DOCUMENT SEGMENTATION METHODS: A COMPREHENSIVE REVIEW

Ayush Sinha[1], Arpita Saha Biswas[2], Soumyadeep Mandal[3],Suparna Biswas[4], Koushik Pal[5], Anurima Majumdar[6]

[1,2,3] Student of Electronics and Communication Engineering, Guru Nanak Institute of Technology
[4,5,6] Faculty of Electronics and Communication Engineering, Guru Nanak Institute of Technology

Corresponding author: Suparna.biswas@gnit.ac.in

***Abstract*—** Document Segmentation Analysis (DSA) is concerned with transformation of any information presented on paper document into an equivalent symbolic representation, accessible to computer information processing. DSA basically deals with quantitative measurement from an image, to produce a description. The aim of document segmentation is to automatically recognize and extract textual or graphical information from digitized documents. Document Segmentations addresses the problem of separation of text and graphics and their identification and recognition. It is basically the process of extracting and representing information from the image or the process of partitioning a digital image into multiple segments or set of pixels. Different researchers have developed different segmentation techniques with successful results. In this paper we have reviewed different document segmentation techniques in details.

Keywords—Document Segmentation, Optical Character Recognition,

I. INTRODUCTION

Document segmentation and interpretation deals with the variety of documents such as letters, books, newspapers, magazines etc. Document segmentation analysis and understanding has become an important and a challenging research area. Segmentation is one of the most important operations in computer vision. The goal of image segmentation is the domain independent partition of the images into a set of regions which are visually distinct and uniform with respect to some property, such as gray level, texture or color [1].

Image segmentation[2][3] is an important aspect in most of the automatic pattern recognition and scene analysis problems. It is one of the difficult operations in image processing. In this process digital image is of partitioned into multiple regions or clusters. Image segmentation is typically used to locate objects of interest and boundaries such as lines, curves in an image.

Segmentation of text line from images is one of the major and complicated tasks in Character Recognition (CR) system. A Character Recognition system takes an image as input and generates a character set in editable form as a result. Most of the research works have already been done through last three decades regarding different aspects of Optical CR system. The research firstly started with printed text segmentation and recognition and the works have been extended to segmentation and recognition of handwritten texts. The main application of OCR was firstly to recognize scanned document images [4]. Different document segmentation techniques are available in the literature [5-11]. It includes top-down, bottom-up and mixed approach. Typically a successful approach must cope with as many variations (e.g. shapes of regions skew) in the document as possible. Document segmentation

techniques are widely used in medical purposes, traffic control system and agricultural processes, Satellite imaging.

Pixel-Based Segmentation: Pixel based or point based segmentation is simple to use. Pixel based segmentation results in a bias of the size of segmented objects when the objects show variations in their gray values.

Edge-Based Segmentation: Edge-based segmentation is based on the fact that the position of an edge is given by an extreme of the first-order derivative or a zero crossing in the second-order derivative.

Region-based Segmentation: Some parts are missing from point based segmentation techniques. These are taken care of in region based techniques.

Line-based Segmentation: A line segment is a part of a line that is bounded by two distinct end points, and contains every point on the line between its endpoints. A closed line segment includes both endpoints, while an open line segment excludes both endpoints; a half-open line segment includes exactly one of the endpoints.

Word-based Segmentation: The word segmentation procedure is divided into two steps. The first step deals with the computation of the distances of adjacent components in the text line image and the second step is concerned with the classification of the previously computed distances as either inter-word distances or inter-character distances.

Character-based Segmentation: Character Segmentation is the most crucial step for any OCR (Optical Character Recognition) System. The selection of segmentation algorithm being used is the key factor in deciding the accuracy of OCR system. If there is a good segmentation of characters, the recognition accuracy will also be high. Segmentation of words into characters becomes very difficult due to the cursive and unconstrained nature of the handwritten script.

Rest of the paper is organized as follows: Section II provides the literature review of different document segmentation methods. Section III concludes the paper.

II. LITERATURE REVIEW

In this paper different document segmentation technique like X-Y Cut technique, smearing technique, Fuzzy clustering technique, Artificial Neural Network Technique, Genetic Algorithm Technique, Hough transform techniques are reviewed.

A. X-Y Cut technique

The X-Y cut technique takes care of missed segmentation points to separate the connected components. There are two types of separation, horizontally and vertically. Vertical separation can be done by projection cut vertically direction in top or bottom until facing the border window of structure. If the projection hits black pixel before border windows, the projection stops and turns horizontally to both left and right directions. This projection continues until it hits the black or border of window. If it hits black pixel, it changes to

perpendicular direction. If it hits the border window, the projections stop. Horizontals separation also performs in the same sequence [11].

B. *Smearing technique*

For printed and binarized documents, smearing method such as the Run-Length Smoothing Algorithm can be applied. Consecutive black pixels along the horizontal direction are smeared: i.e. the white space between them is filled with black pixels if their distance is within a predefined threshold. The bounding boxes of the connected components in the smeared image enclose text lines. A variant of this method adapted to gray level images and applied to printed books from the 16th century exist in accumulating the image gradient along the horizontal direction. For this purpose, numerous adjustments in the method concern the tolerance for character alignment and line justification. Some foreground pixels may be skipped if their number does not exceed a predefined value. This matrix is threshold to make pieces of text lines appear without ascenders and descended. Parameters have to be accurately and dynamically tuned. [14]

C. *Fuzzy Clusterin gtechnique*

Clustering includes the task of dividing data into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as not similar as possible. Clustering can also be a form of data, compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. Examples of values that can be used as similarity measures include distance, connectivity, and intensity. In non-fuzzy or hard clustering, data is divided into crisp clusters, where each data point belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster, and associated with each of the points are membership which indicate the degree to which the data belong to the different clusters. The main idea behind fuzzy clustering is that an object can belong to more than one class and does so to varying degrees called memberships. The aim of a clustering is to divide a given set of data or objects into clusters, which represents a group. The partition should have two properties: the data which belong to one cluster should be as similar as possible that is called as Homogeneity inside clusters and the data which belong to different clusters should be as different as possible called as Heterogeneity between the clusters [2].

D. *Artficial Neural Network technique*

An Artificial Neural Network (ANN) is information that process is inspired by the way of nervous systems, such as the brain. The basic parts of this process are the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements working to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. Neural networks take a different approach to solving a problem that of conventional techniques. Conventional techniques use an algorithmic approach i.e. the

computer follows a set of instructions in order to solve a problem. Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements working in parallel to solve a specific problem [2].

E. Genetic Algorithm technique

Color images can increase the quality of segmentation, but also increase the complexity of the problem. A way of handling this complexity is to use genetic algorithms. Genetic algorithms are an optimization technique used in image segmentation. A characteristic of genetic algorithms is their effectiveness and robustness in dealing with uncertainty, insufficient information and noise. If the conception of a computer algorithms being based on the evolutionary of organism is surprising, the extensiveness with which this algorithms is applied in so many areas is no less than remarkable. Its usefulness and gracefulness of solving problems has made it the more favorite choice among the traditional methods. GAs is very helpful when the developer does not have precise domain expertise, because GAs possesses the ability to explore and learn from their domain. [2]

F. Hough Transform technique

In automated analysis of digital images, a frequently arising problem is detecting the simple shapes as straight line, circle or ellipse. In most of the cases an edge detector can be used as a pre-processing stage to obtain image points or image pixels that are on the desired curve in the image space. But due to imperfections in either the image data or the edge detector there may be missing or isolated or disjoint points or pixels on the desired curves as well as there may be spatial deviations between the ideal line or circle or ellipse and the noisy edge points a obtained from the edge detector [15].

G. Contour technique

The contour technique improves the segmentation speed. This technique is very fast since only border pixels information is used in segmentation. The algorithm starts with the searching of the first black pixel. Then normal edge algorithm is performed, where the window is considered as placed on the black pixels if 10 black pixels are in the window. In spite of its fast execution time, the segmented block may contain multiple columns or paragraphs. In order to reduce the multiple connected columns, the window use in contour is rectangular of 16 by 32 pixels instead of square 32 by 32 pixels. The main advantage of contour segmentation is the shape of block can be in any form and its execution speed [11].

H. Grouping technique

These methods involves in building alignments by accepting units in a bottom-up strategy. The units may be pixels or of higher level, such as connected components, blocks or other features such as salient points. Units are then joined together to form alignments. The joining scheme relies on both local and global criteria, which are used for checking local and global consistency respectively. Contrary to printed documents, a simple nearest-neighbor joining scheme would often fail to group complex handwritten units, as the nearest neighbor often

belongs to another line. Every method has to face the initiating alignments. Hence, these methods include one or several quality measures which ensure that the text line under construction is of good quality. When comparing the quality measures of two alignments in conflict, the alignment of lower quality can be discarded. Also, during the grouping process, it is possible to choose between the different units that can be accepting within the same neighborhood by evaluating the quality of each of the formed alignments [12].

Different character segmentation methods are discussed in this paper. There are different methods of character segmentation such as localized histogram multilevel thresholding, Bayes theorem, prior knowledge, feature extraction, dynamic programming, nonlinear clustering, multistage graph search algorithm, segment confidence-based binary segmentation, separator symbol's frame of reference and horizontal-vertical segmentation. All these methods are very useful as a preprocessing step for the OCR. Some of algorithms based on prior knowledge and separator symbol's frame of reference might not be useful for NP segmentation as it is difficult have prior knowledge regarding vehicle NP in advance. Dynamic programming and Segment confidence-based binary segmentation (SCBS) based methods can be really useful for NP character extraction.

III. CONCLUSIONS

Main objective of this survey is to study the different document segmentation techniques in detail and also the importance of document segmentation methods. Here we have discussed total eight different techniques of document segmentation. All these methods are very much useful as preprocessing step of Character segmentation.

REFERENCES

[1] Freixenet, J.et. al. Mu˜noz, X. Raba, D. Mart´I, J. and Cuf, X..Yet, "Another Survey on Image Segmentation: Region and Boundary Information Integration", University of Girona. Institute of Informatics and Applications. Campus de Montilivi s/n. 17071. Girona, Spain.

[2] Singh, Sukhmanpreet et. al. Verma , Deepa and Kumar, Arun. Rekha. Image Segmentation Using Soft Computing.

[3] Gonzalez , Rafael C. and Woods, Richard E. 2007. Digital Image Processing. Prentice Hall.

[4] Namboodiri, Anoop. M. and Jain, Anil K. Jan 2004. Online Handwritten Script Recognition. IEEE Trans. On Pattern Analysis and Machine Intelligence. vol. 26. no. 1. pp. 124-130.

[5] Antonapoulos, A. Page Segmentation using the Description of the Background Computer Vision and Image Understanding. Vol. 70, (1998) 350-369.

[6] Nagy, G. and Seth, S. 1984. Hierarchical representation of optically scanned documents. Proc. of ICPR. 347-349.

[7] Kruatrachue, B. and Suthaphan, P. 2001. A fast and efficient method for document segmentation for OCR. Electrical and Electronic Technology. Proceeding of IEEE Region 10 International conference on, Volume: 1, 19-22 Aug. (2001) 381- 383 vol.1.

[8] Jaekyu, Ha. Haralick, R.M. and Phillips, I.T. Recursive. 15Aug. 1995. XY Cut using Bounding Boxes of connected components. Proceedings of the Third International Conference on Document Analysis and Recognition, Volume:2, 14-952-954.

[9] Jaekyu, Ha. Haralick, R.M. and Phillips, I.T. 15Aug. 1995. Document Page Decomposition by the Bounding- Box Projection. Technique. Proceedings of the Third International Conference on Document Analysis and Recognition , Volume:2, 14-1119-1122. [10] Saitoh, T. and Pavlidis, T. 30 Aug.-3 Sept. 1992. Page Segmentation without Rectangle Assumption. Pattern Recognition Methodology and Systems, Proceedings, 11th IAPR International Conference on , 277 – 280.

[11] Kruatrachue, Boontee. Moongfangklang. Narongchai and Siriboon, Kritawan. Fast Document Segmentation Using Contour and X-Y Cut Technique. King Mongkut's Institute of Technology Ladkrabang Bangkok, Thailand 10520.

[12] Likforman-Sulem, Laurence. Zahour, Abderrazak Taconet, Bruno. 2006. Text Line Segmentation of Historical Documents: a Survey., submitted to Special Issue on Analysis of Historical Documents, International Journal on Document Analysis and Recognition, Springer.