IARJSET

ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Vol. 8, Issue 7, July 2021

DOI: 10.17148/IARJSET.2021.8773

Text Extraction and Recognition from Images

Mrs. Ranjitha K.N[1], Sai Kumar LS[2], Sandesh Naikal[3], Varun R Reddy[4], Vinay Biradar[5]

Assistant Professor, Dept. of Computer Science, K.S. Institute of Technology, Bengaluru, India[1]

Student, Computer Science, K.S. Institute of Technology, Bengaluru, India[2,3,4,5]

Abstract: Text recognition in images is an active field of research that aims to create a computer program with the ability to interpret text from images automatically. There is a huge demand nowadays to store the knowledge available on paper records in a machine-readable form for subsequent use. One convenient way to store data in the computer system from these paper documents is to first scan the documents and then store them as images. It is, however very difficult to read the individual material and scan the contents of these documents line by- line and word-by-word to reuse this information.

Keywords: Text extraction, summarization, Voice over, OCR

I. Introduction

Text identification in natural environments, such as advertisements, street signs, bills, and so on, is an important component of computer vision applications such as robotics and vehicle number plate recognition. There are two steps to text detection. The first phase is to detect text regions in a picture, and the second step is to obtain text information from these regions. The goal of text detection is to figure out where the text is in the input image, and a bounding box is used to indicate the position. Text recognition in photos is a growing field of study that tries to develop a computer programmer that can automatically interpret text from images. Nowadays, there is a tremendous demand for storing knowledge from paper records in a machine-readable format for future use. Scanning the documents and then storing them as images is a convenient approach to save data in the computer system from these paper documents. The process of creating a concise, fluent, and most crucially accurate summary of a lengthier text content is known as text summarization. The main idea behind automated text summarizing is to be able to discover and provide a small subset of the most important information from the complete package in a human-readable style.

II. Methodology

Text detection and localization: Text detection is concerned with identifying the existence of text in the input image, whereas text localization is concerned with positioning the position of the text and forming groups of text regions by removing as much background as possible. Using connected component analysis or region-based approaches, the text detection and localization procedure is carried out. Segmentation: Some techniques for scene text detection utilising segmentation maps are proposed, inspired by segmentation approaches. Pixel link predicts the connections between each pixel and its neighbours that are valid when both linked pixels are text instances. Pixel link succeeds in separating text occurrences that are very close to each other by doing so. Text Snake was recently proposed as a method for detecting text instances by anticipating the text region and centre line, as well as geometry properties.

End-to-end text detectors: The detection and recognition modules are simultaneously trained in an end-toend strategy to improve detection accuracy by utilising the recognition result. Most approaches detect text using words as their unit, however specifying the extents to which a word can be detected is difficult because words can be separated by a variety of factors, including meaning, space, and colour. Summarization: The process of creating a concise, fluent, and most crucially accurate summary of a lengthier text content is known as text summarization. The fundamental goal of automatic text summarization is to be able to extract a small subset of the most important information from a large number of documents and display it in a human-readable fashion. Automatic text summarization systems have the potential to be tremendously valuable as online textual data rises, because more useful information can be read in a shorter amount of time. Text to Speech Module: Using speech synthesis techniques to convert text to vocal output, Text-to-speech is a voice synthesis process that converts text in a computer document into a spoken sound version. For the visually impaired, Text to Speech can make it possible to read information on a computer display.

III. Implementation

A. Text Recognition

The foundation of this project is optical character recognition (OCR), a technology that converts any image, PDF, or document into editable text. The OCR component of this project is used on photos that can be in a variety of formats.

DOI: 10.17148/IARJSET.2021.8773

Tesseract and OpenCV are used to implement the optical character recognition technique in Python. Tesseract is a text recognition engine, and PyTesseract is a Tesseract Python binding that is used in this project to process images. Tesseract can offer suitable output in general picture extraction, but in the worst-case scenario, if an image is noisy and in deform order, the project will require a Tesseract tuning process to improve the quality of the output Tesseract generated. In this case, OpenCV can play a crucial role in fine-tuning the Tesseract technique's output. To solve the problem of noisy images and improper image format, this project uses PyTesseract with OpenCV, which may help tune PyTesseract engine and produce a nice output. OpenCV is used to recognise text in images, reduce noise in images, and develop image output to aid Tesseract in extracting text from images. As a result, the initial component of the pipeline, word recognition and extraction from images, is implemented using a PyTesseract and OpenCV combo.

B. Summarization

After cleaning and separating the text into sentences, the following stage in the Textrank algorithm was vector production of sentences using word embedding, followed by the creation of a similarity matrix to detect similarities between sentences. Following the creation of a similarity matrix, use the Textrank method to provide a summary based on the sentence score. In addition, the highest-scoring sentences were used to create a project overview. The image of the similarity vector matrix was provided, which aided in the use of the text summarization technique. The Textrank algorithm is used using a similarity metric to detect how similar two things are regardless of their size and produce a score in a matrix. The development of text summarization techniques, such as the text rank algorithm and the TF-IDF algorithms, was the final stage of the project after text extraction and text pre-processing. Unsupervised and extractive text summarization algorithms are the two algorithms. The evaluation and results of the applied algorithms to generate a summary of the given image are provided in this part. To construct a summary, use the textrank algorithm to the retrieved text. This is an unsupervised text summarising method that generates results using word and sentence scores. This is an extractive text summarization method that can create an extractive summary that is very relevant. The summary of an image is shown here using the textrank method; the summary is based on the best five sentences of the complete text, as scored by sentence vector and similarity matrix. As a result of the textrank algorithm, numerous summaries of various sentences are provided.

C. Voice Over

The Text-to-Speech module is designed to be a user-friendly application for everyone. Text-to-Speech convertor and Image-to-Text convertor are the two main modules used in this programme. The software provides a multi-functional platform for users to conveniently communicate, listen, or narrate. Users have the option of converting legible photos to text files or reading text directly. The text-to-speech mode converts a text file or entered text to speech, which is then narrated/read using Microsoft SAPI's voice database. The application also includes a narrator to assist users in using the software. This is done by concatenating syllables with phonemes using the optimal coupling technique. The image-to-text mode turns viewable images into a text file that can then be converted to voice. Readable images are those with less intricacy in the foreground, allowing the letters to be extracted in a logical manner. In this software, the user has a number of alternatives from which to choose the mode of operation. When the Text-to-Speech option is chosen, the user must enter text into the text input box or into a text file. The text is parsed, and speech is generated as a result.

IV. CONCLUSION

This project focuses on the specific characteristics of content that help it stand out from other image material, such as recurrence and introduction data. We'll start by cleaning up the image by adjusting the contrast and gradient. The images' objects have now been identified and numbered. These numbered objects are further separated into text and non-text during the text recognition process. Later the recognized text is reconstructed to form a meaningful text present in the image. Also, we are focusing on extracting the text such that certain portion of the images such as logos etc. is retained. This is accomplished by calculating the pixels of the required portion of the image to be retained and then training the system to extract all the text except the portion of the image to be retained, resulting in average accuracy in the work. Although the OCR System's results aren't great, they aren't terrible either, indicating that the OCR Technique isn't overwhelmed. Text detection can be used in real-world scenarios such as optical character recognition, artificial intelligence, separating human and machine inputs, and spam removal. The process of detecting areas in an image where a full text appears is known as text detection. It's a difficult process because the environment in which the image is captured varies. With the ever-growing text data, text summarization seems to have the potential for reducing the reading time by showing summaries of the text documents that capture the key points in the original documents.

V. FUTURE ENHANCEMENT

Implement this project as a mobile application. Mobile applications are growing more, popular, and users prefer them since they are portable and offer a simple user interface. Due to time constraints, we only focused on text images and

DOI: 10.17148/IARJSET.2021.8773

documents with no skew, that contain English Alphabet. As an example, we were unable to extend my work to make my application process text documents in other languages, such as Japanese, Chinese, Arabic, or any language that does not use the English alphabet, will be attempted to be integrated. Add a new feature called paraphrase, which can be used for paraphrase the extracted text. Paraphrase means changing the sentence without changing the meaning of sentence.

REFERENCES

[1]. H. Lin, P. Yang, and F. Zhang, "Review of scene text detection and recognition," Archives of Computational Methods in Engineering, vol. 27, no. 2, p. 433–454, 2020.

[2]. Ankit Kumar, Zixin Luo, Ming Xu, "Text Summarization using Natural Language Processing," in Juniper Networks, 2018.

[3]. Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang, "Convolutional attention networks for scene text recognition," 2019.

[4]. Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, Hwalsuk Lee, "on Computer Vision and Pattern Recognition (CVPR)," in Proceedings of the IEEE/CVF Conference, 2019.

[5]. C.P. Chaithanya, N. Manohar, Ajay Bazil Issac, "Automatic Text Detection and Classification," International Journal of Recent Technology and Engineering (IJRTE), vol. 7, no. 5S3, 2019.

[6]. Dai Y, Huang Z, Gao Y, Xu Y, Chen K Fused text seg-mentation networks for multi oriented scene text detection, 2017.

[7]. Deng D, Liu H, Li X, Cai D PixelLink: detecting scene text via instance segmentation. In: Proceedings of association for the advancement of artificial intelligence, 2018.