RESEARCH ARTICLE OPEN ACCESS

Word Extraction Using X-Y Cut Algorithm

Simple Batra
Assistant Professor IT, Delhi School of Professional Studies and Research,affiliated to Guru Gobind Singh Indraprastha University, Dwarka, New Delhi

ABSTRACT
Digitization of printed documents is the motivating factor today to work more on text of scanned documents. Conversion of hand written scanned or printed documents into electronically readable form enables to store, exchange and process the valuable information. Text Recognition aims to recognize the text from printed or handwritten document to desired format. Several steps of text recognition include preprocessing, segmentation, feature extraction, classification, post processing. Preprocessing refers to the basic conversion operation of gray Scale image into Binary Image and removal of noisy signal from image. Segmentation does the segment the document image into line by line and extracts each word from segmented line. Feature extraction is calculating the characteristics of character. A classification contains the database and further processing of them. The paper proposes the approach to extract words from based on a set of properties for each connected component in the whole binary image of the document which is independent of languages.
Keywords-OCR, Word Extraction, binary image, digitization

Date Of Submission:21-12-2018 Date Of Acceptance:05-01-2019

I. INTRODUCTION

Technology is driving the society now through the digitization and use of online systems all around us. Information which is available either in hardcopy i.e. printed, handwritten or scanned. The current technology and various applications in which text extraction is useful include digital libraries, multimedia systems, Information retrieval systems, and Geographical Information systems. Optical Character Recognition (OCR) is software that converts scanned or printed image of the document into a text document through Pre-processing, Segmentation and Recognition. Text Recognition usually involves a program designed to translate images of typewritten text (usually captured by a scanner) into machine editable text. OCR began as a field of text recognition used in real life applications where we want to collect some information from text written image. People wish to extract text from the scanned document images available in different formats. This paper proposes the method to extract words from specific lines on the basis of bounding boxes.

II. METHOD-

Optical character recognition, or OCR, is a process which enables us to convert text based images can be produced by scanners, cameras, read only files, etc. into editable electronic documents. Optical Character Recognition software Optical character recognition software follows several steps to convert an image file into an editable document. Specific algorithms are used at each step to change, enhance, and interpret the images found within a file. Every step involved in this process and its precision of result is critical to the overall success of OCR.Optical character recognition firstly translates the text image into editable character codes such as ASCII. Any OCR implementation consists of a number of preprocessing steps as shown in Figure-

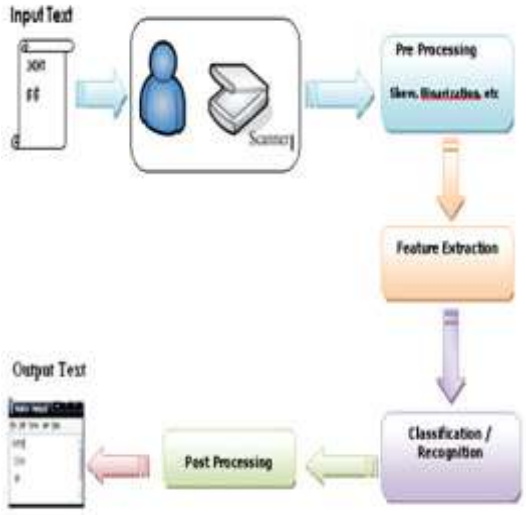

Fig. 1-Process of OCR

Firstly a printed document is scanned and converted into a gray scale image.

Binarization is the technique used to do this conversion. To identify the objects of interest from the rest in a document image it involves separating the pixels forming the printed text or diagrams

(foreground) from the pixels representing the blank paper (background).

The binarization is the process of separating the foreground and background information. This separation of text from background is a prerequisite to further process the input image. Scanned documents often contain noise which is essential to be removed that arises due to printer, scanner, print quality, age of the document, etc. Further skew is unavoidable due to the degrees of tilt of an original document fed to the scanner. Identification of document image skew angle and rotation of image according to the detected skew angle is done next.

Binarized and skew corrected, document image now becomes an input image for actual text extraction module. The top down approach is used by implementation of the X-Y tree method which segments a document in multiple steps into a hierarchical tree structure consisting nested rectangular blocks. The method presents an intuitive approach to segment the document into big blocks through horizontal and vertical cuts. And then the same process is repeated for in each one of the sub-blocks. A logical order of the document can be created by giving recursive cuts in the document to find paragraphs in a document and further lines in it [1]. The X-Y cut method is used for finding horizontal projection profile of the document image to extract the lines from the document. If the lines are well extracted, the horizontal projection will have separated peaks and valleys, they serve as the separators of the text lines. These valleys once detected can be used to determine the location of boundaries between lines [1]. Similarly, gaps in the vertical projection of a line imageare used in extracting words in a line as a separate module which will enable to save words in separate files for being processed further

III. ALGORITHM OF WORD SEGMENATION

Module is implemented by taking a line as an input and a function named as "word_extract" is called, which works recursively to extract the words from that line and stores it in a separate directory.

This process has some steps that are very similar to the problem of line segmentation [1]. The process starts by computing the projection profile of the image, iterating across vertical lines. We assume that two words are separated by a significant amount of pixels. With the projection profile we can find the lengths of white spaces that are in the line. Each component is decomposed into smaller regions by analyzing its vertical and horizontal projection profiles, and finally each of the small regions satisfying certain heuristic constraints is labeled as text.

Steps involved in the word segmentation module-

1. Reads an Input Document Image having an extracted line from the paragraph or a block.
2. It then performs the vertical as well as horizontal projection of an image.
3. It finds out the continuous white pixels in the row.
4. It then marks the ending of word by pointing the black pixel.
5. Then does the extraction of that word.
6. Extracted word is stored in the separate directory.
7. These extracted words can be taken as an input from the file for character segmentation work.

IV. RESULTS-

To implement word segmentation any scanned document is given as an input and word_extract module is implemented on it using MATLAB. Firstly the paragraphs from the document are extracted and stored in separate directories, then these paragraphs are given as input and lines are segmented and stored in separate directories [1]. Finally these lines are given as input and a function named as "word_extract" is called, which works recursively to extract the words from that line and stores it in a separate directory.

[9

अमित को साथी क्लर्क से ऐसी बात सुनकर आश्चर्य हुआ। आफिस की गपशप से अलग इस बात ने उसके मानस को झकझोरा। वह स्वयं उलझन में था और राह की खोज में था।

''भाई गोयल, आपने मेरे अन्तस को छुआ है। मैं इस शहर में नया हूँ और यह सम्पर्क मेरे लिए हितकारी होगा। आप मुझे कुछ ऐसे व्यक्तियों से मिलवा दें जो चित्रकला में कार्यरत हों। इससे मुझे प्रेरणा मिलती रहेगी।'' अमित ने आभार के स्वर में कहा।

गोयल ने उसे साथ लिया और शहर के जाने-माने कलाकारों से मिलवाया। सबसे पहले वह अँबूजी से मिलाने ले गया। अँबूजी अच्छे खासे जमे हुए चित्रकार माने जाते थे। आदमी सज्जन थे और सहायता करने में पीछे नहीं रहते थे।

''आइये गोयल साहब, आइये। आज कैसे कष्ट उठाया आपने। और आज आपके साथ कौन सज्जन पधारें हैं। हमें भी इनसे मिलवाइये।'' पान की पीक थूककर अँबूजी ने आवभगत की।

पुरानी बस्ती के इस दबे हुए कमरे में चित्रकारी के उपकरण बिखरे पड़े थे। कुछ ताजा पर अधूरे चित्र एक कोने में झाँक रहे थे। रंगों के कर्मी आमने-सामने बैठे और गोयल ने परिचय कराया ''अँबूजी, ये हैं मि० अमित। हमारे कार्यालय में ही काम करते हैं। इस नगर के लिए नये हैं। इनकी रुचि भी चित्रकला में है और अच्छा हाथ चलता है। सोचा आपसे मिलवा दूँ। मिल बैठेगें दीवाने दो.....।'' अँबूजी ने स्वभाववश ठहाका लगाया और अमित भी हँसे बिना नहीं रह सका।

''तो अमित भाई भी कलाकार हैं। भाई गोयल हमारा तजुर्बा कुछ इस तरह का है कि जिसे दुनिया में कहीं प्यार नहीं मिलता वह कला का प्रेमी बन जाता है। कलाकार को पैदाइश ही कहीं में —उपेक्षाओं में है। देख रहे हो मुझे। आज जीवन के तीस साल इस हुनर में खपाकर भी फटेहाल हूँ। इस मुल्क को कद्र करना ही नहीं आता। क्यों अमित भाई, ठीक कह रहा हूँ न।'' अँबूजी ने जीवन-सार फेंका।

अमित को लगा कि कहीं न कहीं अँबूजी ने उसकी दुखती रग को टटोल लिया है और उसे होले से दबा दिया है। वह बोला—'अँबू साहब, आप तो बुजुर्ग हैं। मैं तो अभी बालक हूँ इस क्षेत्र में। परन्तु, यदि व्यक्ति परिश्रम करता रहे तो उसे सफलता अवश्य मिलती है। आप अच्छे चित्रकार हैं और अनेक ऐसे अवसर भी आये होंगे

Fig. 2-This is the input image for the extraction of the blocks from it

Fig. 3- Directory showing all the paragraphs obtained from Input Image

Fig. 4- Input paragraph to extract the lines from it

Fig. 5- Directory showing all the lines extracted from the paragraph

Fig. 6- showing an input line given to module "word_extract"

Fig. 7- Directory showing all the words extracted from the line using "word_extract"

V. CONCLUSION-

Very promising results are achieved by using this technique to extract words.Segmentation of the various parts of the document Images and saving them in isolated files for the further use has been implemented using recursive approach to perform the horizontal and vertical projections and cuts are found within the whitespaces based on the varied intensities of the pixels.. It can give out Paragraphs, lines and then words extracted at three levels. Thresholds are necessary to find out the inter-line or inter-block distances. We need to find out the coordinates of the segmented regions in order to plot the bounding box around that segmented regions. As per the varied document styles and depending upon the style of documents it becomes complex to define the optimal cuts in the document. The method can be applied to the printed documents and are independent of language.

In future the work can be done for the enhancement of the segmentation of paragraphs, lines and words from the multicolumn documents and which include graphics positions at different places in order to enhance the efficiency of the modules for any OCR. Even research work can be done to look out for some another mechanism, which can work on different font styles and thus makes it independent of font styles even. Defining the normalized bounding boxes for the segmented regions can be proposed in the future. In word segmentation the complexity issue related to finding the dynamic mechanism to define bounding box of each word in the same image instead of storing them separately in separate files.

REFERENCES-

[1]. Batra Simple, "XY Cut Modular approach for Segmenting pages", International Journal of Scientific Research in Computer Science and Engineering", Vol.6, Issue.2, April 2018 edition.

[2]. Faisal Shafait, Daniel Keysers, Thomas M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images", IEEE, Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06) 2006

[3]. Antonacopoulos1, B. Gatos2 and D. Karatzas, "Page Segmentation Competition" Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003) 2003 IEEE.

[4]. Faisal Shafait, Daniel Keysers, and Thomas M. Breuel, "Performance Evaluation and Benchmarking of Six Page Segmentation Algorithms" DRAFT, November 30, 2007

[5]. Song Mao and Tapas Kanungo, "Empirical Performance Evaluation Methodology and Its Application to Page Segmentation Algorithms", IEEE transactions on pattern analysis and machine intelligence, vol. 23, no. 3, March 2001

[6]. Zhixin Shi and VenuGovindaraju, "Multi-scale Techniques for Document Page Segmentation", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 2005 IEEE

[7]. Jean-Luc MeunierXerox Research Centre Europe, « Optimized XY-Cut for Determining a Page Reading Order", Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05) 2005 IEEE

Simple Batra" Word Extraction Using X-Y Cut Algorithm" International Journal of Engineering Research and Applications (IJERA) , vol. 8, no.12, 2018, pp 60-63