Text Recognition from Documented Image using Image Segmentation

D.Mony Babu
PG Scholar, IARE, JNTUH, India

N. Chandra Sekhar Reddy, PhD
Professor CSE, IARE, JNTUH, India

ABSTRACT

The image segmentation is often used to trace the thing and bounds like line and curves in a picture. The segmentation of the text dependability is critical to perform the classification and Recognition. the most aim of segmentation is to partition the document image into varied homogenous regions like text block, image block, line and word. during this paper we\'ve got introduced a cluster based mostly} neighbor technique and Direction based line segmentation technique for the image segmentation. First, scan the input image and take away the noise. Second, apply the highest down phaseation approach to segment a document image into text lines. Third, the results of segmentation is about of segments that conjointly cowl entire pictures.

Keywords

Image segmentation, Histogram based Algorithm, Edge Detection algorithm Preprocessing, Image acquisition.

1. INTRODUCTION

The Segmentation subdivides a picture into its constituent region or objects. the extent to that the subdivision is carried depends on the matter being solved . that\'s segmentation ought to stop once the thing of interest in Associate in Nursing application are isolated. The segmentation of nontrivial pictures is one among the foremost troublesome tasks in image process. Segmentation accuracy determines the ultimate success or failure of computerised analysis procedures. The text character contain within the document image will be any grey value, low resolutions, variable size and embedded in advanced background. several issues encountered within the segmentation, these includes the distinction within the skew angle between lines, characters or maybe on an equivalent text line, adjacent text line, overlapping words and touching characters.

1.1 Propose

In this paper the segmentation is planned in 3 stages:

- Line segmentation during which we have a tendency to determine the road within the documents
- Word segmentation during which we have a tendency to determine the words within the documents
- Character segmentation during which we have a tendency to determine the character within the documents

The goal of the segmentation is to change or amendment the illustration of the image into one thing that's additional meaty and easier to investigate.

1.2 Scope

There are several algorithms introduced for document image segmentation. This paper presents 2 algorithms for Document image segmentation, namely

- Direction primarily based line segmentation formula.
- Clustering primarily based nearest neighbor technique

2. CONNECTED WORK

There are several document image segmentation algorithms, few of them are:

2.1 Compression based Algorithm

Compression primarily based algorithms postulate that the best segmentation is that the one that minimizes the general attainable segmentation, secret writing length of the info. The association between these 2 ideas is that segmentation tries to seek out patterns in a picture and any regularity within the image will be accustomed compares it. The formula describes every phase by its texture and boundary form. This formula was enforced by W.J Teahan, Yingying cyst, Rodger Mcnab and local area network .

2.2 Histogram based Algorithm

This formula was enforced by Tony bar chart primarily based ways ar terribly economical when put next to alternative image segmentation ways as a result of they generally needs only 1 suffer the picture element. during this technique, bar chart is computed from the whole picture element within the image and also the peaks and valleys within the bar chart ar accustomed find the cluster within the image. Intensity will be used because the live [2]. A refinement of this technique / this system is to recursively apply the histogram-seeking method to cluster within the image so as to divide them into smaller clusters. This method is recurrent with smaller and smaller cluster till no additional cluster ar fashioned. one among the disadvantages of bar chart seeking technique is that it should be troublesome to spot important peaks and valleys within the pictures. Selim Esedoglu, chan and kangyu metallic element department of mathematic, University of Michigan victimisation Wasserstein Distance. The Wasserstein distance between 2 operates is that the least work that is needed to maneuver the region lying underneath the graph of 1 of the function thereto of the opposite [3].

3. PLANNED WORK

Graph partitioning ways will effectively be used for image segmentation. In these ways, the image is shapely as a weighted, directionless graph. typically a picture element or a bunch of pixels ar related to nodes and edge weights outline the (dis)similarity between the neighborhood pixels. The graph (image) is then divided in step with a criterion designed to model \"good\" clusters. every partition of the nodes (pixels) output from these algorithms ar thought-about Associate in Nursing object phase within the image. Some fashionable algorithms of this class ar normalized cuts, random walker, minimum cut, isoperimetric partitioning and minimum spanning tree-based segmentation. Aleix M. Mart_ınez,a, Pradit Mittrapiyanuruk Department of Electrical and laptop Engineering, The Ohio State University have enforced and recommended an alternate implementation of the k-way Ncut approach for image segmentation.The below mentioned algorithms are enforced within the project

3.1. Partial Eight Direction primarily based Line Segmentation Formula (PEDPBLSF)

In this section, a prime down phaseation approach to segment Associate in Nursing epigraphically document image into text lines is conferred. The planned technique consists of 3 steps. process Base Lines and Supplementary Reference Lines, Portioning of Core text line regions and derivation non-linear ways.

3.2. Nearest Neighbor Clump Primarily Based Technique (NNC)

In this section, a completely unique approach for line Associate in Nursingd character segmentation in an epigraphically script supported nearest neighbor clump technique is conferred. The planned formula scans the given input image from the left corner. once it encounters the primary black picture element, it identifies the entire character through connected element. This character is segmental and placed at totally different location. The focused of the character is computed. equally the second character is known and also the focused is computed. The euclidian distance between the centroids is computed to grasp whether or not the character belongs to an equivalent line or next line. this is often determined supported the brink that is predicated on the idea that the house between the text lines is bigger than that between the characters. this way, the text lines and characters ar segmental that may well be used for the classification method. Mr. Praveen Dasigi applied the Spectral partitioning technique to phase the writing pictures. The Segmentation uses a spectral partitioning approach that tries to maximise the proximities among the partitions whereas minimizing the proximities across them. This category of algorithms computes a try wise similarity matrix designed over each try of elements (pixels) from the image. The idea is to seek out Associate in Nursing indicator vector from the spectrum of this matrix which may be threshold to partition the set.

3.3 Steps in Image segmentation

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

1. Image Acquisition

2. Preprocessing

3. Segmentation

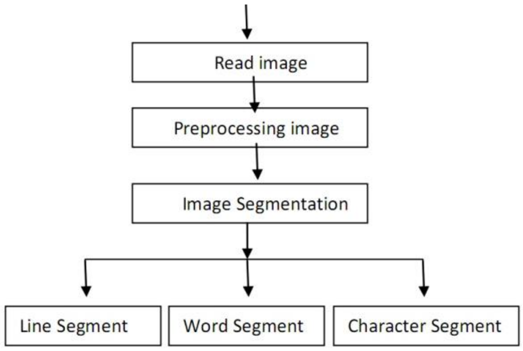

Fig 1: Design for Image Segmentation

3.3.1 Image Aquistion

This is the stage where the image into thought is taken. at intervals the case of on-line recognition system a specialised hardware is implemented as explained earlier whereas for offline systems, the pictures ar obtained either through a scanner or a camera. Whenever an image is nontransmissible, there'll be some variations at intervals the intensity levels on the image. put together noise gets superimposed to the image. so preprocessing is required for adjusting the intensity levels and to denoise the image.

3.3.2 Preprocessing

Preprocessing is that the foremost vital a district of a additional strong arts recognition system. throughout this stage, the nontransmissible image is processed to urge obviate any noise that may have incurred into the image throughout the time of acquisition or throughout the time of transmission. a colored image then it'll be converted to a gray image before continued with the noise removal procedure. The denoised image is then converted to a binary image with applicable threshold.

3.3.3 Segmentation

Segmentation refers to a technique of partitioning an image into groups of pixels that ar homogenised with connection some criterion. Segmentation algorithms ar area directed instead of constituent directed. The results of segmentation is that the rough of the image into connected areas. therefore segmentation is bothered with dividing an image into meaty regions. Image segmentation are going to be broadly classified into a pair of types.

i. Native Segmentation: It deals with the segmenting sub photos that are small windows on whole image.

ii. World segmentation: It deals with the pictures consisting of relatively sizable quantity of pixels and makes enumerable parameter values for world segments further robust.

For character segmentation, initial the image should be segmental row-wise (line segmentation), then each rows have to be compelled to be segmental column-wise (word segmentation). Finally characters are going to be extracted exploitation applicable algorithms like edge detection technique; bar graph based totally ways that or connected part Associate in Nursingalysis Connected part Associate in Nursingalysis is an algorithmic application of graph theory, where subsets of connected parts ar unambiguously labeled supported a given heuristic. Connected part analysis is used in portable computer vision to search out connected regions in binary digital photos, the color photos and information with higher-dimensionality might also be processed. once integrated into an image recognition system or human-computer interface, connected part labeling can treat a variety of data technique.

4. EXPERIMENTAL RESULTS

The different types of techniques used for image segmentation are mentioned at intervals the previous chapters. Throughout this chapter the variety of the experimental results obtained are shown.

4.1 Image Acquisition

The three pictures captured are shown in the following figures Fig 2, Fig 3 and Fig 4 respectively. Fig.2 shows the captured picture of the printed characters (Synthetic Picture), Fig 3 shows the captured image of test picture and Fig 4 shows the captured image of Transcription (Handwritten Text). These pictures are further processed as described in the algorithm.

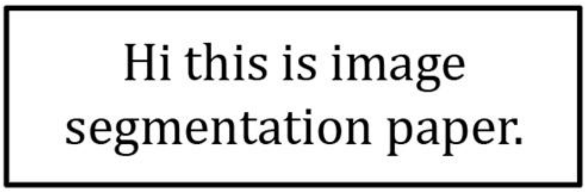

Fig 2: Captured Image of Printed picture

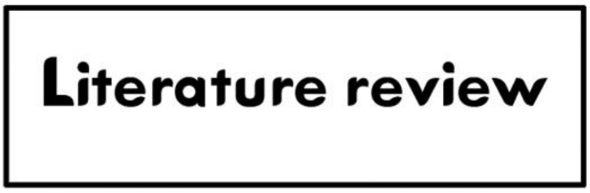

Fig 3: Captured Image of Test picture

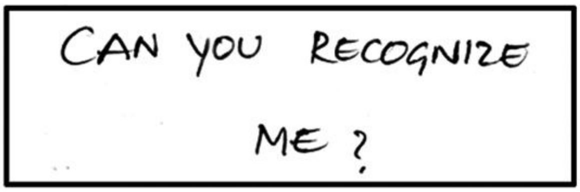

Fig 4: Captured Image of Hand Written Text

4.2 Pre Processing

The captured image inverted and is cropped to the required size. The cropped image is converted to digital kind. The pre-processed transcription (synthetic picture) is shown in Fig.5, The Preprocessed transcription (Test Picture) and Handwritten Text Pictures are shown in Fig.6 and Fig.7 respectively.

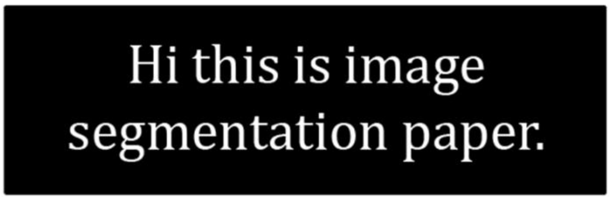

Fig 5: Preprocessed Image of Printed(Synthetic) picture

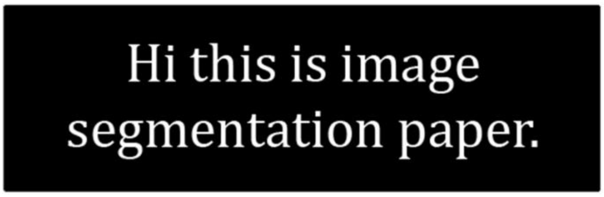

Fig 6: Preprocessed Image of Printed(Test) picture

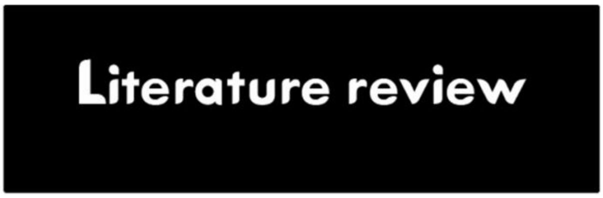

Fig 7:Preprocessed Image of Handwritten (Captured) picture

4.3 Segmentation

The above preprocessed input pictures are further segmented into line, word and character as shown below.

4.3.1 Line segmentation

The preprocessed pictures are segmented row-wise (line segmentation). The resulted pictures of the row segmentation for the figures are shown below.

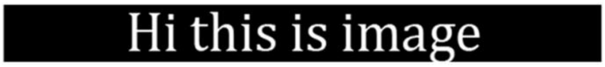

Fig 8: Line Segmented Image of Printed Text

4.3.2Row/ word segmentation

In the Row segmented image each image is segmented into words by considering space as a seperator as shown in the figures 9, 10 and 11.

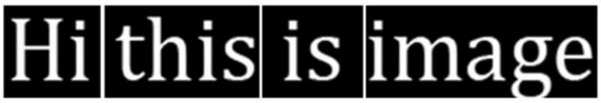

Fig 9: Word Segmented Image of Printed (Synthetic) Text

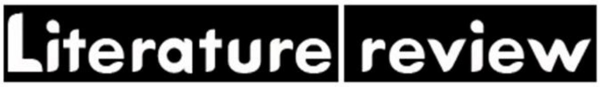

Fig 10: Word Segmented Image of Printed(Test Pic.) Text

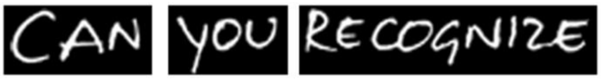

Fig 11: Word Segmented Image of Handwritten (Captured) Text

4.3.3 Character Extraction

The following figures 4m, 4n, 4p shows the characters which are extracted from the row segmented pictures are shown below.

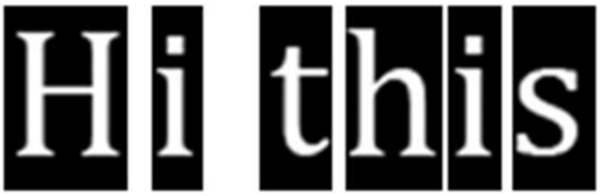

Fig.12: Extracted Characters from Printed (Synthetic Picture) Text

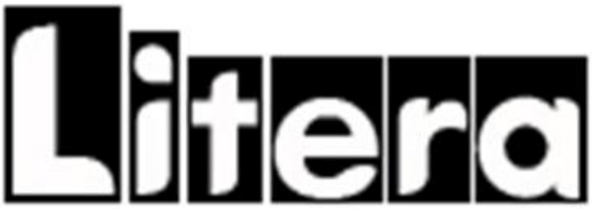

Fig.13: Extracted Characters from Printed (Test Picture) Text

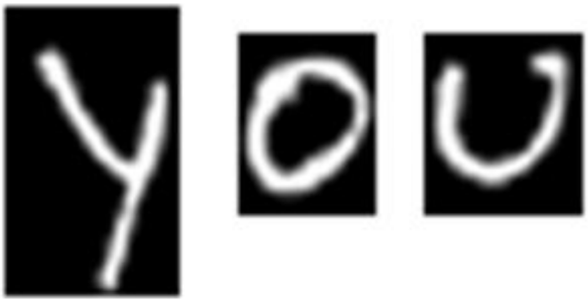

Fig.14: Extracted Characters from Handwritten (Captured Picture) Text

5. CONCLUSION

The planned image segmentation technique ar tested on form of documented image and hand written, written photos. we've got a bent to use a bunch of measure measurements for the image segmentation. The system is supposed in such how that, the text at intervals the documented image is detected and segmental automatically. Line segmentation is completed by exploitation horizontal projection profile and vertical projection profile analysis. Character segmentation is completed by exploitation Connected part Analysis (CCA) and Vertical Projection Profile Analysis Experiments and results show that, this application yield ninety 2.99% efficiency for line segmentation and eighty eight.5% efficiency for character segmentation. so the long haul work includes this to be implemented for an online system. put together this should be modified therefore it works for every distinct and continuous written characters at constant time.

6. 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 Thresholding Using Fuzzy Sets”

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

[4] Jayarathna, Bandara, “A Junction Segmentation Algorithm for Offline Handwritten Connected Character Segmentation”, 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, 2007

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

[7] Jian and S. Bhattacharjee, “Text using gabor filters for automatic document processing,” Machinecat.,Visvol. .5, ppAppli.169– 184, 1992.