Segmentation and Feature Extraction for Cursive English Handwriting Recognition

Pritam Dhande, Reena Kharat
Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, Pune, India,
Savitribai Phule Pune University, Maharashtra.
dhandespritam@gmail.com, reenakharat@yahoo.com

Abstract—This paper aims to represent the various techniques of optical character recognition for cursive English handwriting. OCR is a challenging research area in pattern recognition and image processing. In OCR, images of printed, handwritten or typed text are scanned and converted into machine readable text. There are scanners with inbuilt OCR for printed documents but not for handwritten documents. Character recognition of handwritten cursive English script is a very challenging task. In cursive English handwriting, the characters in a word are connected to each other. So the segmentation and feature extraction of cursive English script is much difficult.

Index Terms—Optical character recognition, Feature extraction, Handwritten cursive English script, Segmentation.

I. INTRODUCTION

Optical character recognition is also called as optical character reader and it is abbreviated as OCR. OCR translates the images into machine readable format such as ASCII or Unicode. Character recognition can be classified into two types based on the type of the text i.e. machine printed text and handwritten text. Character recognition of handwritten text is more challenging than machine printed text. Because, machine printed characters are straight with uniform alignment and spacing. While the handwritten characters are not uniform and greatly varies in shape and size.

There are many advantages of OCR. When a printed text is converted to machine readable text then we can search through it with keywords, compress, edit and send it, and can store in much less space. OCR has numerous applications. It is used by blind and visually impaired persons. In banking and legal department, it is used to digitize the documents. Barcode recognition technique is used in retail industry which is also related to OCR. It is widely used in education, finance and automatic detection of number plate.

The main challenge in the recognition of handwritten characters is that every person on the earth has different handwriting. There are various other factors also which causes difference in handwriting such as multi-orientations, skewness of the text lines, overlapping characters, connected components, pressure points etc. Many scripts are there with their intrinsic variations. A single character can be written in many forms, so it is a challenging task to recognize a particular handwritten character.

There are six steps in optical character recognition process. They are as follows

- Image acquisition
- Pre-processing
- Segmentation
- Feature extraction
- Classification
- Post-processing

II. REVIEW OF LITERATURE

In the middle of 1940s, the first character recognition system is appeared with the development of digital computers.

In literature of cursive English handwriting recognition, earlier study highlighted that an off-line handwritten document...
analysis through segmentation, skew recognition and writing pressure detection for cursive handwritten document. The proposed segmentation method is based on modified horizontal and vertical projection that can segment the text lines and words even if the presence of overlapped and multi-skewed text lines. The proposed method was tested on more than 550 text images of IAM database and sample handwriting image which are written by the different writer on the different background. Using the proposed method 95.65% lines and 92.56% word are correctly segmented from the IAM dataset. Proposed work also normalizes 96% lines and words perfectly with very small error rate. Proposed skew normalization method deals with the exact skew angle and extremely efficient with compare to on hand techniques [1].

Each and every pixel in an image represents some information. The pixels which contributes to the text has more information energy. Based on this information energy, the text-lines are segmented with 92% accuracy. Artificial Neural Network is used to recognize the characters [2].

The study includes the performance of convex hull based feature set i.e. 125 features are computed by considering various bays attributes of the convex hull of a pattern, for effective recognition of isolated handwritten Bangla basic characters and digits. The recognition rate is 76.86% for handwritten Bangla characters and 99.45% for Bangla numerals [3].

In this work, a novel technique is used for recognition of Bangla cursive words. MLP classifier is used for recognition purpose. Before recognition, the words are divided into three parts i.e. upper, middle and lower parts. The various features are designed by considering octant-centroid features, modified shadow features and longest-run features. The experiments are carried out on 300 samples with accuracy 80.58% [4].

The work includes the study of different segmentation techniques for handwritten character recognition. Three levels of segmentation are presented i.e. text-lines, word and character segmentation. The need and factors which affects the segmentation process are discussed [5].

The work contains the new approach which uses the sequence of segmentation and recognition algorithms for the OCR of cursive handwriting. Hidden Markov Model (HMM) is used for recognition with accuracy 92.3% with lexicon size 50. Lexicon and HMM are combined for word-level segmentation [6].

In this work, various segmentation levels are discussed. Hough transform is used for text-line segmentation. For the separation of vertically connected components, skeletonization is used. The experiments are carried out on IC-DAR2007 dataset [7].

In this work, the novel connectivity strength function is used for segmentation process. Connectivity strength parameter is used to decide the components of the text-line. It is a language adaptive approach with accuracy 97.30% [8].

III. SYSTEM ARCHITECTURE

The proposed system consists of six modules.
C. Segmentation:

Segmentation is of three types i.e. line, word and character segmentation. Line segmentation separates the lines from a paragraph. Word segmentation separates the words from a line and character segmentation separates the characters from a word.

D. Feature Extraction:

Feature extraction is an important step in the recognition process. In this process, all the essential information about a character which is present in an image is extracted.

E. Classification:

In classification, an unknown sample is assigned to the pre-defined class. According to the extracted features, characters are classified and recognized.

F. Post-processing:

To achieve more accuracy, various post-processing techniques are used, for example, matching a recognized word with a dictionary word.

IV. ALGORITHM FOR PROPOSED SYSTEM

Horizontal projection method is used to segment a line from a paragraph [1]. In this, first horizontal histogram of an image is created. The average height of a rising section is assumed as threshold. Then the height of each rising section is checked whether it is greater or equal to the threshold, then each line is segmented from a binary image.

A. Algorithm for Line Segmentation:

1) Read a handwritten document image as a multi-dimensional array.
2) Check the image is a binary image or not. If binary image then stores it into a 2-d array IMG[ ][ ] with size MN and go to Step 4, otherwise go to Step 3.
3) Convert the image to binary image and store into a 2-d array IMG[ ][ ].
4) Construct the horizontal projection histogram of the image IMG[ ][ ] and store into a 2-d array HPH[ ][ ].
5) Measure the height, starting row position and ending row position of each horizontally rising section of horizontal projection histogram image and store into 3d array LH[ ][ ][ ] sequentially.
6) Count the number of rising section by counting the rows of the 3-d array LH[ ][ ][ ]. Then measure the threshold (Ti) value by calculating average height of rising sections from the 3-d array LH[ ][ ][ ].
7) Select each rising section from 3-d array LH[ ][ ][ ] and check the height of that rising section is less than the threshold or not. If yes then this rising sections is not considered as a line and go to Step 9, otherwise rising section is treated as a line and go to Step 8.
8) Find the rising sections starting and ending rows number from the array LH[ ][ ][ ]. Let starting and ending row are r1 and r2 respectively. Extract the line segment between r1 and r2 from the original binary image denoted by IMG[ ][ ].
9) Go to Step 7 for next rising sections till all rising section are not under consideration, otherwise go to next Step.
10) End.

After line segmentation, to normalise the skew angle, orthogonal projection method is used [1].

B. Algorithm for Word Segmentation:

1) Read a segmented binary line as 2-d binary image LN[ ][ ].
2) Construct the vertical projection histogram of the line LN[ ][ ] and store into a 2-d array LVP[ ][ ].
3) From the vertical projection histogram (LVP[ ][ ]), measures width of each inter-word and intra-word gaps and store the width into 1-d array GAPSW[ ].
4) Count total number gaps as TGP by calculating the size of GAPSW[ ]. Add width of all gaps by adding the elements of GAPSW[ ] and store into TWD.
5) Calculate the threshold (Ti) as follows: Ti = TWD/TGP
Where, Ti is the threshold value denoting average width of inter-word gaps, TWD denotes total width of all gaps and TGP denotes the total number of gaps.

6) For each \( i \), if \( GAPSW[i] > T_i \) then this gap is treated as inter-word gaps, otherwise gaps is treated as an intra-word gaps. Depending on inter-word gaps width, words are segmented from the line.

7) End

C. Algorithm for Feature Extraction:

For feature extraction, convex hull algorithm is used [3]. A convex hull is the smallest convex polygon which contains all the points in a given plane. Convex hulls have local controllability and used for affine matching. Area of a convex hull can be calculated as

\[
A = \frac{1}{2} \sum_{i=1}^{L} (x_i y_{i+1} - x_{i+1} y_i) 
\]

L = Number of order vertices, \((x_i, y_i)\) coordinates of the order vertices.

The centroid \((C_x, C_y)\) of the convex hull can be calculated as

\[
C_x = \frac{1}{6A} \sum_{i=1}^{L} (x_i + x_{i+1})(x_i y_{i+1} - x_{i+1} y_i) 
\]

\[
C_y = \frac{1}{6A} \sum_{i=1}^{L} (y_i + y_{i+1})(x_i y_{i+1} - x_{i+1} y_i) 
\]

The points which lies inside the convex hull but doesn’t belongs to an object is called as deficit of convexity. There are two types of convex deficiencies. They are as follows:

- Lakes: The region which is completely surrounded by an object.
- Bays: The region which is present between the perimeter of convex hull and the object.

By considering various bays attributes of a convex hull, 125 features are designed. A distance measure dcp is calculated. Dcp is the total count of pixels which are present from the boundary of convex hull to the pixel of nearest character in either horizontal or vertical direction.

Dcp is calculated from the four boundaries of an image. So based on that dcp in a specific direction, six topological features are calculated.

- Maximum dcp
- Average dcp
- Total numbers of rows having dcp > 0
- Total numbers of rows having dcp = 0
- Number of visible bays in that direction
- Mean row co-ordinate having dcp > 0

Along the perimeter of convex hull one more feature is calculated i.e., the number of perimeter pixels having dcp = 0. The total feature count is 125 i.e. 25 features for the overall image and 100 features in all four sub-images.

TABLE I

<table>
<thead>
<tr>
<th>SEGMENTATION TECHNIQUES</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horizontal and vertical projection method</td>
<td>Line-95.65% Word-92.56%</td>
</tr>
<tr>
<td>segmentation based on information energy</td>
<td>Line-95% Character-94%</td>
</tr>
<tr>
<td>Connectivity Strength Function</td>
<td>Line-97.30%</td>
</tr>
<tr>
<td>Hough transform</td>
<td>Line-98.9% Word-99.1%</td>
</tr>
</tbody>
</table>

V. Mathematical Model

The mathematical model for proposed architecture is as follows

Let S be a system, \( S = \{s, e, I, O, P\} \)

Where,

s = Start of the program
1. Upload the handwritten document image file.
2. Preprocessing
3. Segmentation
4. Feature Extraction
5. Classification
6. Post-processing

e = End of the program.

\[
F(x) = F(1 \rightarrow O/P) 
\]

\( F(x) \) is a function, mapping set of input images \( I \) to set of output strings \( O \), given a set of operations \( P \).

The definitions of sets are as follows:

I= Input of the program (Image file)
O= Output of the program (String)
P = \{P_1, P_2, ..., P_n\}

Where \( P_1 \ldots P_n \)

are the operations followed in a linear fashion.

VI. Experimental Results

Fig. 5. Input Image
VII. CONCLUSION AND FUTURE WORK

The work deals with the study of various techniques which are used in optical character recognition for cursive handwriting recognition. For segmentation of cursive handwriting, horizontal and vertical projection based methods are used. Convex hull algorithm is used for feature extraction and for classification and recognition, SVM is used. The experiments are carried out on CCC benchmark dataset.

Future work can be done by improving the recognition accuracy and speed in much more better way. It can be improved further to get accurate result in noisy environment. Document retrieving can also be done as future work.

VIII. APPLICATION

The main thrust area of the project is medical domain, in which OCR can be used for recognizing the medicines name from the doctor’s prescription. The other applications can be Historical document recognition, Automatic reading of bank cheques, Automatic postal code identification, Converting handwriting in real time, Extracting data from filled in forms etc.

IX. ACKNOWLEDGMENT

It is a great sense of satisfaction that I represent my real venture on practical computing in the form of project work. I have successfully completed the work under the guidance of Prof. Reena Kharat and its my first duty to express my hearty gratitude to her. I also offer a sincere thanks to Mr. Vivek Kulkarni sir for giving me the opportunity to work with Persistent Systems Limited, Pune for their project and guiding at every stage with valuable suggestions.

REFERENCES


[7] G. Louloudisa, B. Gatosb, I. Pratikakisb, C. Halatsisa, “Text line and word segmentation of handwritten documents” a Department of Informatics and Telecommunications, University of Athens, Greece Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research Demokritos, 15310 Athens, Greece
