# Statistical Approach for Segmenting Unconstrained Handwritten Text lines

Gomathi Rohini. S
Department of Computer Science, Sri Ramakrishna Engineering College
Coimbatore 641022, India

Umadevi. R.S
GR Govindarajulu School of Applied Computer Technology
Coimbatore 641 004
India

Mohanavel. S
Dr. N.G.P. Business School, Dr. N.G.P. Institute of Technology
Coimbatore 641 048, India

## ABSTRACT
The segmentation of unconstrained handwritten text lines into words is an important stage in word recognition systems. This paper addresses a methodology to overcome the challenges, which are amplified by the non-uniform spaces between words and overlapping components by using a few statistical approaches. The system was developed using Java 2 and ImageJ tool. In this approach, a text line image is scanned vertically, holding only the spatial information. A scheme based on distance metrics and gap classification into inter-word gap and intra-word gap is presented. The threshold value is determined by using arithmetic mean, inter-quartile mean or trimmed mean based on the variation in the text. A pre-processing of removal of noise and correction of skew angle and dominant slant angle were done to improve the recognition accuracy. The system was illustrated with a few cases. A quantitative analysis of the experiment done on the system by using 1100 text lines from IAM database achieved an accuracy of 96.72% and found the system faster and reliable. Further, the proposed method is compared with the contour based and non-contour based techniques.

## General Terms
Document Image Processing.

## Keywords
Inter-quartile mean; projection profile; connected component; distance metrics.

## 1. INTRODUCTION
Segmentation of unconstrained handwritten text lines into words is an important preprocessing stage in handwritten document recognition, due to high variability and uncertainty of human writing style. Therefore, the efficiency of the document image analysis methods is affected by the precision of the word segmentation process. The approaches used for segmenting printed lines into words fail in the case of freestyle handwritten documents since the texts are curvilinear and gaps between the words are non-uniform. The proposed approach begins with preprocessing of scanned image of the handwritten text by enhancing some of the features and eliminating some of the inconsistencies to increase the accuracy of recognition. Variations like skew and slant angle, non-uniform spacing between words, existence of overlapping components, existence of punctuation marks and variable character size are a few challenges in word extraction. This paper introduces a relatively simple method, which is more tolerant to such cases. Most of the earlier methods considered a spatial measure of the gap between successive connected components (CC) and defined a threshold to classify within and between word gaps[6]. These measures are sensitive to shape of the CC like dominance of ascender and descender of a character.

This paper is organized as follows: Section 2 describes the related works, Section 3 describes the pre-processing and proposed segmentation method, Section 4 presents the experiments and results and Section 5 describes the conclusion and future work.

## 2. RELATED WORKS
This section briefs the review of related works carried by researchers on word segmentation of handwritten document images. Many approaches have been reported for the segmentation of unconstrained handwritten words. Histogram projection CCs, stroke bounding boxes, recognition based segmentation and holistic approaches are a few to mention. Most of the word segmentation approaches consider text line images to extract words. Each CC belongs to only one word and the gaps between words greater than the gaps between characters are the major assumptions made by the researchers. Marti and Bunke [1] employed the convex hull distance to estimate the gap metrics between successive CCs. This is calculated based on the horizontal distance between the left most and right most black pixels in each text line. Then the threshold is used to classify the candidate gaps into inter or intra words. In [3] and [4], a similar method was proposed to evaluate eight different spatial measures between the pairs of CCs for locating words in hand written text. It also proposed a method to combine the results of the minimum run-length with vertical overlapping of two successive CCs. The method used in [4] was based on a scale space approach for noisy historical documents. The line image was filtered with an anisotropic Laplacian at different scales to produce blobs, which correspond to portions of characters at small scales and words at larger scales. It has shown that the optimum scale is equal to 10% of the text-line height. The method was applied on George Washington's 100 manuscript samples. The error rate was 17%. The segmentation method in [5] was based on a gap metric, which exploited the objective function of a soft-margin linear support vector measure to separate the successive CCs. In [7] and [8], the Euclidean distance between overlapped components was used as the distance metric and a threshold was calculated by using several characteristics of the whole document image. In [9], the use of Gaussian mixture modeling was presented for gap classification and combination of two different distance metrics for distance computation. The segmentation of text lines into words based on the cooperation among digital data and symbolic knowledge was suggested in [10]. In [11], contour of the word is considered to determine the word

boundary. It is analyzed along with the threshold for inter-word gaps to extract words with high confidence.

## 3. WORD SEGMENTATION

In this paper, a simpler method is used to segment the handwritten text lines into words. In this approach, a text line image is taken as input and scanned vertically to remove the foreground pixels, holding only the spatial information. With the spatial information, the number of gaps and average gap width are calculated. Then the threshold value for all gaps in the text line is estimated using arithmetic mean, trimmed mean or inter-quartile mean (IQM) based on the size to classify the inter-character gap (ICG) and intra-word gap (IWG). When the gap is less than the threshold, then the gap is considered as ICG else it is IWG. The system was developed using Java 2 and ImageJ tool. Before word segmentation, the input image must be preprocessed. Instead of mapping the segmentation points directly into the image, the image is converted as a binary array. Using projection profile, gaps are identified and stored an array.

### 3.1 Pre processing

The pre-processing procedure concerns the removal of noise and correction of skew angle and dominant slant angle to eventually improve the recognition accuracy. A sample cursive line image is obtained from IAM database [2]. This gray scale image is converted into binary form [12]. The skew of the image is corrected by using the MATLAB function rotate() and its width and height are read by using standard Java function getWidth() and getHeight() respectively.

### 3.2 Word segmentation

The word segmentation begins from finding the distance between CCs. A CC can be a word or part of a word. At first, vertical projection profile is obtained for the pre-processed image. In order to compute the distance between successive CCs, vertical scan is performed from $0^{th}$ pixel to height of the image. If no black pixels are encountered, then the scan is denoted by 1, otherwise by 0. This process is continued throughout the width of the image. These scan results are stored in a one dimensional array called distance metric (DM) array, which holds foreground and background information of the image (see Figure 1).

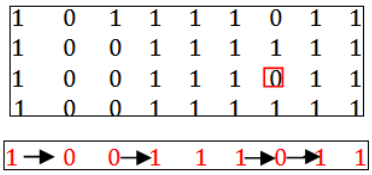


**Fig. 1: 2D binary representation of an image and its DM array**

In the DM array, the presence of 0 indicates CC and 1 indicates gap. To find the number of gaps (NG), total number of transitions from 0 to 1 or 1 to 0 in the array is divided by 2. The width of each gap ($g_w$) is the count of consecutive white runs and the sum total gap width GW is,

$$GW = \sum_{w=1}^{NG} g_w \qquad (1)$$

With this gap metrics, the computed gaps between adjacent CCs are classified into inter-word gaps or intra-word gaps. For the gap classification, a threshold for each text line is used. All distances above the threshold are considered as inter-word gaps, and below the threshold are considered as intra-word gaps. The proposed approach is illustrated with the following three cases:

#### *Case-1: Distance between characters is less than distance between words*

In this case (see Figure 2), the vertical projection profile for the input image is constructed (see Figure 3). With this, the DM array is extracted to identify the characters and gaps in the input image.

**Fig. 2: Inter-word gap greater than intra-word gap**

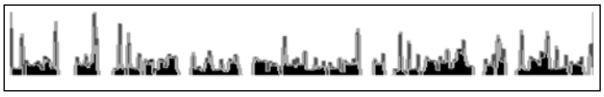

**Fig. 3: Vertical projection profile**

The gaps identified in the DM array are compared with the threshold ($T_{am}$), which is calculated using the formula given below.

$$T_{am} = 1/NG \sum_{w=1}^{NG} g_w \qquad (2)$$

If the gap is greater than the threshold $T_{am}$, then the point is marked as word boundary else it is ignored as gaps within the characters.

#### *Case-2: Distance between characters greater than or equal to distance between words*

In this case (see Figure 4), since the gaps between characters are greater than the gaps between words, it is difficult to classify gaps as "within" or "between" words. So arithmetic mean cannot be used as threshold.

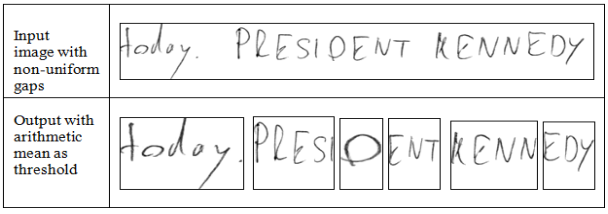


**Fig. 4: Word segmentation with arithmetic mean as threshold**

Due to the big variance in the gap width, some normalization is needed. For this case, IQM is used as threshold ($T_{iqm}$), which sorts the width and cuts off the two extreme values of gap width and gives a normalized mean that suits for this type of images containing non-uniform spaces (see Figure 5). The formula for IQM is given below.

$$T_{iqm} = 2/NG \sum_{w=\left(\frac{NG}{4}\right)+1}^{(3NG)/4} g_w \qquad (3)$$

**Fig. 5: Input image with non-uniform gaps**

Figure 6 shows the inter-word gap by filled rectangular box (orange) and intra-character gap by hollow rectangular boxes (red). The widths of these boxes are classified by comparing with the threshold value $T_{iqm}$.


**Fig. 6: Inter-word gap greater than or equal to intra-word gap**

*Case-3: Segmentation of Overlapping Components*

In case of denser and skewed documents, since the dominance of ascender and descender is more, IQM method fails to spot some valid segmentation points (see Figure 7).

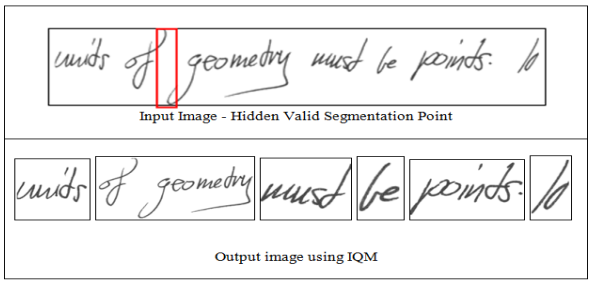


**Fig. 7: Word segmentation with IQM as threshold**

To overcome this drawback, the segmentation is restricted to core region only (Figure 9). For the given input image, horizontal projection profile is constructed (see Figure 8). Core region is plotted in the constructed profile, as the area in which the pixel distribution is greater than the threshold (see Figure 9). Here the trimmed mean ($T_{tm}$) is used as the threshold value for core boundary. The formula for trimmed mean is given below.

$$T_{tm} = \frac{1}{NG\text{-}2k} \sum_{w=k+1}^{NG-k} g_w \qquad (4)$$

where k is αNG and α is degree of freedom, which is chosen based on the number of gaps identified.

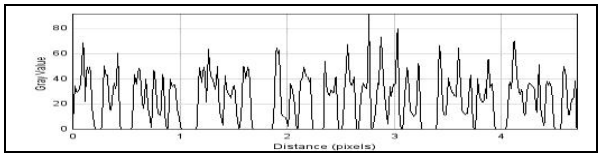


**Fig. 8: Horizontal projection profile**

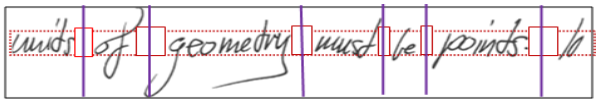

**Fig. 9: Word segmentation in core region**

For the identified core region, trimmed mean is calculated and set as threshold to classify the gaps between or within words. If the gap inside the core region is greater than the threshold then it is inter-word gap else it is intra-word gap. This might result in negligible loss of foreground pixels. But it has less impact in the recognition phase, since the pixels only in the ascender or descender may vanish.

Due to variability in spacing between adjacent words, two kinds of errors may occur. Error caused by merging adjacent parts of words, referred as under-segmentation is one and error caused by splitting a single word into two or more words, referred as over-segmentation is another one. In some cases both may occur together. The under-segmentation error can be rectified by applying the proposed approach in a recursive manner, whereas over-segmentation can be rectified by validating the segmentation points.

## 4. Experiments and Results

### 4.1 Dataset

In this section, we present a quantitative analysis of our approach over a large database. The IAM handwriting database contains forms of handwritten English text, which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments. It contains scanned unconstrained handwritten text forms, at a resolution of 300 dots per inch with 256 gray levels and saved as PNG images. Table 1 below shows some statistics of IAM Database 3.0. The Figure 10 below shows some samples of words.

**Table 1. Statistics about IAM database 3.0**

| IAM Database 3.0 | Pages | Sentences | Text lines | Words |
|---|---|---|---|---|
| | 1539 | 5685 | 13353 | 115320 |

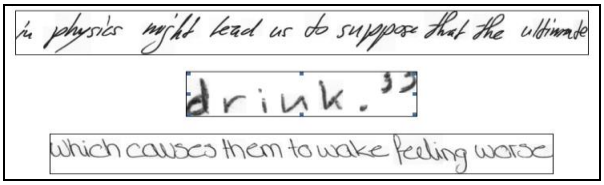

**Fig. 10: Handwriting samples in IAM database**

### 4.2 Implementation

We have implemented the proposed approach in Java platform on a PC with Intel core 2 duo, 1.6GHz, 2 GB RAM and Windows XP and evaluated its performance. The approach for segmenting the words has successfully solved problems like variation in gaps "between" and "within" words. It shows that, even when dealing with overlapped characters, the approach segmented properly

The experiments were performed on the handwritten documents, randomly selected from the IAM database. The segmentation system incorporates the above three cases. The common mistakes concerned with punctuation marks are rectified and correct outputs were produced as ground truth. For all the images, corresponding ground truth in terms of text lines was described and segmentation results were manually checked for errors (see Figure 11). Experimental results showed the improved performance of projection profile, when integrated with suitable statistical approaches for distance metrics and gap classification (Table 2).

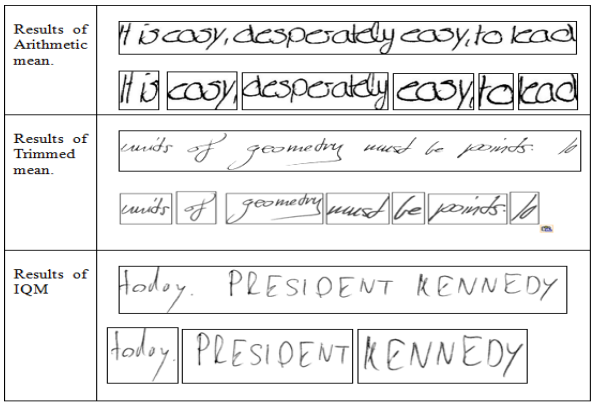


**Fig. 11: Results of the word segmentation**

We compared our approach with median white run-length as the threshold, in segmentation of handwritten line samples taken from IAM database. From the table below, we infer that the proposed approach performs slightly better than the other two.

**Table 2. Segmentation Accuracy**

| Threshold Calculation by | Accuracy of Segmentation |
|---|---|
| Median white run-length (Convex hull) | 94.23% |
| Median white run-length (Bounding Box) | 96.35% |
| The proposed approach – IQM | 96.72% |

## 5. CONCLUSION AND FUTURE WORKS

In this paper, a new approach for unconstrained handwritten word segmentation has been presented and applied to the IAM database. The input text line image was pre-processed and the text lines were segmented into words. Vertical projection profile was integrated with the system to find distance metrics. Following this, gaps were classified by threshold estimation. Threshold value was determined by arithmetic mean, IQM or trimmed mean. The results obtained by this segmentation, thus show that the system, capable of locating accurately the words in text lines.Future work mainly concerns the improvement of this word segmentation method, by using feedbacks from character segmentation and recognition modules.

## 6. REFERENCES

[1] Marti. U.V. and Bunke, H. 2001. Text line segmentation and word recognition in a system for general writer independent handwriting recognition. In: Proceedings of International Conference on Document Analysis and Recognition, 159-163.

[2] Marti, U.V. and Bunke, H. 2002. The IAM-Database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, 5, 39-46.

[3] Seni, G. and Cohen, E. 1994. External word segmentation of offline handwritten text lines. Pattern Recognition, 41-52.

[4] Manmatha, R. and Rothfeder, J.L. 2005. A scale space approach for automatically segmenting words from historical handwritten documents. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 1212-1225.

[5] Papavassiliou, V., Stafylakis, T., Katsouros, V., and Carayannis, G.: Handwritten document image segmentation into text lines and words. Pattern Recognition, 43, 369-377.

[6] Simistira, F., Papavassiliou, V., Stafylakis, T., Katsouros, and V., Carayannis, G. (2011). Enhancing handwritten word segmentation by employing local spatial features. In: Proceedings of International Conference on Document Analysis and Recognition, 1314-1318.

[7] Louloudis, G., Gatos, B., Pratikakis, I. and Halatsis, C.: Line and word segmentation of handwritten documents. In 1st International Conference on Frontiers in Handwriting Recognition (ICFHR), 247-252.

[8] Louloudis, G., Gatos, B., Pratikakis, I., and Halatsis, C.: Text line and word segmentation of handwritten documents. Pattern Recognition Journal. Special issue on Handwriting Recognition.

[9] Louloudis, G., Stamatopoulos, N., and Gatos, B. 2009. A Novel two stage evaluation methodology for word segmentation technique. In: Proceedings of International Conference on Document Analysis and Recognition, 686-690.

[10] Lemaitre, A., Camillerapp, J., and Couasnon, B. 2011. A perceptive method for handwritten text segmentation. Document Recognition and Retrieval, XVIII.

[11] Kuniawan, F., Khan, A. R., and Mohamad, D. 2009. Contour vs Non-Contour based word segmentation from handwritten textlines: an experimental analysis. In: International Journal of Digital Content Technology and its Applications, 3(2).

[12] Otsu, N. 1979. A Threshold selection method from gray level histograms. IEEE Transaction on Systems, Man and Cybernetics, 9(1), 62-66.