A New Approach for Sentence Recognition Using Multiple Features

Shruthi Shri .D1, Dr. Jharna Majumdar2, Laxmidevi Noolvi3

1Department of CSE (PG), Nitte Meenakshi Institute of Technology, Bangalore, India
2Dean R&D, Professor and Head CSE (PG), Nitte Meenakshi Institute of Technology, Bangalore, India
3Department of CSE (PG), Assistant Professor, Nitte Meenakshi Institute of Technology, Bangalore, India

Abstract: This paper presents the work towards recognition of sentence by using segmentation based on bounding box for character recognition. After segmentation various variant and invariant features are extracted like Correlation method, Shadow feature, horizontal and vertical profiling, Chain code and Projection methods which increase the recognition capability of a character. Classification techniques are used for improving the recognition of sentence. Dictionary search for each sentence is performed, after first character is recognized and matched. The recognized sentence is spelled out by the system.

Keywords: Segmentation, Feature Extraction, classification, Character Recognition

1. Introduction

Nowadays, sentence recognition systems are used in many fields that have different applications in robotics, license plates, business cards, invoices, ID cards, driver licenses. The character recognition started from the recognition of machine printed characters and then it was developed to the recognition of the machine printed words and sentences. Gradually, handwritten digit, character and word recognition were introduced into this domain. Several research works have been focusing towards evolving the newer techniques that would reduce the preprocessing time and to provide higher recognition accuracy.

Text objects occurring in sentence images contain much information related to characters. Therefore, the text information has to be extracted efficiently, by performing pre-processing on the input sentence image. After segmenting sentence into words, words to isolated characters the features of each characters is extracted. Feature Extraction plays an important role in identification of character hence, selection of good feature set is most important aspect of printed character pattern recognition so Correlation method[6], horizontal, vertical profiling, Shadow feature[5], chain code and projection methods increases the recognition capability. The method as provides the ease of implementation and recognition [4]. Here we use the concept of clustering in identification of characters in a sentence.

2. Proposed Method

The proposed method consists of 6 main steps.

i. Preprocessing
ii. Segmentation
iii. Feature extraction
iv. Reduced Data
v. Clustering
vi. Recognition

A. Preprocessing

In the scanning process, some distorted images may be introduced due to light machine printing, quality of paper on which the character printed. Some Pre-processing steps are
performed for rectification of distorted images, improving the quality of images and size normalization. The goal of this pre-processing is to distinguish between image pixels that belong to text and those that belong to the background. Every image is first converted into a grayscale image to preserve memory and speed up further processing. After threshold is applied, the grayscale image converts into a binary image with only two possible values, black (0) and white (1).

B. Segmentation
Segmentation is an integral part of any text based recognition system. It assures efficiency of classification and recognition. Accuracy of character recognition heavily depends upon segmentation phase. Incorrect segmentation leads to incorrect recognition. Segmentation phase include line segmentation, character and word segmentation. It is important to obtain complete segmented character without any noise to ensure quality feature extraction.

The pre-processed image is considered for segmentation of words in a sentence. We use vertical and horizontal scanning to obtain the bounding box coordinates of each character. We can separate a word and character by space between them. The spacing between the word and the character will vary so drawing a bounding box for words and isolated characters in a sentence is quite challenging. Space between the characters ‘WO’, ‘OW’ ‘AO’,‘OV’, ‘VO’ is perceived by the human eye we find some space between characters but a machine cannot identify the space and segment characters. Here in this paper we propose a concept of segmentation using bounding box to segment words and characters in a sentence as shown in Figure 2.

![Figure 2: Bounding box Segmentation for word and character](image)

C. Feature Extraction
It is finding the set of parameters that define the shape of character precisely and uniquely. Feature set plays one of the important role in recognition. Transforming the input data into the set of features is called Feature extraction. There are several feature extraction methods that are based on geometric, diagonal features and pixel based feature.

1. Correlation Function
The correlation coefficient is one of the popular metric used in the literature to provide comparison of two images. For our empirical knowledge it is fundamentally of co-varying things. We consider the skeletonized character image for correlation function. The character images size depends on font size of a character, define each character into different segments, i.e. of size 21*21. The two elementary shapes chosen are horizontal and vertical line. These entire row elementary shapes were located at central position the 21*21 window. Our feature extraction stage identifies and determines the type and location of the elementary shapes in the thinned character samples. To compute correlation, we need to define the correlation points in the skeletonized character image size.

Horizontal axis of normalized character image (m) …. (1)
Vertical axis of normalized character image (n) …. (2)

Combining the axes in equation 1 and 2, we obtain following correlation points around which we take the different segments of each character:

Now, we divide the character into different segments. Each segment is of size 21*21 with its centre at the point of ‘mn’ vector in equation(3). The graphical form of this whole procedure is shown in figure 3 which contains two elementary shapes on left side and example handwritten character on the right side. The example handwritten is divided into segment windows the centre of each segment window is indicated by a cross.

Equation for Correlation:

\[
CF(m,n) = \frac{\sum_{x \in X} \sum_{y \in Y} f(x,y)g(x,y)}{\left( \sum_{x \in X} \sum_{y \in Y} f(x,y)^2 \right) \left( \sum_{x \in X} \sum_{y \in Y} g(x,y)^2 \right)^{1/2}}
\]

(1)

We find the normalized correlation value of different segments with elementary shapes by the following procedure:

Algorithm
Input: Thinned Character Image
Output: Correlation function based feature values

Step 1: Initially thinned character image is considered as input.
Step 2: The templates of size 30*30 is created i.e vertical line, horizontal line, positive slant line and negative slant line..
Step 3: Pixel by pixel move is made to match the template using correlation function based features from the thinned character.
Step 4: Correlation equation is calculated for every pixel move, if exact match for template is found then count is taken.
Step 5: Occurrence of template is counted for different templates.
Step 6: Counts are noted for different templates.
This gives a total horizontal template match of 54. An important property of these features, since they are based on variant in the character images.

2. **Shadow Feature of a character**

The shadow features for a character are computed globally and intersection features, line fitting features are computed by dividing the character image into 8 different segments. 16 shadow features are extracted from eight octants of the character image. The rectangular boundary enclosing the character image is divided into eight octants, for each octant shadow of character segment is computed on two perpendicular sides so a total of 16 shadow features are obtained shown in Fig. 4(a).

![Figure 4: (a) Projections on sides (b) Character image is divided into eight octants, Shadow feature Extracted.](image)

Shadow is basically the length of the projection on the sides as shown in figure. These features are computed on scaled image. Character recognition technique uses features obtained from Shadow, intersection and chain code histogram. After segmenting the character image into eight octants, each octant has the part of character black pixels i.e each octant’s black pixel count is obtained. Then final step is repeated to all octants in the image and count is obtained. Shadow feature is extracted.

3. **Projection Method**

The projection method does the compression of the data through a projection. Black pixel counts are taken along parallel lines through the image area generating marginal distributions. The direction of projection can be horizontal axis, vertical axis and diagonal axis. Initially character is extracted with the boundary points. Character image is segmented into four quadrants such that each and every line of a character image is scanned and black pixel counts are taken along parallel lines. As shown in Fig 4.

The direction of projection can be horizontal axis, vertical axis and diagonal axis.

![Figure 5: (a) Projection Method, (b) Values of each quadrant for diagonal projection.](image)

Here the feature of character boundary black pixel count for horizontal and vertical projection are equal. So, instead we consider the diagonal black points as shown Figure 5. As a feature of character.

4. **Horizontal and Vertical Profiling**

The four profiles of one connected component encompassed with a bound box are obtained by counting the white pixels in the four directions, rightward, leftward, upward and downward respectively until the black pixels are encountered. Calculating right and left profile gives the horizontal profiling value, upward and downward profile gives the vertical profiling values. These two values are scale invariant for any character size image.

```
Horizontal Profile = (Right view)/(Left view)
Vertical Profile = (Upside view)/(Downside view)
```

5. **Chain Code Method**

Chain Code, is a feature extraction method is investigated based on 4-neighborhood or 8–neighborhood methods. 8–neighborhood method has been implemented which allows generation of eight different codes for each character. These codes have been used as features of the character image.

**Algorithm**

**Input:** Character Boundary Image  
**Output:** Normalized chain code values.

1. **Step 1:** Find out starting point which has non zero values and store it in array.
2. **Step 2:** Initialize 0-7 total eight directions.
3. **Step 3:** Travels all 8 neighbours.
4. **Step 4:** Find first nonzero value.
D. Reduced Data
PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. PCA is that once we have found these patterns in the data, and we compress the data, i.e., by reducing the number of dimensions, without much loss of information. This method of extracting information from a higher dimensional data by projecting it to a lower dimension. The principal components as a whole form an orthogonal basis for the space of the data. Mathematically, PCA transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate, the second greatest variance on the second coordinate. Each coordinate is called a principal component.

E. Clustering
The goal of the clustering analysis is to divide a given set of data or objects into a cluster, which represents subsets or a group. Clustering is the process of identifying each character and assigning it the correct cluster group. K-means clustering is simple unsupervised learning method which can be used for data grouping or classification when the number of the clusters is unknown. Initial clusters are assigned internally i.e., by supervised method. The cluster centers are calculated, steps are repeated until convergence. A point is assigned to the cluster which is close in Euclidean distance to the point. Then characters are clustered.

F. Recognition
This paper present an approach for sentence recognition, here each sentence and every character in a sentence is recognized by integrating feature extraction values, reduced data and clustering method.

The recognition of very first character in a sentence by using features and clustering method, the search i.e., dictionary search is performed for a first character match and lists out the vocabulary of all sentences and words from the database. It compares for all characters in a sentence, the exact match is found in a database then a sentence is recognized.

3. Result and analysis
The proposed system analysed by calculating the recognition rate for characters, words and sentence images. Recognition rate (RR) calculated by:

\[
RR = \frac{\text{Number of Recognised Sentence}}{\text{Total Number of Testing Sentences}} \times 100
\]

<table>
<thead>
<tr>
<th>Input Image</th>
<th>Output analysis for proposed method</th>
<th>Recognition Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters</td>
<td>26 characters</td>
<td>99.73%</td>
</tr>
<tr>
<td>Words</td>
<td>40 words</td>
<td>95%</td>
</tr>
<tr>
<td>Sentences</td>
<td>60 sentences</td>
<td>95%</td>
</tr>
</tbody>
</table>

In proposed system, comparison of recognition method is done based on the time taken for recognition of sentence. Table (3) shows the comparison of the clustering of with and without PCA.

Table 3, illustrates different clustering methods, with including and excluding PCA. From observation of recognition rate in table K-means with PCA integrated is efficient compared to other clustering method.

<table>
<thead>
<tr>
<th>Sl no.</th>
<th>Clustering Method</th>
<th>Input Sentence</th>
<th>Time to Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>K-mean</td>
<td>Go Straight</td>
<td>0.44999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Follow Right</td>
<td>0.42800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U Turn</td>
<td>0.17609</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Go Slow</td>
<td>0.21700</td>
</tr>
<tr>
<td>2.</td>
<td>Bachelors</td>
<td>Go Straight</td>
<td>0.30400</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Follow Right</td>
<td>0.33909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U Turn</td>
<td>0.15907</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Go Slow</td>
<td>0.18900</td>
</tr>
<tr>
<td>3.</td>
<td>K-mean with PCA</td>
<td>Go Straight</td>
<td>0.02987</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Follow Right</td>
<td>0.04508</td>
</tr>
<tr>
<td></td>
<td></td>
<td>U Turn</td>
<td>0.01680</td>
</tr>
</tbody>
</table>
4. Conclusion

The experimental results show that the algorithm successfully recognises sentence by implementing multiple features and clustering methods. The algorithm worked well for printed characters. The achieved time to identify is remarkable and will make the technique to apply in concerned application. It is difficult to achieve a robust algorithm that would be good in all segments and all cases of sentence recognition without fault. This research on sentence recognition improves recognition performance without faults and accuracy of recognition system by multiple feature extraction and clustering methods.

References


