Character Recognition for Cursive English Handwriting to Recognize Medicine Name from Doctor's Prescription

Pritam S. Dhande
Dept of CE, PCCOE
Pune,India.
dhandespritam@gmail.com

Reena Kharat
Dept of CE,PCCOE
Pune,India.
reenakharat@yahoo.com

***Abstract*—This paper aims to represent the work related to character recognition for cursive English handwriting recognition. Recognition of cursive characters is very challenging because the characters are connected to each other. In the proposed architecture, horizontal projection method is used for text-line segmentation and vertical projection histogram method is used for word segmentation. Convex hull algorithm is used for feature extraction and SVM is used for classification. Proposed work is specifically for medical domain to recognize the medicine name from doctor's prescription. The segmentation experiments are carried out on different doctor's prescription samples and achieved 95% accuracy for text-line segmentation and 92% accuracy for word segmentation. The recognition accuracy of proposed system is 85%.**

Keywords*—Character recognition system; Segmentation; feature extraction; classification; convex hull.

I. INTRODUCTION

Character Recognition (CR) is a method which identifies and recognizes the text which stored in an image of the format jpeg, gif etc. The text in an image is converted into a machine readable format viz. ASCII or Unicode. The pixel representation of a letter is converted into its equivalent character representation using CR. It is the machine simulation of human reading. Character recognition is the study of how machines can observe the environment, distinguishes character of interest from their background and makes the correct decisions about the characters. CR system has numerous applications in pattern recognition, computer vision and artificial intelligence. The task of character recognition broadly separated into two categories:

- Hand-written documents and
- Machine printed documents.

The characters in machine printed documents are straight with uniform alignment and there is uniform spacing between the characters. While handwritten characters are not uniform and their size, shape varies because each person on the earth has different handwriting. In general, there are six steps in the process of character recognition of handwritten cursive script. These steps are as follows:

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

Handwritten text is an integral part of human life. In daily life whatever we write on blackboards, postal system, doctor's prescription etc. everything is handwritten. And character recognition problem is yet unsolved. Hence, there is a need of character recognition system which will accurately recognize the handwritten characters.

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 also a challenging task to recognize a particular handwritten character.

II. LITERATURE REVIEW

Abhishek Bal et al.[1] highlighted 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 projection and vertical projection that can segment the text-line 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% words 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.

978-1-5386-4008-1/17/$31.00 ©2017 IEEE

Kanchan Keisham et al.[2] proposed the text-lines segmentation based on information energy which is calculated for each and every pixel and for
classification and recognition, Artificial Neural Network(ANN) is used. The recognition has an accuracy of 92%.

Nibaran Das et al.[3] used the convex hull algorithm for feature extraction. For one character, total 125 features are computed by using different bays attributes of a convex hull. The experiments are carried on a dataset of Bangla basic characters and digits.The recognition rate is 76.86% for handwritten Bangla characters and 99.45% for Bangla numerals.

Subhadip Basu et al.[4] used multi-layer perceptrons (MLPs) for classification. To recognize the complex characters, feature set is designed by using three types of topological features i.e. longest-run features, modified shadow features and octant-centroid features. The experiments are carried out to segment and recognize Bangla characters from handwritten Bangla document.The recognition rate observed with the two-pass approach on the 300 samples is 80.58%.

Namrata Dave et al.[5] discussed various techniques which are used to segment the text-line. Three levels of segmentation i.e. text-line, word and character segmentation are explained. The various factors which affects the segmentation are explored.

Nafiz Arica et al.[6] proposed a segmentation and recognition algorithms, which is used for offline cursive handwriting recognition. For segmentation, first the image is converted into gray scale and then into binary image.Hidden Markov Model (HMM) is used for recognition of the characters .

G. Louloudisa et al.[7] discussed various segmentation techniques and levels of segmentation. Hough transform method is used for text-line and word segmentation of cursive characters which are connected to each other.

III. PROPOSED ARCHITECTURE

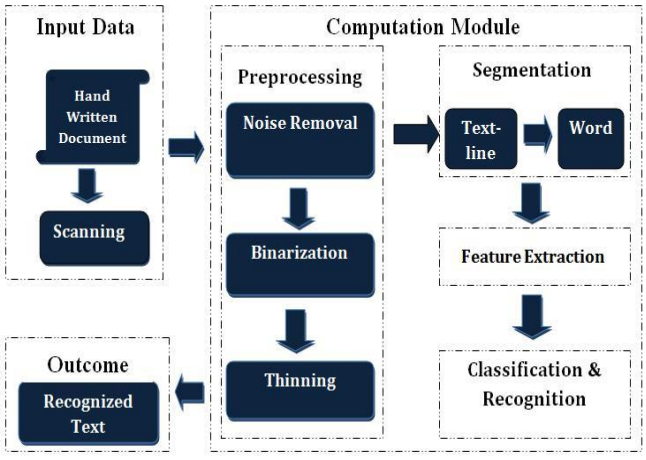

Fig. 1. Block Diagram of Proposed System

The input to the system is a Cursive English handwritten document. The image of a document is acquired by taking a photograph or scanning the document. Then pre-processing is performed on that image. First the noise is removed by using various filters and morphological operations. Then that image is converted into a binary image. After that thinning is performed on the binary image by using Guo-Hall algorithm to reduce the extra information present in the document and maintains only one pixel wide length.

Segmentation is carried out on preprocessed image. Line segmentation is carried out by horizontal projection histogram method and word segmentation is carried out by vertical projection histogram method. For feature extraction, convex hull algorithm is used. By using this algorithm, 125 features for each character is extracted. SVM is used for classification and recognition. The output of the proposed architecture is recognized text or string. In post-processing, the recognized word will be matched with the dictionary to improve the accuracy.

IV. MATHEMATICAL MODEL

F(x) = F (I→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)
- $O=\{C_1, C_2\}$
- C_1 = **{a,b,…..,z}**
- C_2 = **{A,B…..Z}**
- $P = \{P_1, P_2, \ldots . P_n\}$

Where $P_1..P_n$ are the operations followed in a linear fashion.

- P_1 = Image Acquisition
  - represented as M×N matrix

$$(f) = \begin{vmatrix} f_{0,0} & f_{0,1} & \cdots & f_{0,N-1} \\ f_{1,0} & f_{1,1} & \cdots & f_{1,N-1} \\ \cdots & \cdots & \cdots & \cdots \\ f_{M-1,0} & f_{M-1,1} & \cdots & f_{M-1,N-1} \end{vmatrix}$$

- P_2 = Filtering
  - Median filter
- P_3 = Binarization
  - f(i, j) **€ {0,1}**
- P_4 = Thinning
  - Guo-Hall Algorithm

- P_5 = Line Segmentation

- Threshold

$$Ti = \frac{1}{N} \sum_{i=1}^{N} (h1 + \ldots\ldots + hn)$$

for (i=1 to N)

(h_i >= T_i) then line is extracted.

- P_6 = Word Segmentation

- Threshold

$$Ti = \frac{TWD}{TGP}$$

for each i(1<=i<=sizeof(GAPSW[ ])),

if (GAPSW[i] >= Ti)

then inter-word gap

else

intra-word gap

- P_7 = Feature Extraction

- Area of convex hull

$$A = \frac{1}{2} \sum_{i=1}^{L} (X_i Y_{i+1} - X_{i+1} Y_i)$$

- Centroid calculation

$$C_x = \frac{1}{6A} \sum_{i=1}^{L} (X_i + X_{i+1})(X_i Y_{i+1} - X_{i+1} Y_i)$$

$$C_y = \frac{1}{6A} \sum_{i=1}^{L} (Y_i + Y_{i+1})(X_i Y_{i+1} - X_{i+1} Y_i)$$

- Success : S € O
- Failure : Words are not correctly recognized.

V. ALGORTHMS OF 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 M×N 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 (T_i) 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 r_1 and r_2 respectively. Extract the line segment between r_1 and r_2 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.

For word segmentation, first inter-word and intra-word distance is calculated. Vertical projection histogram calculates the threshold value. If the distance is less or equal to the threshold value then it is considered as intra-word gap. Vertical projection method is used for word segmentation [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(1<= i<= sizeof(GAPSW[ ])), if GAPSW[i]> = T_i then this gaps 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].For the given set of points in a plane, convex hull is the smallest convex polygon which contains all the points. Convex hulls have local controllability and used for affine matching.

The area A of a convex hull is given by

$$A = \frac{1}{2}\sum_{i=1}^{L}(X_i Y_{i+1} - X_{i+1} Y_i)$$

Where, L = Number of order vertices, (x_i, y_i) coordinates of the order vertices forming polygon.

The centroid (C_x, C_y) of the convex hull can be expressed 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 regions which are totally enclosed by an object.
- Bays : The regions which are present between the convex hull perimeter and the object.

Based on different bays attributes of a convex hull, 25 features are designed. A distance measure dcp is calculated. Dcp is the total count in number of pixels from the convex hull boundary pixel to the nearest character pixel in either horizontal or vertical direction. Dcp is calculated from the top , bottom, right and left boundaries of an image. Based on the dcp measure in a particular direction, six different topological features are calculated as follows:

- 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 as 125 i.e. 25 features for the overall image and 100 features in all four sub-images.

VI. EXPERIMENTAL RESULTS

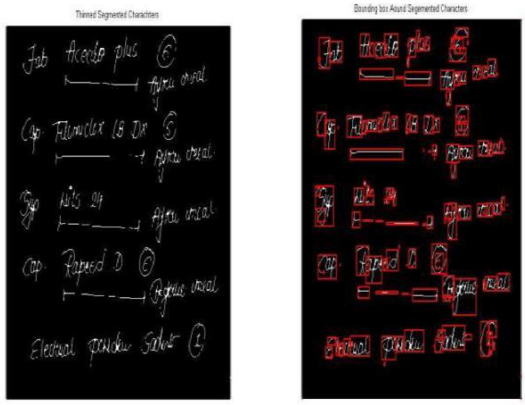

Fig.2.Pre-processed Image Fig.3.Segmented Image

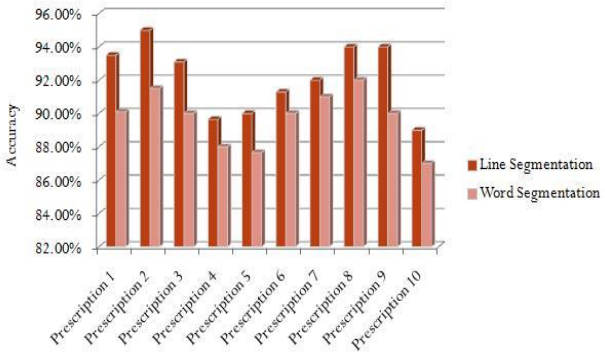

Fig.4. Segmentation result on 10 prescription samples

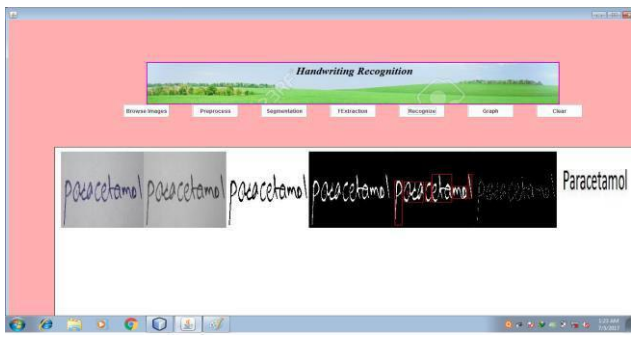

Fig.5. Recognition of a medicine name

VII. CONCLUSION AND FUTURE SCOPE

The work deals with the study of various techniques which are used in character recognition of cursive English handwriting. 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. Horizontal and vertical projection methods are implemented on 50 different prescription samples and achieved 95% accuracy for text-line segmentation and 92% accuracy for word segmentation . And the recognition accuracy of the system is 85%.

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. Future work is to develop a mobile application which will take the image of doctor's prescription and give reminder to the user for taking the medicine.

VIII. APPLICATIONS

The main thrust area of the project is medical domain, in which CR system 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 , For client application to get reminder for taking medicine etc.

REFERENCES

[1] Abhishek Bala and Rajib Saha, "An Improved Method for Handwritten Document Analysis using Segmentation, Baseline Recognition and Writing Pressure Detection",6th International Conference On Advances In Computing Communications, ICACC 2016, 6-8 September 2016, Cochin, India,Elsevier-2016.

[2] Kanchan Keisham and Sunanda Dixit,"Recognition of Handwritten English Text Using Energy Minimisation", Information Systems Design and Intelligent Applications, Advances in Intelligent Systems and Computing, Bangalore, India,Springer-2016.

[3] Nibaran Das,Sandip Pramanik,Subhadip Basu,Punam Kumar Saha, "Recognition of handwritten Bangla basic characters and digits using convex hull based feature set", 2009 International conference on Artificial intelligence and pattern recognition(AIPR-09).

[4] Subhadip Basu, Nibaran Das, Ram Sarkar, Mahantapas Kundu, Mita Nasipuri, Dipak Kumar Basu,"A hierarchical approach to recognition of handwritten Bangla characters", Elsevier-2009.

[5] Namrata Dave,"Segmentation Methods for Hand Written Character Recognition", International Journal of Signal Processing, Image Processing and Pattern Recognition Vol. 8, No. 4 (2015), pp. 155-164.

[6] Nafiz Arica, Student Member, IEEE, and Fatos T. Yarman-Vural, Senior Member, IEEE,"Optical Character Recognition for Cursive Handwriting", IEEE transactions on pattern analysis and machine intelligence, vol. 24, no. 6, june 2002.

[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, 15310Athens,Greece.

[8] Bilan Zhua, Arti Shivramb, Venu Govindarajub and Masaki Nakagawaa,"Online Handwritten Cursive Word Recognition Using Segmentation-free and Segmentation-based Methods", 978-1-4799-6100- 9/15/31.00-2015 IEEE.

[9] Hong Lee n, Brijesh Verma,"Binary segmentation algorithm for English cursive handwriting recognition "2011 ElsevierLtd.

[10] Subhash Panwara Neeta Naina,"A Novel Segmentation Methodology for Cursive Handwritten Documents", IETE JOURNAL OF RESEARCH,VOL 60-NO 6- NOVDEC 2014.