ISSN 2395-1621

Survey On: Text Extraction from scene Images

[#1]Ankit B Prasad, [#2]Harshal A Gunjal, [#3]Yogesh A Rajguru, [#4]Rushikesh K Ugale,
[#5]Prof. Borkar B.S.

[1]ankitprasad3948@gmail.com,
[2]harshalgunjal34@gmail.com,
[3]iamrajguru007@gmail.com,
[4]rushikeshugale23@gmail.com,
[5]bharat.borkar@avcoe.org

[#12345]Department of Information Technology, AVCOE Sangamner,
Savitribai Phule Pune University, Maharashtra.

ABSTRACT

The successful analysis of image is currently in great demand because scenes in images is a major source of data in our lives. The text is a direct source of information, while recent surveys on the detection and recognition mainly focuses on extracting text scene pictures. Here, this paper presents a survey of text detection from images. This paper surveys different proposed generic framework that successively implements text detection, recognition and extraction in images, the text can be of any type, both caption text and scene text. The video will be retrieved based on scene and text content. We surveyed different offered approach for content-based image segmenting and retrieval in large image collection of different papers. Subsequently, we extract textual metadata by applying Optical Character Recognition (OCR) technology on key-frames.

Keywords: AI, OCR, Computational Vision, OpenCV.

ARTICLE INFO

Article History

Received: 20th March 2019

Received in revised form :

20th March 2019

Accepted: 22nd March 2019

Published online :

23rd March 2019

I. INTRODUCTION

Now-a-days, demand for the software systems is growing to recognize characters in computer system when scanned image through paper documents. Storing the information available in paper format into a storage disk and then later reusing that information in efficient way. One simple way to store these paper documents into computer is to scan the documents and save it in image or pdf format. The text in images cannot be edited by the user and sometimes it is difficult to interpret. Searching text from image is very difficult for computer system, user need to read the individual contents from these documents line-by-line and word-by-word. Thus to develop some text recognition algorithm is an essential need to perform Document Image Analysis (DIA) which transforms documents in paper format to electronic format that can be easily edited. In the processing of scanned image one of the initial step is preprocessing. The processed image is checked for noise, skew, slant etc. There are chances of image getting skewed with either left or right orientation or with noise such as Gaussian. Here the image is converted into grayscale and then into binary. Hence we get image suitable for further processing. Preprocessing is the operation on input image which includes functions like binarization which convert grayscale image into Binary Image, noise reduction removes the noisy signal from image. Segmentation stage is for segment the given image into line by line and segment each character from segmented line. Future extraction calculates the characteristics of character and ROI of image.

II. LITERATURE SURVEY

The Text recognition and extraction from images is an active research in the field of pattern recognition and image processing. The issues related to text recognition and extraction: Many researchers have proposed different technologies and we have proposed a survey for their technology in this paper, each approach or technology tries to address the issues in different way. A detailed survey of approaches proposed to handle the issues related to text recognition and extraction are as follows.

Badawy, W. et al. [1] has discussed the Automatic license plate recognition (ALPR) is the extraction of text

from scene image or a sequence of images containing vehicle license plate information. The extracted text from image or sequence of images can be used with a database in many applications, such as e-payment systems (toll payment, parking fee payment), freeway and arterial monitoring systems for traffic surveillance. The ALPR uses a color, black and white or infrared camera to take images.

Gur et al. [2] has discussed some problems in text recognition and retrieval. Automated optical character recognition (OCR) tools doesn't have a complete solution and in most cases human interaction is required. They recommend a novel text recognition algorithm based on fuzzy logic rules relying on statistical data of the analyzed font. This approach combines letter statistics with correlation coefficients in a set of fuzzy based rules, enabling the recognition and extraction of distorted letters that may or may not be retrieved focused on rashi fonts associated with commentaries of the bible that are actually handwritten calligraphy.

Yang et al. [3] has proposed a novel adaptive binarization method based on wavelet filter is proposed. This approach was processes faster, so that it is more suitable for real-time processing and applicable for mobile devices. They evaluated this adaptive method on complex scene images of ICDAR 2005 database.

Sankaran et al. [4] has proposed a novel recognition approach that result in a 15% decrease in word error rate on heavily degraded Indian language document images.

Jawahar et al. [5] has proposed a recognition scheme for the Indian script of Devanagari. They used approach does not require word to character segmentation, which is one of the most common reason for high word error rate. They have been reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

Malakar et al. [6] has described that recognition and extraction of text from images is one of the important steps in the process of an Optical Character Recognition (OCR) system. In case of handwritten images, presence of skewed, touching or overlapping text line(s) makes this process a challenge in front of the researcher.

Rhead et al. [7] has considered real w world UK number plates and relates these to ANPR. It considers aspects of the relevant legislation and standards when deploying them to real world number plates. The varied manufacturing techniques with varied specifications of component parts are also noted. The varied fixing methodologies and fixing locations are discussed as well as the impact on image capture.

III. ALGORITHM

1. Start
2. Take input image.
3. Do preprocessing like segmentation, noise removal, skew correction, etc.
4. Do Text recognition by feature extraction and classification.
5. Train images from system training module consisting supervised learning.
6. Load the DATABASE containing training module.
7. Compare training and testing module
8. Do Post processing module.
9. Store text in proper format.
10. End.

IV. APPLICATIONS

Text recognition and extraction may be applied throughout the entire spectrum of industries, revolutionizing or in vanishing the document management process. This technology enable scan documents to become more than just image files, turning into fully searchable and editable documents with text content that is recognized by computers. By using this people no longer need to manually retype important documents when entering them into electronic databases. Instead, Text recognition system extracts text and enters it automatically. The result is accurate, efficient information processing in less time which ultimately reduces human stress and time. We will overview some applications of text recognition system, some are as follows.

1. Legal[8]: In the legal industry, significant movement to digitize paper documents has been initiated. In order to save space and eliminate the boxes of paper files which are currently available, documents are being scanned and entered into computer databases. Image text recognition and extraction simplifies the process by making documents text-searchable and editable, so that they are easier to locate, find and work with once entered in the database.

2. Healthcare[8]: Healthcare uses image text recognition technology to process paperwork prescribed by doctors. Healthcare system always have to deal with large volume of forms for each patient, including insurance forms as well as general health forms done by patients. By using image recognition and extraction they are able to extract information from forms and put it into databases which are searchable and editable later on when demanded, so that every patient's data is promptly recorded.

V. FIGURES AND TABLES

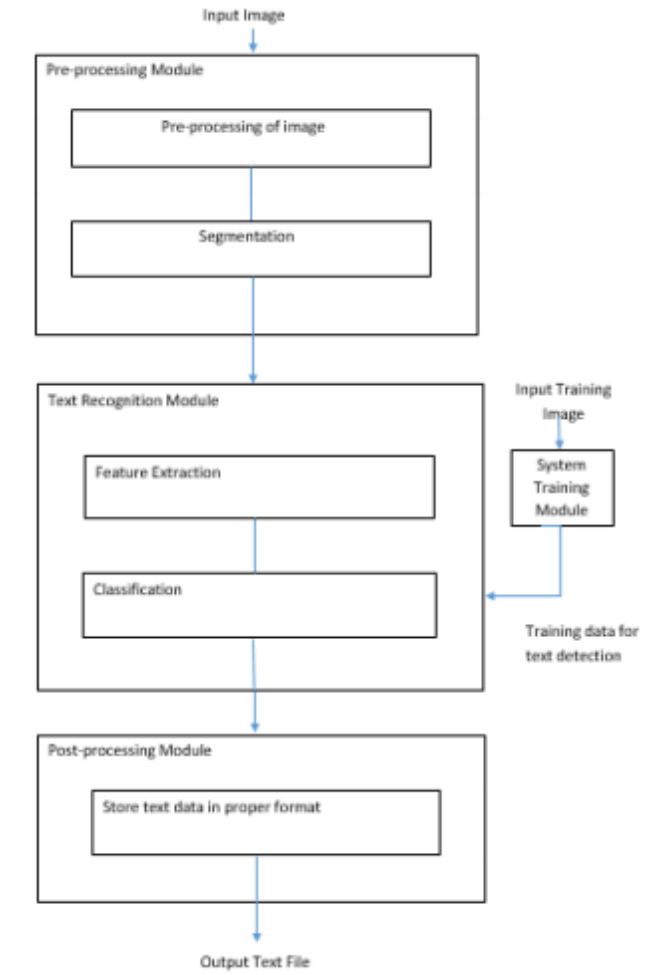

Fig. 1. Block Diagram Of Text Recognition

VI. CONCLUSION

In this paper we proposed different issues faced by researcher for character recognition and algorithm to solve them. We served a medium to solve different issues in text extraction by this survey paper. We also surveyed different approaches of text recognition and extraction such as different approaches of OCR, etc.

REFERENCES

[1] Badawy, W. "Automatic License Plate Recognition (ALPR): A State of the Art Review." IEEE International Conference on Document Analysis and Recognition, 2012

[2] Gur, Eran, and ZeevZe lavsky, "Retrieval of Rashi Semi Cursive Handwriting via Fuzzy Logic," IEEE International Conference on Frontiers in Handwriting Recognition (ICFHR), 2012.

[3] Yang, Jufeng, Kai Wang , Jiaofeng Li, Jiao Jiao, and Jing Xu, "A fast adaptive binarization method for complex scene images," 19th IEEE International Conference on Image Processing (ICIP), 2012.

[4] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, "Robust Recognition of Degraded Documents Using Character N-Grams," IEEE, 2012.

[5] Naveen Sankaran and C.V Jawahar, "Recognition of Printed Devanagari Text Using BLSTM Neural Network," IEEE, 2012.

[6] Malakar, Samir, et al. "Text line extraction from handwritten document pages using spiral run length smearing algorithm," IEEE International Conference on Communications, Devices and Intelligent Systems (CODIS), 2012.

[7] Rhead, Mke, "Accuracy of automatic number plate recognition (ANPR) and real world UK number plate problems." IEEE International Carnahan Conference on Security Technology (ICCST), 2012.

[8] Application of OCR, from http://www.cvisiontech.com.

