Text Detection From Documented Image Using Image Segmentation

Santosh, Dr. Jenila Livingston L.M.

Research Scholar at VIT, Chennai.
Associate Professor, School of Computing Science and Engineering, VIT University - Chennai Campus, Vandalur - Kelambakkam Road, Chennai – 600127
M:+91 9444337733
Email id: santoshswami410@gmail.com, jenila.lm@vit.ac.in

Abstract: The image segmentation is typically used to trace the object and boundaries such as line and curves in an image. The segmentation of the text reliability is necessary to perform the classification and Recognition. The main aim of segmentation is to partition the document image into various homogeneous regions such as text block, image block, line and word. In this paper we have introduced a clustering based neighbor method and Direction based line segmentation method for the image segmentation. First, Read the input image and remove the noise. Second, apply the top down segmentation approach to segment a document image into text lines. Third, the result of segmentation is set of segments that collectively cover entire images. *

Keywords: Image segmentation, Histogram based Algorithm, Edge Detection algorithm Preprocessing, Image acquisition

I. Introduction

The Segmentation subdivides an image into its constituent region or objects. The level to which the subdivision is carried depends on the problem being solved. That is segmentation should stop when the object of interest in an application have been isolated. The segmentation of nontrivial Images is one of the most difficult tasks in image processing. Segmentation accuracy determines the eventual success or failure of computerized analysis procedures. The text character contain in the document image can be any gray scale value, low resolutions, variable size and embedded in complex background. many problems encountered in the segmentation, these includes the difference in the skew angle between lines, characters or even along the same text line, adjacent text line, overlapping words and touching characters.

1.1 propose

In this paper the segmentation is proposed in three stages:

- Line segmentation in which we identify the line in the documents
- Word segmentation in which we identify the words in the documents
- Character segmentation in which we identify the character in the documents

The goal of the segmentation is to simplify or change the representation of the image into something that is more meaningful and easier to analyze.

1.2 Scope

There are many algorithms are introduced for document image segmentation. This paper presents two algorithms for Document image segmentation, namely

- Direction based line segmentation algorithm.
- Clustering Based nearest neighbor method

II. Related Work

There are many document image segmentation algorithms some of those are:

2.1 Compression Based Algorithm

Compression based algorithms postulate that the optimal segmentation is the one that minimizes the overall possible segmentation, coding length of the data. The connection between these two concepts is that segmentation tries to find patterns in an image and any regularity in the image can be used to compares it. The algorithm describes each segment by its texture and boundary shape. This algorithm was implemented by W.J Teahan, Yingying Wen, Rodger Mcnab and Lan H [1].

2.2 Histogram based Algorithm

This algorithm was implemented by Tony Histogram based methods are very efficient when compared to other image segmentation methods because they typically requires only one pass through the pixel. In this technique, histogram is computed from the entire pixel in the image and the peaks and valleys in the histogram are used to locate the cluster in the image. Intensity can be used as the measure [2]. A refinement of this technique is to recursively apply the histogram-seeking method to cluster in the image in order to divide them into smaller clusters. This process is repeated with smaller and smaller cluster until no more cluster are formed. One of the disadvantages of histogram seeking method is that it may be difficult to identify significant peaks and valleys in the images. Selim Esedoglu, chan and kangyu Ni department of mathematic, University of Michigan using Wasserstein Distance. The Wasserstein distance between two functions is the least work that is required to move the region lying under the graph of one of the function to that of the other [3].

1.3 Edge Detection Algorithm

Edge detection is well developed field on its own within image processing. The region Boundaries and edge are closely related, since there is often a shape adjustment in intensity at the region boundaries. Edge detection techniques have therefore been used as the base of another segmentation technique. The edge identified by edge detection is often disconnected. To segment an object from an image however, one needs closed region boundaries. Salem Saleh Al-amril, Dr N.V kalyankar and dr.

khamitkar S.D implemented image segmentation by using Edge detection .They did a comparative study using seven technique of the edge detection segement.They are sobal,Roberts,canny, laplacian, krish and edge maximum technique on the Saturn original image and found that EMT and Perwitt techniques respectively are the best techniques for edge detection.

III Proposed Work

Graph partitioning methods can effectively be used for image segmentation. In these methods, the image is modeled as a weighted, undirected graph. Usually a pixel or a group of pixels are associated with nodes and edge weights define the (dis)similarity between the neighborhood pixels. The graph (image) is then partitioned according to a criterion designed to model "good" clusters. Each partition of the nodes (pixels) output from these algorithms are considered an object segment in the image. Some popular algorithms of this category are normalized cuts, random walker, minimum cut, isoperimetric partitioning and minimum spanning tree-based segmentation. Aleix M. Mart_ınez,a, Pradit Mittrapiyanuruk and Avinash C. Kak of Department of Electrical and Computer Engineering, The Ohio State University have implemented and suggested an alternative implementation of the k-way Ncut approach for image segmentation.The below mentioned algorithms have been implemented in the project

3.1. Partial Eight Direction Based Line Segmentation Algorithm (PEBLS)

In this section, a top down segmentation approach to segment an epigraphically document image into text lines is presented. The proposed method consists of three steps. Defining Base Lines and Supplementary Reference Lines, Portioning of Core text line regions and deriving non-linear paths.

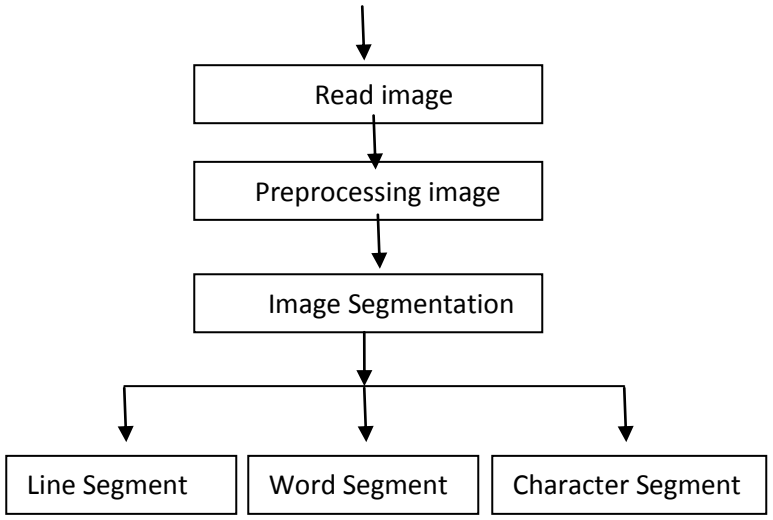
3.2. Nearest Neighbor Clustering Based Method (NNC)

In this section, a novel approach for line and character segmentation in an epigraphically script based on nearest neighbor clustering method is presented. The proposed algorithm scans the given input image from the left corner. When it encounters the first black pixel, it identifies the complete character through connected component. This character is segmented and placed at different location. The centered of the character is computed. Similarly the second character is identified and the centered is computed. The Euclidean distance between the centroids is computed to know whether the character belongs to the same line or next line. This is determined based on the threshold which is based on the assumption that the space between the text lines is greater than that between the characters. this way, the text lines and characters are segmented which could be used for the classification process. Mr. Praveen Dasigi applied the Spectral partitioning method to segment the documental images. The Segmentation uses a spectral partitioning approach that tries to maximize the proximities within the partitions while minimizing the proximities across them. This class of algorithms computes a pair wise similarity matrix built over every pair of components (pixels) from the image. The idea is to find an indicator vector from the spectrum of this matrix which can be threshold to partition the set.

3.3 Steps in Image segmentation

The general steps that are involved in Image segmentation systems are,

1. Image acquisition
2. Preprocessing
3. Segmentation



Fig; Architecture for image Segmentation

3.3.1 Image Acquisition

This is the stage where the image under consideration is taken. In the case of online recognition system a specialized hardware is implemented as explained earlier whereas for offline systems, the images are obtained either through a scanner or a camera. Whenever an image is acquired, there will be some variations in the intensity levels along the image. Also noise gets added to the image. Hence preprocessing is required for adjusting the intensity levels and to denoise the image.

3.3.2 Preprocessing

Preprocessing is the most important part of a better performing recognition system. In this stage, the acquired image is processed to remove any noise that may have incurred into the image during the time of acquisition or during the time of transmission. A colored image then it will be converted to a gray image before proceeding with the noise removal procedure. The denoised image is then converted to a binary image with suitable threshold.

3.3.3 Segmentation

Segmentation refers to a process of partitioning an image into groups of pixels which are homogeneous with respect to some criterion. Segmentation algorithms are area oriented instead of pixel oriented. The result of segmentation is the splitting up of the image into connected areas. Thus segmentation is concerned with dividing an image into meaningful regions. Image segmentation can be broadly classified into two types[5]

i. Local Segmentation: It deals with the segmenting sub images which are small windows on a whole image.

ii. Global segmentation: It deals with the images consisting of relatively large number of pixels and makes estimated parameter values for global segments more robust.

For character segmentation, first the image has to be segmented row-wise (line segmentation), then each rows have to be segmented column-wise (word segmentation). Finally characters can be extracted using suitable algorithms such as edge detection technique; histogram based methods or connected component analysis Connected component analysis is an algorithmic application of graph theory, where subsets of connected components are uniquely labeled based on a given heuristic. Connected component analysis is used in computer vision to detect connected regions in binary digital images, although color images and data with higher-dimensionality can also be processed. When integrated into an image recognition system or human-computer interface, connected component labeling can operate on a variety of information method.

IV. EXPERIMENTAL RESULTS

The different types of techniques used for image segmentation are discussed in the previous chapters. In this chapter the some of the experimental results obtained are shown. Graphical User Interface (GUI) are also showned

4.1 Image acquisition

The four images captured are shown in the figures 4.1, 4.2 and 4.3. The figures 4.1 is the shows the image of the printed characters (synthetic image). The printed test image is shown in the fig 4.2.These images are further processed according to the algorithm

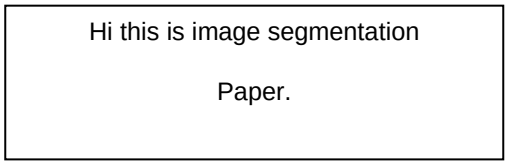


Figure 4.1: Captured image of printed text (synthetic

Figure 4.2: Captured image of printed text (Test image)

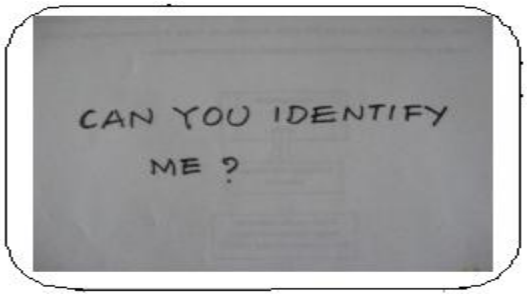


Figure 4.3: Captured image of Handwritten text

4.2 Pre-processing

The captured image inverted and it is cropped to the required size. The cropped image is converted into digital form. The pre-processed printed text (synthetic image) is shown in the figures 4.4. The preprocessed printed text (test image) and handwritten text images are shown in the figures 4.5 and 4.6 respectively.

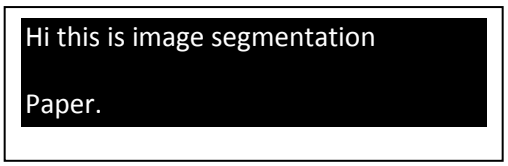


Figure 4.4: Pre-processed image of printed text (synthetic image)

Figure 4.5: Pre-processed image of printed text (test image)

Figure 4.6: Pre-processed image of Handwritten text

4.3 Segmentation

Images are segmented into line, word and character for the given preprocessing input image.

4.3.1 Line segmentation

The preprocessed images are segmented row-wise (line segmentation). The resulted images of the line segmentation for the figures 4.4, 4.5 and 4.6 are shown in the figures 4.7, 4.8 and 4.9.

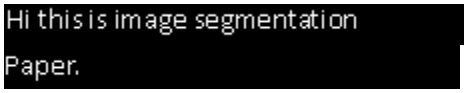


Figure 4.7: Line segmented image of printed text (synthetic image)

Figure 4.8: Line segmented image of printed text (test image)

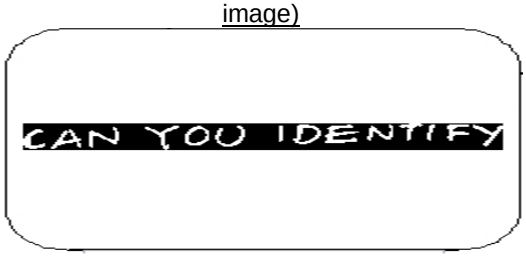


Figure 4.9: Line segmented image of Handwritten text

4.3.2 Word segmentation

In the line segmented image each word is segmented. The figure 4.10 shows the words segmented from the lines of the figure 4.7 (synthetic image of printed text). The word segmented images for the printed text (test image) and handwritten texts are shown in the figures 4.11 and 4.12.

Figure 4.10: word segmented image of Printed text (Synthetic image)

Figure 4.11: Word segmented image of Printed text (test image)

Figure 4.12: Word segmented image of Handwritten text

4.3.3 Character Extraction

The characters extracted from the word in the captured images are shown in the figure 4.13 to 4.16. These each characters are extracted using connected component analysis.

Figure 4.13: Characters extracted from a Printed text

Figure 4.14: Characters extracted from a printed text

Figure4.16: Characters extracted from a handwritten text

V. CONCLUSION

The proposed image segmentation method have been tested on a number of documented image and hand written, printed images. We use a set of quantitative evaluation measurements for the image segmentation .The system is designed in such a way that, the text in the documented image is detected and segmented automatically. Line segmentation is done by using horizontal projection profile and vertical projection profile analysis. Character segmentation is done by using Connected Component Analysis (CCA) and Vertical Projection Profile Analysis Experiments and results show that, this application yield 92.99% efficiency for line segmentation and 88.5% efficiency for character segmentation. Hence the future work includes this to be implemented for an online system. Also this has to be modified so that it works for both discrete and continuous handwritten characters simultaneously.

VI. REFERENCES

[1]. W.J Teahan, Yingying Wen, Rodger Mcnab and Lan H "A Compression-based algorithm for Chinese Word segmentation",**acl.ldc.upenn.edu/J/J00/J00-3004.**pdf

[2]. Orlando J. Tobias, Member, IEEE, and Rui Seara, Member, IEEE "Image Segmentation by Histogram Thresholding Using Fuzzy Sets"

[3]. N. Senthilkumaran and R. Rajesh**"** Edge Detection Techniques for Image Segmentation"

[4]. Jayarathna, Bandara, "A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation", IEEE Nov. 28 2006-Dec. 1 2006, 147 – 147.

[5]. Dr.-Ing. Igor Tchouchenkov, Prof. Dr.-Ing. Heinz Wörn, "Optical Character Recognition Using Optimisation Algorithms", Proceedings of the 9th International Workshop on Computer Science and Information Technologies CSIT'2007, Ufa, Russia, 2007

[6]. Robert Howard Kasse, "A Comparision of approaches to online handwritten character recognition", submitted to the department of EE&CS for the degree of Ph.D at MIT, 2005.

[7]. Jian and S. Bhattacharjee, "Text segmentation using gabor filters for automatic document processing," Machine Vis. Applicat., vol. 5, pp.169–184, 1992.

[8]. Keechul Jung, Kwang In Kim and Anil K. Jain, "*Text information extraction in images and video: A Survey"*, Elsevier, Pattern Recognition, vol.37 (5), pp 977–997, 2004.

[9]. Gonzalo A. Ruz, Pabltevez and Claudio A.Perez, "A neurofuzzy color image segmentation method for wood surface defect detection", *Forest Products Journal*, Vol.55, No.4, April 2005, pp.52-58.

[10]. Mausumi Acharyya and Malay K. Kundu, " Image Segmentation Using Wavelet Packet Frames and Neurofuzzy Tools", *International Journal of Computational Cognition*, Vol.5, No.4, December 2007, pp.27-43.

[11]. Ibrahiem M. M. El Emary, "On the Application of Artificial Neural Networks in Analyzing and Classifying the Human Chromosomes", *Journal of Computer Science*,vol.2(1), 2006, pp.72-75.

[12]. Bouchet A, Pastore J and Ballarin V, "Segmentation of Medical Images using Fuzzy Mathematical Morphology", *JCS and T*, Vol.7, No.3, October 2007, pp.256-262.

[13]. Mantas Paulinas and Andrius Usinskas, "A Survey of Genetic Algorithms Applicatons for Image Enhancement and Segmentation", *Information Technology and Control*, Vol.36, No.3, 2007, pp.278-284.

[14]. Xian Bin Wen, Hua Zhang and Ze Tao Jiang, "Multiscale Unsupervised Segmentation of SAR Imagery Using the Genetic Algorithm", *Sensors*, vol.8, 2008, pp.1704-1711.

[15]. Daniel L. Schmoldt, Pei Li and A. Lynn Abbott, "Machine vision using artificial neural networks with local 3D neighborhoods", *Computers and Electronics in Agriculture*, vol.16, 1997, pp.255-271.

[16]. M. Padmaja, J. Sushma, "*Text Detection in Color Images"*, International Conference on Intelligent Agent & Multi-Agent Systems,Chennai, 22-24 July 2009, pp. 1-6, 2009.

[17]. Mohieddin Moradi, Saeed Mozaffari, and Ali Asghar Orouji, **"***Farsi/Arabic Text Extraction from Video Images by Corner Detection"*, 6th, IEEE,Iranian conference on Machine Vision and image processing ,Isfahan, iran,2010. IEEE.

[18]. Dr.N.Krishnan, C. Nelson Kennedy Babu 2, S.Ravi 3 and Josphine Thavamani, "Segmentation Of Text From Compound Images", International Conference on Computational Intelligence and Multimedia Applications", Tamil nadu , India, vol. 3, pp.526-528, 2007.