A study on structural method of feature extraction for Handwritten Character Recognition

Vijay Prasad, Y. Jayanta Singh

Dept. of Computer Science and IT, Don Bosco College of Engineering and Technology, Assam Don Bosco University, Airport Road, Azara, Guwahati, Assam – 781 017
vijay@dbuniversity.ac.in, jayanta@dbuniversity.ac.in

Abstract— This paper presents the study reports of major process involved in a handwritten character recognition system. We focus on the various feature extraction techniques as the recognition mainly depends on the features extraction. After studying the various features we have modified an existing feature extraction technique by introducing two more feature vectors. After the introduction of these two new vectors we found a considerable increase in the percentage of recognition.

Keywords— Handwritten, character, recognition, feature extraction, optical scanner, classification

I. INTRODUCTION

Handwritten character recognition is a technique of a computer application recognizes handwritten character or sentence from sources such as paper photographs, documents, touchscreens and similar devices by a computer. The image of the handwritten sentence/word/character can be gathered either offline or online. In off line technique it can be from a scanned image of a paper. In online the detecting motion of the pen tip, for example by a pen-based computer screen surface. There are certain steps of processing before an image can be recognised. First in this paper those steps are discussed and then the main part that is the feature extraction techniques is explained broadly.

![An image of a handwritten sheet (grayscale).](image)

II. STEPS INVOLVED

An offline handwritten character recognition system in general involves the following major steps:

A. Pre-processing
B. Segmentation
C. Feature Extraction

D. Classification

E. Post-processing

The fig. 2 gives a block diagram of the whole recognition system.

![Block diagram of the recognition system](image)

Fig. 1 An image of a handwritten sheet (grayscale).

Fig. 2 Block diagram of the recognition system

A. Pre-processing

When a document is scanned or say the raw data it may require some preliminary processing. This Pre-processing helps in producing the final document which will be processed finally by a handwritten character recognition system. The main objectives of pre-processing are:

- Noise reduction
- Binarization
- Skew correction
- Stroke width normalization
Noise Reduction – Normalization

Application of noise reduction techniques increases the quality of the document. The major methods are:
- Filtering (masks)
- Morphological Operations (erosion, dilation, etc)

Fig. 3 Filter applied to a scanned image to get a noise reduced image.

After normalization generally it reduces the amount of data to be processed. For eg by thinning the shape information of a character can be gathered without losing the data.

Fig. 4 Stroke width normalization.

Binarization

In Binarization a gray scale image if transformed to a binary image. Two categories of binarization are:

1. Global: In this global one threshold value for the entire document image is picked. This picking value is based on an estimated value the intensity of the image.
2. Adaptive: It uses different values for each pixel[1]

Skew Correction and Slant Removal

Skew correction methods are used to align the coordinate system of the scanner with that of the document. Its main approaches include correlation, projection profiles and hough transform etc.

Fig. 5 Hough transformation applied to a raw scanned image.

The slant of any handwritten text(s) varies from one user to another. Normalization of characters are done by using the slant removal methods.

Some of most popular deslanting techniques are given below:

a. Calculation of the average angle of near-vertical elements.


Fig. 6 Calculation of the average angle of near-vertical elements.

Fig. 7 Bozinovic – Shrihari Method (BSM).

The dominant slope of the character is found from the slope corrected characters used in this study[2]. The vertical histogram projection is calculated for a range of angles ± R. The slope of a character, \( a_m \), can be found

\[
\alpha_m = \min_{a \in [-R,R]} H \\
H = \sum_{i=1}^{N} p_i \log p_i
\]

Character correction is done by using:

\[
x' = x - y \tan(a_m) \quad y' = y
\]

B. Segmentation

It involves two major steps in the given sequence:

a. Text Line detection, for which we may use Hough Transform, projections, smearing, etc

Fig. 8 Text line segmentation

b. Word extraction for which we can use vertical projections, connected component analysis, etc.

and finally

Fig. 9 Word segmentation

For segmentation of word we may use wither explicit segmentation or implicit segmentation technique.

The explicit approaches tries to identify the smallest possible word segments that may be smaller than letters, but surely cannot be segmented further. However during the recognition process these segments are assembled into letters. It is robust and quite straightforward.

Fig. 10 Explicit segmentation
The implicit approaches recognized the words entirely without segmenting them into letters. It is effective only when the set of possible words is small and known in advance [3].

C. Feature Extraction

In this phase each character is represented as a vector. It extract a set of features. Latter these features will be used to optimised the recognition percentage. General feature extraction methods considers three types of features

i. Statistical
ii. Structural
iii. Global transformations and moments

i. Statistical Method

Some of the major statistical features used for character representation-(a)Zoning, (b)Projections and profiles and (c)Crossings and distances

Zoning: The character image is divided into N x M zones. Features are extracted from each zone at local characteristics.

Zoning-Density Features: The number of foreground pixels in each cell is considered a feature. Darker squares indicate higher density of zone pixels as shown in the following figure

Zoning-Direction Features: It is based on the contour of the character image. For each zone the contour is followed and a directional histogram is obtained by analysing the adjacent pixels in a 3 x 3 neighbourhood.

Based on the skeleton of the character image, one can distinguish individual line segments.
Crossings and Distances:
Crossings count the number of transitions from background to foreground pixels. It counts along the vertical and horizontal lines through the character image. The distances calculate the distances of the first image pixel detected from the upper and lower boundaries of the image.

ii. Structural Method
Characters can be represented by structural features with high tolerance to distortions and style variations. It is based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

Three types of features:
- Horizontal and Vertical projection histograms
- Radial histogram
- Radial out-in and radial in-out profiles

D. Classification
Classification is the problem of identifying which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. The individual observations are analyzed into a set of quantifiable properties[4].

There are no best classifier, however use of classifier depends on many factors, such as available training set, number of free parameters. Some of the important methods which can be used for classification are-k-Nearest Neighbour (k-NN), Bayes Classifier, Neural Networks (NN), Hidden Markov Models (HMM), Support Vector Machines (SVM), etc

E. Post-processing
It refer to the processing done on the image after the classification process is over to refine the result.
It also incorporated the context and shape information in all the stages of HCR systems for meaningful improvements in recognition rates.
III. OUR STUDY

We studied one of the feature extraction methods, viz., Structural: Horizontal and Vertical projection histogram. Normally in this method of feature extraction people use only two feature vectors. However by using these feature vectors we generated another two more projection and we named them as Mean H Histogram and Mean V Histogram. By introducing these two extra feature vectors we found the increase in the accuracy level of the recognition. These new feature vectors were generated by

a. an optimal span is taken, e.g. say the normalised image is 45 x 45 pixel in dimension, so we may divide the image in span of 9,

b. The mean for each span is calculated.

c. Each member in the span is initialised with the calculated mean in the previous step.

<table>
<thead>
<tr>
<th>Horizontal Histogram</th>
<th>Mean H Histogram</th>
<th>Vertical Histogram</th>
<th>Mean V Histogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>37</td>
<td>47</td>
<td>51</td>
</tr>
<tr>
<td>51</td>
<td>53</td>
<td>54</td>
<td>36</td>
</tr>
<tr>
<td>32</td>
<td>46</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>46</td>
<td>45</td>
</tr>
<tr>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>49</td>
<td>48</td>
<td>48</td>
<td>48</td>
</tr>
<tr>
<td>46</td>
<td>48</td>
<td>48</td>
<td>39</td>
</tr>
<tr>
<td>36</td>
<td>39</td>
<td>39</td>
<td>35</td>
</tr>
</tbody>
</table>

Fig. 23 A screen shot of the proposed feature vectors test engine

IV. RESULTS

For testing our implementation of these feature vectors, we created an engine, by which we can test the result with and without these new feature vectors. Example: when we test out the study with using these new feature vectors we found a considerable increase in the accuracy level of the recognition compared to that study performed without extra feature vectors. On the introduction of these new vectors it partially incorporated the zoning method which is a kind of statistical method of feature extraction.

V. CONCLUSIONS

This paper reports the basic concept of a Handwritten Character Recognition. Our study process of HCR developed certain feature vectors which enhanced the recognition. We have used the basic feature extraction technique. By using two new feature vectors that we called as Mean H Histogram and Mean V Histogram in process of feature formation; we found the increase in the accuracy level of the recognition.

ACKNOWLEDGMENT

We want to acknowledge all the faculty members of our department with special thanks to our Principal and Director of Research.

VI. REFERENCES


