Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Language Independent Text-Line Extraction Algorithm for Handwritten Documents

**Nikita Vijay Borse, Prof. Imran R. Shaikh**
Department of CSE, S.N.D.C.O.E.R.C Yeola,
Dist-Nashik,Savitribai Phule University of Pune,
Maharashtra, India

***Abstract—Text-line extraction in handwritten documents is an important step for document image understanding, and a number of algorithms have been proposed to address this problem. In order to overcome this limitation, we develop text-line extraction algorithm for cursive handwriting. Our method is based on connected components (CCs), however, unlike conventional methods, we analysed strokes and partition under-segmented CCs into normalized ones. Due to this normalization, the proposed method is able to estimate the states of CCs for a range of different languages and writing styles.***

***Keywords— Connected Components, Text-Line Extraction , Trained Dataset.***

## I. INTRODUCTION

TEXT-LINE extraction in document images is an essential step for various document image processing tasks such as layout analysis and optical character recognition (OCR).Therefore, there have been a lot of researches in this area, and a number of algorithms have been proposed for the extraction of text-lines in machine-printed document images. However, text-line extraction in handwritten documents is still considered a challenging problem: the scale and orientation of characters are spatially varying, inter-line distances are irregular, and characters may touch across words and/or text-lines. Handwriting detection is a technique or ability of computer to receive & interpret intelligible handwritten input from source. Handwriting recognition is comparatively difficult, because different people have different handwriting style.

In optical character recognition, segmentation is a significant phase and accuracy of character recognition highly depends on accuracy of segmentation. Incorrect segmentation leads to incorrect character recognition. Segmentation phase includes text line, word, and character segmentation. Text line detection and separation in digital image documents is a challenging job for handwritten document analysis and character recognition. The problem becomes compounded if the text lines in the text image are connected or overlapped. Emergence of these problems is common in handwritten documents in comparison of printed documents because of individual's varying handwriting styles. Researchers are continuously working on these problems for different languages.

Text-line extraction in handwritten documents is an important step for document image understanding, we develop a language-independent text-line extraction algorithm. However, most conventional work focused on specific character sets. That is, conventional algorithms address the variations caused by individual writers by exploiting language- specific features. The situation is worse for Indian scripts where most characters are connected. On the other hand, character components are placed in a one-dimensional way in cursive Latin-based and Indian scripts, allowing us to develop horizontal bottom-up clustering rules.

Our method is based on connected components (CCs), however, unlike conventional methods; we analyze strokes and partition under-segmented CCs into normalized ones. Due to this normalization, the proposed method is able to estimate the states of CCs for a range of different languages and writing styles. From the estimated states, we build a cost function whose minimization yields text-lines. We develop an effective CC segmentation method: by partitioning under-segmented CCs into normalized ones, we can estimate states reliably in a variety of documents.

## II. LITERATURE SURVEY

This section describes the work done carried out by the various researchers so far in the field of handwritten text line detection in OCR. The observations from the work done so far have also been illustrated. The various issues related with text line segmentation in OCR are critically analysed in the literature survey and these help the researchers to understand and carry out the work further in this field. A wide variety of text line segmentation methods for handwritten documents has been reported based on projection profiles, Hough transform, smearing method, fuzzy run length and many others.

Nicolaou et al. (2009) proposed technique to segment handwritten document images into text lines by shredding their surface with local minima tracer. It is assumed that there exists a path from one side of the image to other that

 

traverses only one text line. Image is blurred first and then uses tracers to follow the white-most and black-most paths from both left to right and right to left direction in order to shred the image into text line areas.

Xiaojun Du et al. (2009) presented a new text line segmentation approach based on the Mumford–Shah model. The algorithm is script independent, use piecewise constant approximation of the MS model to segment handwritten text images. In addition, morphing is used by the author to remove overlaps between neighbouring text lines and connect broken text lines.

G. Louloudis et al. (2008) presented a text line detection method for handwritten documents. The proposed technique is based on a approach that consists of three distinct steps. The first step includes image pre-processing and connected component extraction, division of the connected component domain into three spatial sub-domains and average character height estimation. Secondly, author used a block-based Hough transform for the detection of potential text lines while third step is to correct feasible splitting, to detect text lines that the previous step did not expose and, finally, to disconnect vertically connected characters and assigns them to text lines.

Yi Li et al. (2008) proposed an approach based on density estimation and a state-of-the-art image segmentation technique, the level set method. A probability map is estimated from an input document image where each element represents the probability of the underlying pixel belonging to a text line. Then level set method is developed to determine the boundary of neighbouring text lines by evolving an initial estimate.

VassilisPapavassiliou et al.(2010) presented two approaches to extract text lines and words from handwritten document. The line segmentation algorithm is based on locating the optimal succession of text and gap areas within vertical zones using Viterbi algorithm. A text-line separator drawing technique is applied and then finally the connected components are assigned to text lines.

## III. EXISITING SYSTEM

Presently there are systems available for text extraction from input files. But all are working on the predefined dataset. Which means it is only suitable for printed text, if user tried to write a character in cursive handwriting or some new format, then system will not prompt an output or resulting in some wrong character in result.

Previous text line segmentation methods, such as connected component based methods, work directly on the input image (generally, a binary image). For connected component analysis, each pixel is treated equally and a change of one pixel may result in a significantly different result. If two neighboring text lines touch each other through even a single handwritten stroke, the segmentation algorithm fails. Hence we need to develop systems that overcome the drawbacks of existing system.

**Disadvantage** :

1. In handwritten documents, majority of writing patterns are not straight which cause problems in locating header line and base line. Space between lines is uneven.
2. Doesn't give output on handwritten text.
3. Accuracy is very low.

## IV. PROPOSED SYSTEM

We proposed a language-independent text-line extraction algorithm for the processing of handwritten document images. By introducing stroke lengths, we split under-segmented CCs into several pieces so that we can have better representations for text components. Then with the help of text line extraction we can estimate the line spacing and orientation of every CC.

We propose a system that will trace out trained data from input file. Input file is consisting of handwritten text and printed text. Text is separated using CC segment and then next in text line extraction, We estimate the line spaces and overlapping text lines. We using Optical Character Recognition algorithm. Resulting keywords are matched with the trained dataset. The keywords which pass the set threshold level of frequency will be added in output file as a result of handwritten recognition system. Trained dataset contains number of different samples of each handwritten character.

**Advantages:**

1. We are using Trained DATASET.
2. Dataset contains hundreds of samples for each & every character
3. Handwritten text is easily traced out.
4. Accuracy is increased.

## V. SYSTEM DESIGN

Fig.1 Represent the system architecture of handwritten as well as printed text. It consists of Input File. Input file is consisting of handwritten text and printed text. It is first recognize weather it is printed text or handwritten text. If it is printed text then text is extracted using Optical Character recognition.

If it is Handwritten Text then it is separated using CC segmentation, then text line is extracted, we get resulting keyword are match with Trained Dataset. Here we apply the threshold value that means we set the threshold level, if the matched character pass the set threshold level of frequency will be added in output file.

 

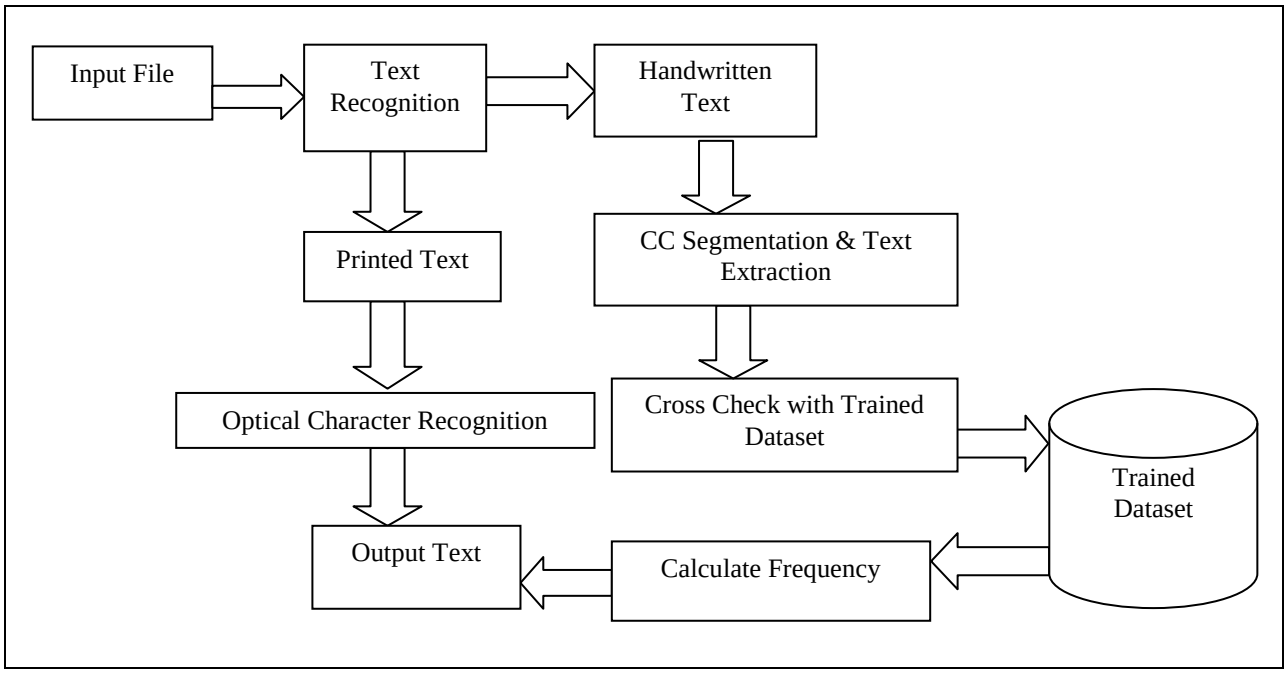


Fig 1: System Architecture

**CC segmentation:-**

Our method is based on connected components (CCs), however, unlike conventional methods; we analyse strokes and partition under-segmented CCs into normalized ones. The scales and orientations of characters (text components) differ between documents, and moreover, they are not fixed, even in a single document.

CC Partitioning consists of two steps:

I. To select CCs that should be segmented and

II. To partition selected CCs into smaller ones

**Text Extraction:-**

An English text line can be divided into 3 zones: Upper zone, Middle zone or Busy zone, and Lower zone. The Busy zone for an English text line is the zone between the mean line and the base line as shown in Fig. 2.

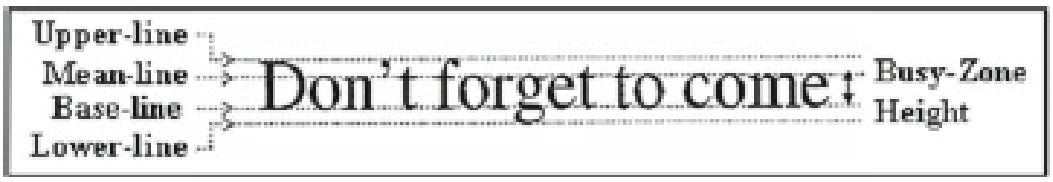


Fig. 2 Different parts of the text line

In text line extraction, we estimate line spacing and orientation of every CC.

**Module Description**

A. Text Extraction from printed Text.

B. Trained Dataset Creation.

C. Text Extraction from Handwritten Text.

***A. Text Extraction from printed Text***

Text extraction from printed documents is achieved using optimal character recognition system. Here we are using the Optical Character Recognition technique for getting the text from input documents**.** We propose a system that will trace out trained data from input file. Input file is consisting of handwritten text and printed text. Printed text is extracted by in build Optical Character Recognition library function. Optical Character Recognition is the electronic conversion of images of a text to the characters.

***B. Trained Dataset***

We proposed a language-independent text-line extraction algorithm for the processing of handwritten document images. Here we use Trained Dataset. Trained Dataset contains number of different samples of each handwritten character.

We have to add the digitized image of each character in different styles of writing techniques. (A-Z from different aspects)After adding that digitized images, it creates a trained dataset which we are using further for the text extraction from input documents.

 

Dataset Creation:

Draw Character in Corel draw
Convert into PNG format.
Collect & store all character into dataset.
Use this dataset as input.

Example :

Fig.3 Example of Trained Dataset of Character A

### *C.* Text Extraction from Handwritten Text

In this handwritten text extraction technique includes following points.

- De-skew – If the document was not aligned properly when scanned, it may need to be tilted a few degrees clockwise or counterclockwise in order to make lines of text perfectly horizontal or vertical.
- Despeckle – remove positive and negative spots, smoothing edges.
- Binarization –Convert an image from color or greyscale to black-and-white (called a "binary image" because there are two colors). In some cases, this is necessary for the character recognition algorithm; in other cases, the algorithm performs better on the original image and so this step is skipped.
- Line removal –Cleans up non-glyph boxes and lines Layout analysis or "zoning"
- Line and word detection –Establishes baseline for word and character shapes, separates words if necessary.
- Script recognition – We propose a system that will trace out trained data from input file. Input file is consisting of handwritten text and printed text.

## VI. RESULT

Text extraction from printed documents is achieved using optimal character recognition system. Here we are using the Optical Character Recognition technique for getting the text from input documents**.** Handwritten text is extracted by using CC segmentation and Text-Line extraction algorithm. First we apply CC segmentation then each character is separated. By using text-extraction we estimate line spacing. We get the resulting keywords, then this resulting keyword are matched with the trained dataset. If we found the character in trained dataset then, we apply threshold for frequency measure. We set the threshold level of frequency; keywords which pass the set threshold level of frequency will be added in output file as a result of handwritten recognition system.

## VII. CONCLUSION

We have proposed a language-independent text-line extraction algorithm for the processing of handwritten document images. By introducing the notion of stroke lengths, split under-segmented CCs into several pieces so that we can have better representations for text components.

Extraction of text lines from the handwritten/printed document images is one of the important associated problems of the OCR systems. Presence of skewed text lines, which is obvious in handwriting documents, always makes it difficult for accurate extraction of the text lines from the handwritten documents than that of the printed ones. In this context, the present work develops a simple and effective partitioning based text line extraction technique by estimating the line contours of the handwritten document images. The technique produces a reasonably good result as well as accuracy also increased. With the help of Trained Dataset we can get resulting matched character.

## ACKNOWLEDGMENT

Authors are cordially giving thanks to the researchers of different model for Text Extraction from printed Document, Text Extraction from Handwritten Document. All other who have tried hard to make their work easy to accomplish.

## REFERENCES

[1] G. louloudis, B. Gatos, I. Pratikakis, C.Halatsis, ”Text Line Detection in handwritten documents,” PatternRecognition vol.41, pp. 3758 – 3772, 2008.

[2] Yi Li, Yefeng Zheng, David Doermann, Stefan Jaeger,” Script-Independent Text Line Segmentation in Freestyle Handwritten Documents.” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 8, Aug.2008.

[3] Fei Yin, Cheng-Lin Liu, ”Handwritten Chinese text line segmentation by clustering with distance metric learning,” Pattern Recognition 42, pp. 3146 – 3157, 2009.

[4] Xiaojun Du, Wumo Pan, Tien D. Bui,” Text line segmentation in handwritten documents using Mumford–Shahmodel,” Pattern Recognition vol. 42, pp. 3136 – 3145, 2009.

[5] A. Nicolaou, B. Gatos, “Handwritten Text Line Segmentation by Shredding Text into its Lines,” 10th International Conference on Document Analysis and Recognition, IEEE Computer society, 2009, pp. 626-630.

[6] M. K. Jindal, R. K. Sharma, G. S. Lehal, “Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts,” International Journal of Computational Intelligence Research, Vol.3, No.4, pp. 277–286, 2007.

 

[7] Dhaval Salvi, Jun Zhou, Jarrell Waggoner, Song Wang, "Handwritten Text Segmentation using Average Longest Path Algorithm," Applications of Computer Vision(WACV), IEEE Workshop, pp. 505-512, 2013.

[8] VassilisPapavassiliou, ThemosStafylakis, VassilisKatsouros, George Carayannis," Handwritten document image segmentation into text lines and words," Pattern Recognition, vol. 43, pp. 369 – 377, 2010.

[9] Nikolaos Stamatopoulos, Basilis Gatos, Stavros J. Perantonis,"A method for combining complementary techniques for document image segmentation," Pattern Recognition vol. 42, pp. 3158 – 3168, 2009.

[10] Zhixin Shi, VenuGovindaraju, "Line separation for complex document images using fuzzy runlength," First International Workshop on Document Image Analysis for Libraries, p. 306, 2004.

[11] B. Gatos, A. Antonacopoulos, N. Stamatopoulos, ICDAR2007 handwriting segmentation contest, in: 9thInternational Conference on Document Analysis and Recognition (ICDAR'07), Curitiba, Brazil, Sept. 2007.

 