Novel Segmentation and Feature extraction Techniques for Handwritten Text Recognition.

Hemappa B.

Intel Technology India Pvt. Ltd.,
#136, 3rd Floor, Airport Road,
Bangalore-560017
India.

Email ID :hems_bs@hotmail.com

Dr.N.V.Subba Reddy

Professor and Head,
Department of Computer Science
Manipal Institute of Technology
MAHE (Deemed University)
Manipal-576 119, India.
Email Id: dr_nvsreddy@rediffmail.com

Abstract: *In this paper, we proposed novel slant detection and correction techniques, segmentation and feature extraction methods for handwritten text recognition. The segmentation and feature extraction module works on data, which is to be preprocessed. This is necessary because of the output of the scanned document is noisy and usually contains too many unwanted impression happened due to humidity and uneven ink distribution on paper. The novel techniques are used to detection and correction of any rotation that may occur during the writing or scanning process. The slant correction algorithm then generates a non_slant image by rotating the blocks, rather then the individual pixels. The proposed segmentation algorithm will count the block pixels in horizontal lines of the binaraized image and estimate the width and base line of each text line in the document. This ensures easy to segment the document into text of sentences. Each sentence again segmented into words. Each word processed and extract the prominent feature for further classification. The proposed system based on neural network feature extraction gives promising results in terms of classification and recognition.*

Keywords: *pattern recognition, Preprocessing, Trimming, Segmentation, Feature extraction, Slant detections, Slant correction, Neural network.*

1. Introduction

The area of handwriting recognition can be divided into online and offline [2,3]. In online recognition the writer is connected via an electronic pen or a mouse to the computer and the handwriting is recorded as a function of time. By contrast, in offline recognition the handwriting is captured by means of a scanner and becomes available for processing and recognition in the form of an image. Because of the availability of temporal information, online recognition is often considered the easier problem [1]. Automatic processing of handwritten forms requires form analysis [4, 5], field extraction [6,7], handwritten recognition, and so on. Texts/Lines separation has two non-trivial problems: one is to detect and remove the lines, and the other is to restore some character strokes distorted by the deleted lines [8]. A number of methods have been reported for sentences/lines separation [9, 10], but most of them can handle lines with in ±5° because of computational burden of their algorithm and fails in some exceptional situations where a straight line is curved a little or has a non-uniform thickness.

2. Developed system

The structure of the developed text recognition system shown in fig 1. Notice that both the segmentation and feature extraction modules works on data, which is preprocessed. This is necessary because of the output of the scanned document is noisy and usually contains too many unwanted impression happened due to humidity and uneven ink distribution on paper. The novel techniques are used to detection and correction of any rotation that may occurred during the writing or scanning process. The volume of the data to processed can be reduced and prominent features of patterns (Texts) can be extracted by incorporating trimming and removal of underlines, by which efficiency of preprocessor improves and it will have influence on the over all performance of the pattern recognition system(11). The proposed segmentation algorithm will count the block pixels in horizontal lines of the binaraized image and estimate the width and base line of each text line in the document. So that is easy to segment the document into text of sentences. And each sentence again segmented into words. Each words processed and extract the prominent feature for further classification and recognition system and to improve the accuracy of the classification, we will propose neural network based feature extraction system. The proposed

methods work well for multi lingual language documents and pattern recognition system.

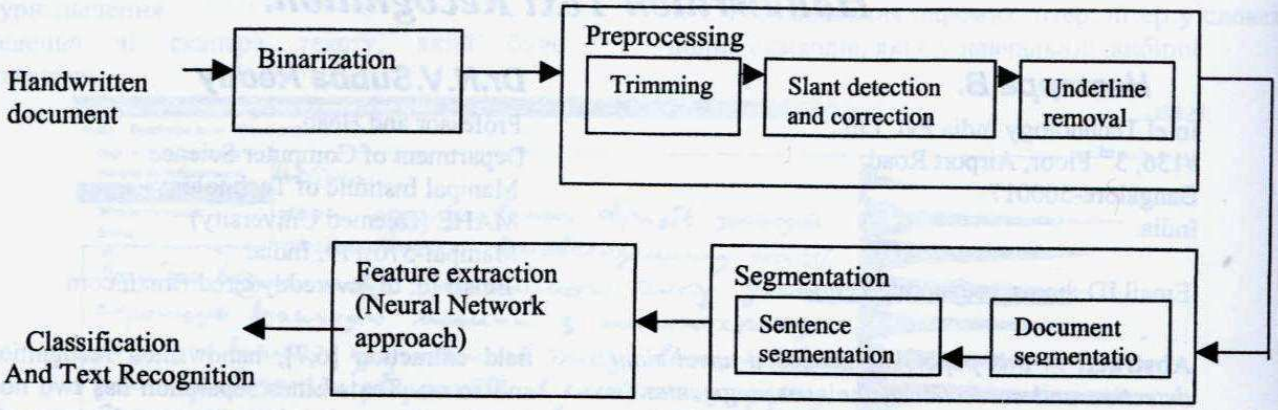

Fig 1. Overview of proposed model for printed and handwritten text recognition system.

3. Preprocessing and Normalization

The preprocessing model takes as input data and gives as output pattern containing the word to be recognized without any other element irrelevant to the recognition process. The tasks performed at this level depend on the data: when the document from which the word is extracted presents a background pattern, the latter should be removed. In some other cases, the words are extracted from forms showing boxes and lines that should be eliminated. For our data, binarization, trimming[11] the pattern to remove the unwanted impression happened due to humidity and uneven ink distribution on paper.

Once the image is preprocessed, it must be normalized. The normalization is deals with removing slant and slope of the document, line of texts and words. The slant detection and correction of the document images is particularly crucial among the document processing operations. Following algorithms explains the slant detection and corrections technique of document , sentence and words.

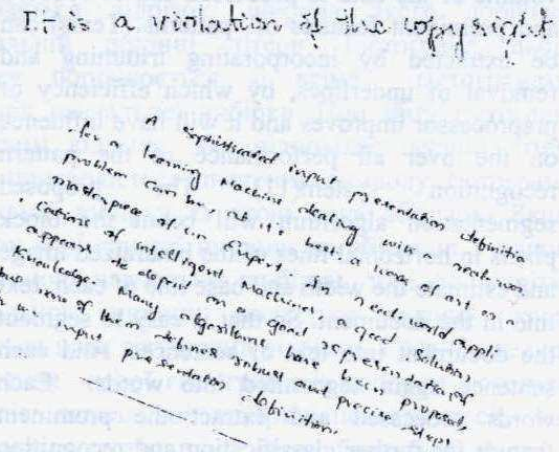

Fig 2. Sample inputs

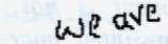

Fig 3. Slanted input

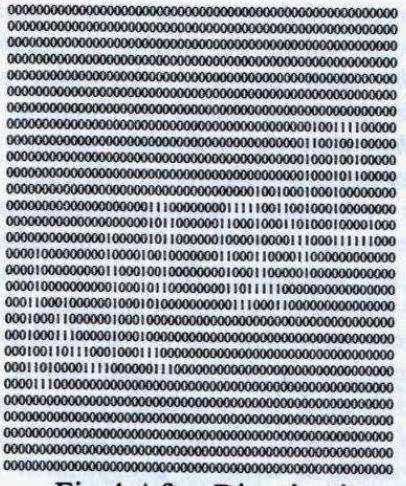

Fig.4 After Binarization

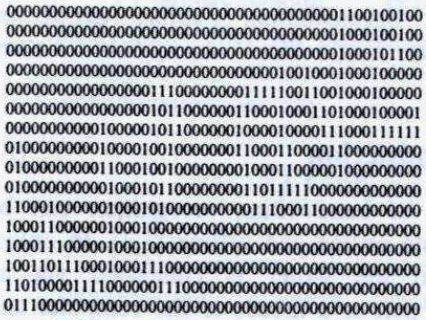

Fig 5 After Trimming

3. 1. Slant Detection Algorithm

The Slant detection algorithm takes preprocessed data as input and gives the degree of the slant towards right or left of the image.

Algorithm:

1. Initialize the TopLeft, TopRight, BottomLeft, BottomRight flags to zero.
2. Scan the 1/8th of binary pattern row wise from top to bottom.
3. if rows contains continues 0's
   A[I][j] = 1;
   Else
   A[I][j] = -1;
   Where I=0 … int(M x 1/8)
   J=0…N
   M= height of the pattern
   N = width of the pattern.
4. if each row of array A contains 1,s in left side , Set TopLeft =1 else TopRight =1;
5. Scan the 1/8th of binary pattern row wise from bottom to top.
6. Initialize the Array A to 0, Repeat the Step 3.
7. if each row of array A contains 1,s in left side , Set BottomLeft =1 else BottomRight =1;
8. if TopLeft = 1 and BottomRight =1 then Return LeftDownward Slant.
   Else if TopRight = 1 and BottomLeft =1 then return RightDownward Slant
   Else if((TopLeft=1 & bottomRight =0)or(TopRight=1 & BottomLeft =0) then return Partial slant.
   Else return no slant.

3.2 Slant correction

Algorithm :

1. Assign the pattern into M x N matrix
2. if (LeftDownward slant)

```
for(I=0;I<M*0.5;I++){
 for(j=0;j<N;j++){
   if(patten[I][j]! = 1)
    pattern[I][j]=pattern[I+1][j];
  else
  break;
 }
}
```

3. else if(RightDownward slant)

```
 for(I=0;I<M*0.5;I++){
 for(j=N;j>0;j--){
   if(patten[I][j]! = 1)
    pattern[I][j]=pattern[I+1][j];
  else
  break;
 }
}
```

```
00000000000000000111000000001111100100011001111 00
00000000001000001110000000011111001000110010010 0
01000000001000001110000000011111001000100010010 0
01000000001000001110000000011011001000100010110 0
11000000001000001110000000010001001000100010000 0
11000000001000001011000000100001000011100011111 1
11000100000010001010000000001110001100000000000 0
10001100000010001000000000000000000000000000000 0
10001110000010001000000000000000000000000000000 0
10011011100010001110000000000000000000000000000 0
11010000111100000111000000000000000000000000000 0
01110000000000000000000000000000000000000000000 0
```

Fig 6. After Slant correction.

4. Segmentation

Text regions posses a unique texture because they typically follow a specific arrangement rule: each regions consists of text line of the same orientation with approximately the same spacing between them and each text line consists of characters of approximately the same size. The above observations suggest that the primary component of the page layout – Text lines can be discriminated using the following novel segmentation algorithm.

4.1 Page segmentation.

Algorithm:

1. Scan the binarized data row wise.
2. Count the number of on cells in each row.
   D[i]= OC;
   Where I = 0…R
   OC =Total number of on Cells in a row.
   R =Total number of rows in page.
3. Mark the minimum or no on cells rows.
   If (D[I] = 0 || D[I]< Minimum)
   M[j]=I;
   J= 0 …R/4;
4. Find the width between the lines and separate the line of text.

4.2 Segmenting the text lines into words

To find the word boundaries within a text, We first set the distance between each words in a text line. Scan the binarized text line pattern vertically to identify the space between words and divide the text line into individual words to use for feature extraction. Preprocessed text line gives as input to the text line segmentation, which trimmed and underline removed pattern.

1. Set the space between the words
2. Scan the text line pattern column wise
3. if the space between words is equal to assumed space. Then divide the word.

```
00000000000000000110        011111001000110011110 0
00000000001000001110        011111001000110010010 0
01000000001000001110        011111001000100010010 0
01000000001000001110        011011001000100010110 0
11000000001000001110        010001001000100010000 0
11000000001000001011        100001000011100011111 1
11000100000010001010        001110001100000000000 0
10001100000010001000
10001110000010001000
10011011100010001110
11010000111100000111
01110000000000000000
```

Fig 7. segmented and trimmed words

5. Feature extraction

In most systems, the word image is segmented into small parts supposed to be the basic information units. In our feature extraction, preprocessed whole word is segmented into 16 numbers of regions, extraction of the feature from each region and generates the feature

vectors. This feature vectors are especially useful in classification of words using neural network system.

Preprocessed pattern divided into 16 numbers of regions.

Feature value = on cells /total number of on and off cels in a regions.

Feature vector = { G1, G2,G3,G4,G5,G6, D1, D2,D3,D4,D5,D6, M1,M2,M3}

1. Divide the preprocessed pattern into number of regions.

2.find the total number of On cells in each regions

3.Find the total number of ON and OFF cells in a region,

4.Calculate the feature value

Feature value = Total number of ON cell in a region/Total number of ON and OFF cells.

Repeat the step 4 for all the regions and generate the feature vector.

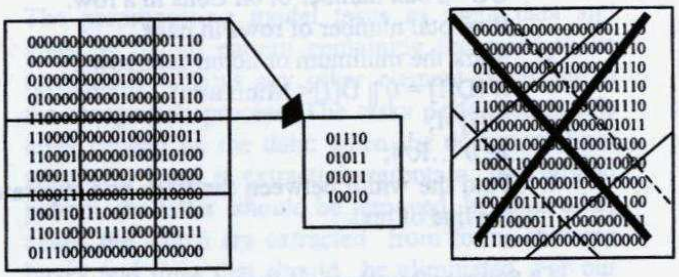

Fig 8. Segment the word into Grids(G1,G2..G6).
Doted lines segments the word as D1,D2,D3 regions, Solid lines divide as D3,D4,D5 regions
Solid Thick line segments M1,M2,M3,M4. regions

Conclusion:

The segmentation and feature extraction mechanisms developed are very useful in the recognition of text. The segmentation and feature extraction are two main components of the preprocessor. The overall performance of the text recognition system will have influence on better use of these two components. In this work proper care has been taken to improve the overall development of the text recognition system.

Reference:

1. U.V.Marti and H. Bunke, "*Using A Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System*," Inter Journal of pattern recognition and Artificial Intelligence, Vol.15, No. 1(2001) pp.65-90.
2. I Guyon, M. Schenkel and J.Denker, "*Overview and synthesis of on-line cursive handwriting recognition techniques*," Handbook of character Recognition and document Image Analysis, eds H.Bunke and P.S.P Wang, World Scientific, 1997, Chap 7, pp. 227-258.
3. J.C.Simon, "*Off-line cursive word Recognition*", Proc. IEEE80, 7(1992) 1150-1161.
4. A. Dengel and G.Barth, High level document analysis guided by geometric aspects, Int'l J. Pattern Recognition and Artificial Intelligence, Vol. 2, No. 4, pp 641-655, 1988.
5. T. Watanable, Q. Luo, and N.Sugie, "*Layout recognition of multi-kinds of table-form documents*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 17, no. , pp. 432-445, 1995.
6. F. Cesarini, M. Gori, S. Marinai and G.Soda, Data extraction from form images, Proc. Conference DEXA 95, pp. 438-448, London(UK), September 1995.
7. L.Y. Tseng and R.C. Chen, "*Recognition and data extraction of form documents based on three types of line segments, Pattern Recognition*", Vol.31,no. 10, pp. 1525-1540, 1998.
8. S.H Kim, S.H. JEONG, H.K. Kwag , "*Line Removal and Character Restoration using Bag Representation of Form Images*" ,Proceedings of seventh international Workshop on frontiers in handwritten recognition, Sept-2000, Amsterdam, ISBN 90-76942-01-3, Nijmegen: International Unipen Foundation, pp 43-52.
9. B. Yu and A.K. Jain, *A Generic system for form dropout*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 18,no 11, pp. 1127-1131, 1996.
10. M.D. Garris, "*Method and evaluation of character stroke preservation on handprint recognition*", Technical Report NISTIR 5687, July 1995.
11. Hemappa B, Dr.N.V.Subba Reddy, "*Preprocessing Technique for Handwritten Pattern Recognition*," National Conference on Intelligent and Efficient Electrical system, January 19-20,2001. No. II-5.