International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

Novel Technique for Layout and Handwritten Character Recoginization in OCR

Mehakanmol Singh[1], Lalit Mann Singh[2]
[1]Computer Science Department & Engg.,Shri Guru Granth Sahib World University, India
Email:mann898@hotmail.com
[2]Computer Science Department & Engg., Shri Guru Granth Sahib World University, India
Email:lalitmann19@gmail.com

Abstract: *In the document image analysis document segmentation is very important step. Document segmentation is the process in which we segment the document which contains the heterogeneous data means data like printed text, handwritten text, graph etc. We do the document segmentation because our optical character recognition system is unable to recognize the whole document with multiple data type so before the recognition we have to apply the document segmentation so to define the each region correctly. We would be using document segmentation on the handwritten bills which contain the heterogeneous content thereby segmenting the text and non- text region and the text into printed text and handwritten text and then we classify the text region into printed text and handwritten text. Information energy approach has been used to segment the text lines into rows that can be embedded into the notepad and command window later which help to save the bill copy in e-format.*

Keywords: Segmentation, Documentation, Layout segmentation, text segmentation, image segmentation

1. Introduction

Image processing system includes treating images as two dimensional signals while applying already set

signal processing methods to them. It is among rapidly growing technologies today, with its applications in various aspects of a business. Image Processing forms core research area within engineering and computer science discipline [1]. In imaging science, image processing is any form of signal processing for which the input is an image, such as a photograph or video frame; the output of image processing may be either an image or a set of characteristics or parameters related to the image. Most image-processing techniques involve treating the image as a two-dimensional signal and applying standard signal-processing techniques to it. Image processing usually refers to digital image processing, but optical and analog image processing also are possible. This article is about general techniques that apply to all of them. The acquisition of images is referred to as imaging [2].Image processing is referred to processing of a 2D picture by a computer. An image defined in the "real world" is considered to be a function of two real variables, for example,

a(x,y) with a as the amplitude (e.g. brightness) of the image at the real coordinate position (x,y).Modern digital technology has made it possible to manipulate multi-dimensional signals with systems that range from simple digital circuits to advanced parallel computers [3]. An image may be considered to contain sub-images sometimes referred to as regions-of-interest, ROIs, or simply regions. This concept reflects the fact that images frequently contain collections of objects each of which can be the basis for a region. In a sophisticated image processing system it should be possible to apply specific image processing operations to selected regions.

1.1. Segmentation: Segmentation is the process which done before the recognition process because it change the object into digital format and then divide it into the proper segment so that it is easy for the recognition to this object [5].
In the natural language processing we divide the segmentation into basically three parts:

a) Image segmentation: Image segmentation is to analysis the image and then partition the image into homogenous r in homogenous region. image segmentation is very important part of the image analysis because show the important and the interesting area of the image for example to segment the MRI image which contain the brain tumor with the help of the image segmentation we can the highlight the brain tumor part for which we are interesting [6].

b) Text segmentation: Text segmentation is the preprocessing step for the optical character recognition. Text segment is the process in which partition the text into the particular reason to easily recognizable. In the text segmentation we include the printed as well as handwritten text.

c) Document segmentation: Mostly of our document contain the heterogeneous data means in the text, image, and graphs together so for to recognition to this kind of table we need to segment the document to the particular region textual data to the textual reason and image to their image region so it is easy to recognition the document [7].

2. Review of Literature

Neumann[1],In this paper concerned about the formation of text line. It kept multiple segmentations of each character till context of each element is not known. At last stage, it calculated various parameter like Region text line positioning, Character recognition confidence, Threshold interval overlap. Then directed graph constructed with corresponding scores. Output of this graph was a word or a sequence of word. To eliminate typographical art fact, a pre-processing used with a Gaussian pyramid.

Shivakumara et al.[2],In this paper had purposed an idea based on GVF(Gradient Vector Flow) and neighbour component grouping that identified arbitrary oriented text from an video images. Sobel edge map of each input frame had been created. Then GVF had applied to find the dominant edge pixels and extracted the corresponding text components called TC (text component).This algorithm proceeded in two stages. First stage eliminate false positive from an image using skeleton concept depending upon junction point. The output named as CTC

(Candidate Text Component). In second stage, this CTC used for determining the orientation of the text. This method did not give any better result for less spacing in textline.

Vassilieva et al.[3],In this paper has proposed a new method of optical character recognition using hierarchical optimization algorithms. Mainly, the existing methods and algorithms for optical character recognition are not suitable for using them in industrial systems, i.e. they are not stable to defects and distortions of the recognized characters. Therefore we have developed a new algorithm which is based on the pattern character recognition algorithms and uses hierarchical optimization. The better recognition results obtained using the proposed algorithm give us a confirmation of a better aptitude of the approach for the industrial environment.

Giri[4],In this paper an algorithm segments regions of individual characters in a complex scene based on colourvariation. Author call it as bilateral regression since it works like bilateral filtering. It divides images pixel into two groups namely foreground and background. Formal group contains pixel corresponding to text and latter group contains other pixel that corresponds to non-text part. Next a word recognition step starts. In this phase, each image is numbered with a lexicon word. During labelling, equivalence classes obtained having three character classes. After that, A String Edit Distance is calculated. An image having least edit distance is given a lexicon word.

3. Documentation Page Segmentation

Document page segmentation is the most important step in the document image analysis to understand the significance of the document page segmentation in document image analysis we understand the concept of document image analysis.Now these days lots of data present on the document so the work of document image analysis to analysis the document which contain the textual and graphical data and extract the useful from it as the human does .there is some further part in document analysis like [11]

1. Document processing in this step we scan the whole document is to be scanned and with the help of document segmentation we can divide it into further textual part and graphical part.

2. Textual processing in this only the textual data is to be processed which contain printed as well as handwritten text. Further the textual processing is divide into two parts OCR and page layout analysis in the OCR we do the text recognition and page layout analysis we analysis the text blocks, text lines, paragraphs.

3.Graphical processing in this we contain graphical data like halftone images, drawing, image, table and process this data. it further divided into the line processing and region symbol processing. Line processing for straight lines and region processing for the filled region [12].

The segmentation of blocks can be handled at segmentation of address blocks on envelopes and mail pieces. It has two types of approaches.

1. Bottom up Approach

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

2. Top down Approach

Top down Approach: Top down method spilt the document images into text regions, text lines and text words and characters. It has some disadvantages that it don't work well on curve and overlapping text lines [7].

Bottom up Approach: This approach merge the small units of images like pixels and text into text lines and text regions. Bottom up grouping can be viewed as a clustering process, which aggregates image components according to proximity and does not rely on the assumption of straight lines. Bottom-up grouping is more intricate in computation than top-down partitioning. on the assumption of straight lines [8].

4. Proposed Methodology

Document segmentation is the preprocessing step in document image analysis step. Document segmentation basically works on the document layout and segment the document into text and non-text component which contain the multiple type of component. Document segmentation gives the homogenous region to the optical character recognition system for the recognition. Now these days it is very important to store the analysis the document because it can store the very important information so for to store and analysis the document we have to process the document and for the processing of document we need document segmentation.

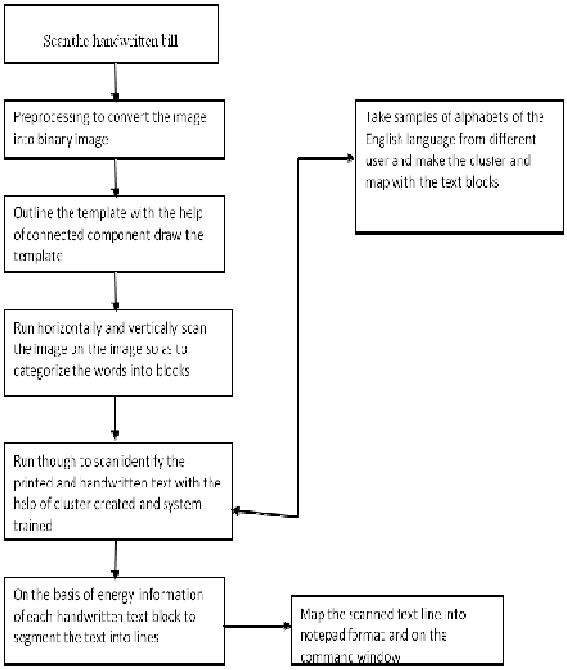

Fig. 1.1 Flowchart of Methodology

We enhance the document segmentation approach to segment the handwritten bill template which contains the heterogeneous component like image, printed text, handwritten text and the graphical image.

5. Experimental Results

The whole scenario is implemented by MATLAB.

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

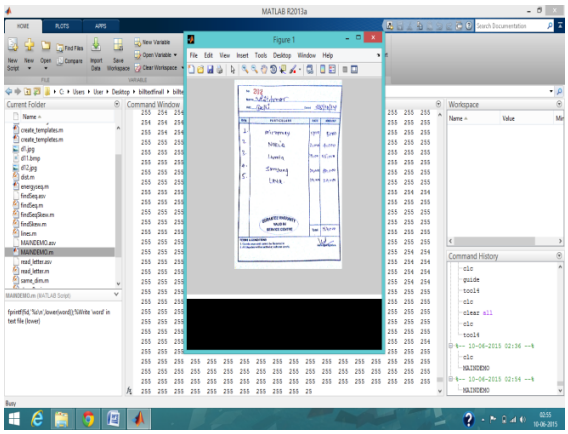

Fig. 5.1: Bill Input

As illustrated in figure 5.1, the input image is loaded. This is the image of handwritten bill. The loaded image will be segmented and after applying the segmentation, image layout will be extracted and handwritten characters are extracted .

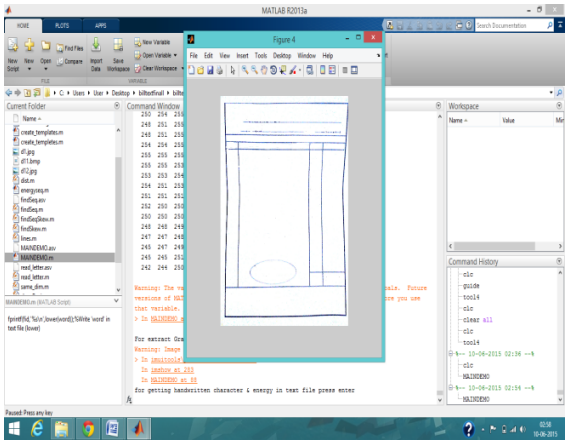

Fig 5.2: Layout Segmentation

As illustrated in figure 5.2, the input image is loaded which is the image of handwritten bill. The loaded image will be segmented and after applying the segmentation , image layout will be extracted and handwritten characters are extracted. The handwritten characters are segmented by applying character wise segmentation. When the characters are segmented, then layout of bill is detected and it will segmented according to the layout.

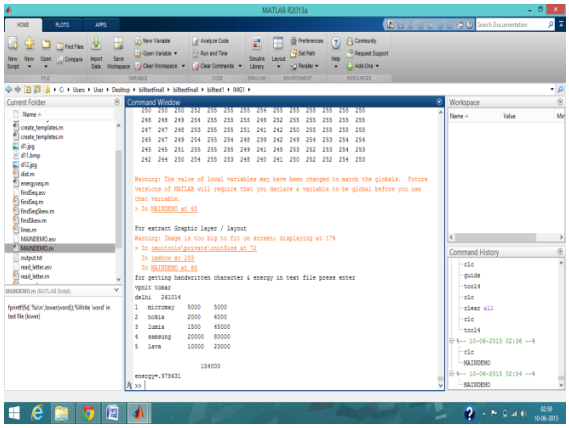

Fig 5.3: Handwritten Data Extraction

As illustrated in figure 5.3, the input image in loaded this is the image of handwritten bill. The loaded image will be segmented and after applying the segmentation, image layout will be extracted and handwritten characters are extracted. The handwritten characters are segmented by applying character wise segmentation. When the characters are segmented, then layout of bill is detected and it will segment according to the layout. The segmented characters will be detected which are in the handwritten bill and shown on the command window.

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 03 Issue 11
July2016

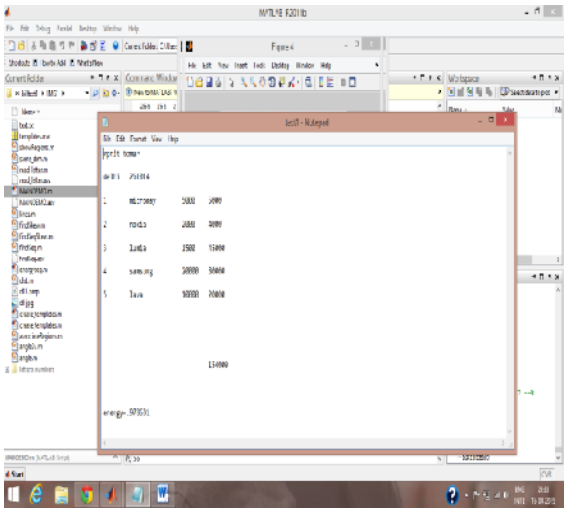

Fig.5.4 Data extracted in notepad

As illustrate in fig.5.4 first of all, data extracted from handwritten documents. After that data sets values are shown in notepad.

6. Conclusion

Document segmentation is the preprocessing step of document image analysis .It play very important role because it is difficult to recognize the document image. But there are lots of works to be done in document segmentation. In the present work, detection is not based upon layout. So it is difficult to detect characters. So main aim is to apply layout detection. In this propose method enhancement of the technique of document segmentation which contains the heterogeneous component is done and apply this segmentation technique on the dataset of shopkeeper bills. In this method segment the bills which contain only the column but with the help of our information energy line detection technique we can find the number of rows and after segmentation and recognition the handwritten text row we can map it on to the command window and notepad. To map the bill on to the command window and notepad for the further analysis make it the user specific approach.

In future, propose a technique which saves handwritten character in different and printed characters in other file

References

[1] L. Neumann (2013), "On Combining Multiple Segmentations in Scene Text Recognition," *in International Conference on Document Analysis and Recognition, 2013, pp. 1020-1024.*

[2] P. Shivakumara, T. Q. Phan, S. Lu, and C. L. Tan (2013), "Gradient Vector Flow and Grouping Based Method for Arbitrarily- Oriented Scene text Detection in Video Images,"*IEEETransactions on circuits and systems for video tecnology, vol. 23, no. 10, pp.1729-1739,2013.*

[3] N. Vassilieva and Y. Fomina (2013), "Text detection in chart images," Pattern Recognit. Image *Anal., vol. 23, no. 1, pp. 139–144, 2013.*

[4] P. S. Giri (2013), "Text Information Extraction and Analysis from Images using Digital Image Processing Techniques," *International Journal on Advanced Computer Theory and Engineering, vol. 3, no. 1, pp. 66–71, 2013.*

[5] P.Barlas, S.Adam, C Chatelaine and T Paquet (2014) "A typed and handwritten text block segmentation system for heterogeneous and complex document", publish in document analysis system, france.

[6] C.A Boiangiu, R.Laonitescu,M.CTanase (2014)"handwritten document text line segmentation based on information energy"publish in int j comput comm.,issn 1841-9836.

[7] Brucu Yidiz, katharina Kaiser and Silvia Miksch(2013) "A method to extract the table from PDF files"

[8] Ankush Gautam(2013)"segmentation of text from image document",publish in international journal of

computer science and information, Vol,4(3), 2013,538-540.

[9] Priyadharshini N, MS VIjaya (2013) "genetic programming for document segmentation and region classification using Discipulas", international journal of Advanced research intelligence,Vol.2,No.2

[10] Rahul Garg and Naresh Kumar Garg (2014), **"**Problems and Review of Line Segmentation of Handwritten Text Document", © 2014, *IJARCSSE* Volume 4 Issue 4,April 2014 ISSN: 2277 128X

[11] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu (2012) "Detecting Texts of Arbitrary Orientations in Natural Images," *in IEEE Conference on Computer Vision and Pattern Recognition, 2012, vol. 8, pp. 1083–1090.*

[12] A. J. Jadhav (2013), "Text Extraction from Images : A Survey," *International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no.3, pp. 333–337,2013. 5*

[13] P. K. Charles, V. Harish, M. Swathi, and C. H. Deepthi (2012) "A Review on the Various Techniques used for Optical Character Recognition," *Int. J. Eng. Res. Appl. 2, vol. 2, no. 1, pp. 659–662, 2012.*