23], we have applied skew Normalization, slant Normalization, size normalization, curve smoothing, to remove all types of variations during the writing and obtain standardized data[24].Then we have performed thinning operation to get better features [25].
4. Feature Extraction

We have extracted feature on our dataset, the feature set consisted of local intensity distribution of gradients computed on 5 X 5 grid using 16 quantized gradient orientations [18,19]. The original binary image was converted to a gray image using Gaussian filter, here feature vector size is 400. The number of blocks are initially 9 X 9 and down sampled to 5 X 5. Here we are using 16 direction level. In order to obtain 16 directions Gaussian filter and a Robert filter are applied to the character image to obtain a gradient image. The arc tangent of the gradient is quantized into 16 directions and the strength of the gradient is accumulated in each direction in each block.

5. Euclidian Distance-Based K-NN Classification

In KNN classification, training patterns are plotted in d-dimensional space, where d is the number of features present. These patterns are plotted according to their observed feature values and are labeled according to their known class. An unlabelled test pattern is plotted within the same space and is classified according to the most frequently occurring class among its K-most similar training patterns; its nearest neighbors. The most common similarity measure for KNN classification is the Euclidian distance metric, defined between feature vectors as:

\[
\rho_{euc}(x, y) = \sqrt{\sum_{i=1}^{f} (x_i - y_i)^2}
\]

Where f represents the number of features. Smaller distance values represent greater similarity [20, 21].

6. Results

We have used 20,000 Hand written sample data files. We have used 1200 for testing. We have organized data in class, separate class is decided for each character (i.e. vowels, consonants without modifiers, consonants with modifiers(%)). Then we have taken result. Our results are given in the table 1.we have computed accuracy of each individual English Character the table shows the average accuracy.

In case of we get high accuracy and rejection and error is also less because they all are unique. But in case of constants we high rejection and error.
A Brief Review and Survey of Segmentation for Character Recognition

Table 1. Average Result Of Our Dataset.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Rejection</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>vowels</td>
<td>98%</td>
<td>3%</td>
<td>2%</td>
</tr>
<tr>
<td>consonants</td>
<td>97.50%</td>
<td>6%</td>
<td>5%</td>
</tr>
<tr>
<td>without modifiers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>consonants</td>
<td>94%</td>
<td>7%</td>
<td>9%</td>
</tr>
<tr>
<td>with modifiers (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

References