TEXT EXTRACTION FROM IMAGES USING OCR

Syed Mohamed Asif(9919004272), S V Nagendra(9919004270), H Nikhil Reddy(9919004106),

S LikhilSrinivas(9919004269), M Jaipal(9919004177)

Abstract:

Optical Character Recognition (OCR) has been a subject matter of hobby for decades. It is defined as the technique of digitizing a document photo into its constituent characters. Despite a long time of intense research, developing OCR with talents corresponding to that of human still remains an open challenge. Due to this hard nature, researchers from industry and academic circles have directed their attentions closer to Optical Character Recognition. Over the previous few years, the range of instructional laboratories and agencies worried in research on Character Recognition has elevated dramatically. This research objectives at summarizing the studies to date completed within the discipline of OCR. It affords an assessment of various factors of OCR and discusses corresponding proposals geared toward resolving problems of OCR. The goal of this studies is to produce a significant summary the use of unsupervised extractive textual content summarization algorithms at the picture extracted textual content using EasyOCR. This studies of extractive textual content summarization from pics can contribute in text analytics and also can cross a step beforehand in making textual content summarization.

I. Introduction:

Data technology is a records-pushed choice making process. In the early level of digital evolution the statistics changed into mainly generated from PCs, however inside the later level data is generating from lots of virtual gadgets. For this large amount of records, people are flooding with the information and facts, due to drastic growth in big-information and net. To cope with this big dependent and unstructured information there are numerous technique in facts technology, in that text analytics is focused on natural language processing .

Text summarization is a method to discover significant precis from a prolonged portions of text. Today's international people are surrounded with the aid of massive quantities of information in the digital area, automatic text summarization techniques can help to get a quick and meaningful precis that may help human to recognize the textual content in much less time, also increase in the first-rate and quantity of information within the brief piece of summarized textual content . There are many strategies for text summarization in natural language processing (NLP) area. On the alternative hand, textual content extraction is likewise part of text analytics, which can be accomplished from the photo. Automatic textual content reputation and extraction from an photograph is part of NLP[Natural Language Processing].

Optical Character Recognition(OCR) is center approach in the back of the text extraction from an image. OCR era can accumulate the text data from any format of an photograph and can be used for the NLP strategies.Here are some of the types used for Extraction of text from an Image.

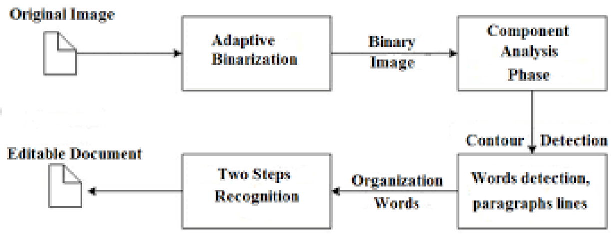
1.Using OpenCV and Tesseract-OCR

2.EasyOCR

3.State-of-the-Art Deep Learning Techniques

Above all methods we are using EasyOCR and PIL Library to obtain Text from Image .

The principal aim of this proposed research is text summarization of the extracted records from photo that is a aggregate technique of system getting to know and natural language processing techniques.Due to very speedy growth of to be had multimedia documents and growing requirement, research within the discipline of sample popularity shows a wonderful quantity of interest in green extraction of text, indexing and retrieval from digital video/report snap shots. Intensive studies initiatives are completed for textual content extraction in pics via many pupils. Text extraction includes detection, localization, tracking, binarization, extraction, enhancement and popularity of the textual content from the given photograph.Several techniques were evolved for extracting the text from an picture. The proposed methods have been based totally on morphological operators, wavelet transform, synthetic neural network,skeletonization operation,edge detection set of rules, histogram technique and so on. The strategies referred to in this paper on text extraction in snap shots are labeled according to one-of-a-kind forms of images.



The goal of this studies is to produce a significant summary the use of unsupervised extractive textual content summarization algorithms at the picture extracted textual content using EasyOCR. This studies of extractive textual content summarization from pics can contribute in text analytics and also can cross a step beforehand in making textual content brief application in an precise manner to make human life greater dependable and time saving in this massive digital data world.

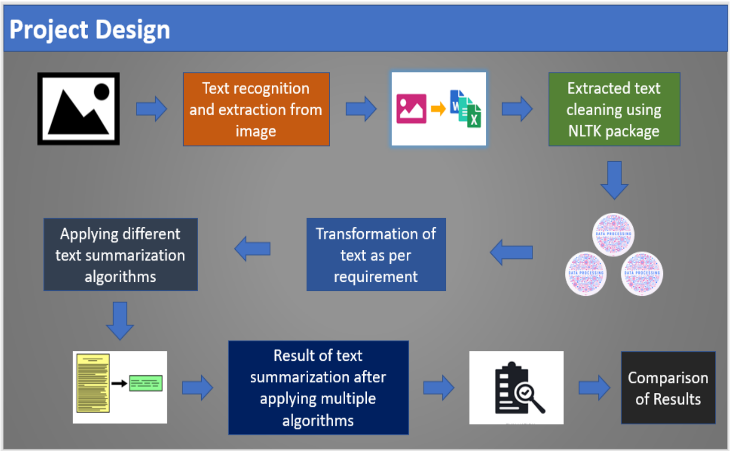
Existed methods of image collection summarization are categorised as follows: visual summarization methods and multi-modal summarization ones. Visual summarization techniques observe most effective visual features of pictures. Generally, these methods utilize numerical metrics for doing the summarization and forget about semantic family members among data. Here, we are applying EasyOCR out of these techniques which will be very useful to obtain the required output.

In the proposed approach, a Easy Optical Character Recognition[OCR] image classification technique was used to get a better output from the image or any input file. As the image classification approach moved toward more general and more specific classifiers, the result of image summarization got easy. There remainded a important point where the summarization performed best. Finally, a group of high-level features was stated based on the classifier output for the classification task.

II. Methodology:

The proposed research is on text extraction which is mainly focused on text summarization of extracted text from an image. Here the process of the research is mainly divided into three parts, text extraction form an photo, textual content pre-processing after which applying different extractive summarization set of rules. Research need to be follow a particular technique to achieve the goal. This research is particularly focused on image processing and text analytics sequentially, after a long research concluding that the easy ocr model is a great manner to produce a end result.

There are different technologies and Python programming language are used to get realistic text extraction. Along with this different libraries and machine learning algorithms are applied in different stage of this research. Python programming language is used in this whole research. In the application layer, OpenCV libraries are used to extract text from an image packages are used to pre-processing and after cleaning the text data different extractive summarization algorithms are applied to get realistic text summarization. After giving input in input queue data processed and produced result in output queue.


Project Design
Text recognition and extraction from image
Extracted text cleaning using NLTK package
Transformation of text as per requirement
Applying different text summarization algorithms
Result of text summarization after applying multiple algorithms
Comparison of Results

The first part of this project is mainly based on optical character recognition (OCR) which is a technology which can transform any image and document into an editable text file. The OCR part of this project is applied on the images which can be in different formats. The optical character recognition technique is implemented by python programming language, and OpenCV. In this situation OpenCV can act as an important part to tune the output technique. OpenCV is used for the textual content detection, noise discount from an picture and can develop the output from image to help extracting the text from an image. So, implementation of first part of the pipeline which is text recognition and extraction from the image is implemented with OpenCV.

Also, to tune the result of OCR, lots of operation have been done like noise removal on the selected image before giving to the optical character recognition for further processing

In this project, as per requirement the total implementation done in python and image to text which is done by optical character recognition. These are the implementation part of this project.

Last and final part of the project methodology is maintaining any project or any application. To keep any project vital part is maintain up to date. In this project this part is not fully functional. The aim of this assignment is to make an interactive application through the usage of OCR and text analytics.

III. Related Works:

This part of the report is fundamentally centered around to examine about the past works, alongside qualities and restrictions of the carried out methods which have as of now done in the past upon the related topic. Also, this section is making an attempt to provide the strengths and limitations of the previous techniques for the chosen topic.

A. Optical-character-recognition

The text extraction approach has been done in various manner by the researchers. The optical character recognition process is the most regular and well known cycle to do the text extraction from a picture. In the year 2007, researchers gave a brief about the EasyOCR engine. The open-source EasyOCR engine for OCR was developed by the HP to use in different digital devices. The architecture of OCR engine divided into two parts, first is text recognition and second is checking the pattern of the text in papers. There are several techniques protected within the EasyOCR engine that are noted by way of the researchers. Line and word finding, Word recognition, Static character classifier, linguistic analysis, adaptive classifier all are techniques of OCR engine to get a better and more accurate result. In referred to method of OCR researchers noted approximately the EasyOCR to growth the accuracy of the text extraction engine (Smith; 2007).

In the recent dates the utilization of OCR is developing quickly to work on the digital technology. Modern world is improving and the use of image in different sectors are also increasing. Image can carry important data, to extract the data from the image is a challenging part in digital world. Textual data are available from different resources like newspapers, images, notes etc. In 2019, Pawar and their group referenced more in short with regards to the text based substance extraction procedures in an research paper from the picture the use of the EasyOCR engine. OCR engine is a section of artificial intelligence which is an advanced text extraction method from image and different sources. At the initial stage, the OCR was expand upon the convolution neural organization, after the improvement of the OCR, presently its for the most part dependent on Long Short Term Memory (LSTM) which is essential for

Recurrent neural organization. EasyOCR is an open source and best for the handwritten text recognition and extraction. The examined research is essentially focus around the EasyOCR model and study revealed with various methodology of OCR model. Those are Connected component based method, Sliding window based method, Hybrid method, Edge based method, Corner based method, Texture based method, Corner based method, Stroke based method, Semi automatic ground truth generation based method. Also, researchers discussed about the advantages of the OCR model, as a data image can take more storage than a text and along with this people still prefer the text report as information in excess of a picture, which made the OCR cycle more important in the new days. Also, OCR can be used to make text data into speech after extracting from the image which can be precious for the blind humans (Pawar et al.; 2019).

The extraction of image text in an perfect way is the difficult task. Also, different images can be in different format, size and colour which can be a difficult task for the OCR.In a review researches presented a methodology which is centered around to remove the picture more clear. Here, researchers introduced the method of image processing than 3cropping and then text extraction. As a end result this method of EasyOCR is implemented on images and produced accurate textual content (Chawla et al.; 2020).Researcher R.R. Palekar and his group cited approximately one of the hard element in text detection after which extraction from the image in 2017. To broaden the result of the EasyOCR researchers used OpenCV and OCR together.OpenCV used to do image processing and EasyOCR used to extract the textual content from the image which is comparatively more correct than the easy OCR engine. After using the following process researchers identified that the text processing before text extraction is an important part. Also, revealed that text processing before text extraction with the different image achieved more accurate result. Researchers used real car number plate to identify the accuracy of the model. Here, OpenCV with the EasyOCR created amazing outcome (Palekar et al.; 2017).

In 2019 a researcher Neha Joshi developed an particular approach to produce a textual content summarization from an photo. The research used EasyOCR with python to extract the text from image and produce summary of the text. The mentioned research following a pipeline of text extraction and then text summarization of that text (Joshi; 2019).

The cited research is the nation of the art of following task where this research seeking to cross a ahead via applying EasyOCR with OpenCV and also applied unsupervised extractive text summarization to produce meaningful summary. These changes can develop the research a step in advance and may full fill the future objective which is making an application of extractive textual content summary.

IV. Results and discussion:

This section of the project report is giving a brief with regards to assessment and consequence of the applied techniques

and algorithms to achieve the goal. As discussed this project implemented OpenCV for optical character recognition (OCR) and after
cleaning two extractive text rundown calculation were applied in this project, the, the
output and assessment of algorithms are being discussed on this segment of the task report.

After implemented optical individual recognition strategies at the pictures studies provided very efficient and accurate output. easyOCR techniques were applied to extract the text from an image. Along with this to check ability of the both techniques, research applied both approach on noisy images and get very convincing results. As consistent with the mission requirement this optical character popularity section become an important element to find textual content summarization. Now, providing result of easyOCR techniques after applying on the original and noisy images to get similar text as present in images

After text extraction and text pre-handling the last piece of undertaking was execution of text summarization. This section is being provided evaluation and results of the applied algorithms to generate a summary of the given image.

The project of extractive text summarization from images is divided into parts.
This project applied two techniques for OCR which are with OpenCV and then compared result of both. Text pre-processing and cleaning has been done by regular expression and natural language toolkit. The result of this project shows that easy OCR with OpenCV perfectly produced text from image but on the noisy image OCR with OpenCV act better than only ocr. Then this project went forward with the result of with OpenCV result. But, the accuracy measurement of generated summary is very challenging which need to overcome in future.

Text recognizer is utilized to distinguish and perceive characters, consequently examining and refreshing the fields of a manually written bills and receipt bills in data set. To extract the printed text and handwritten text from an image and convert into digital format automatically and update in database. The process starts by scanning an image ,compare the extracted text with trained dataset The text extraction is implemented with python using the OpenCV, Python packages. Latest version of both the tools where used as being easy OCR. Using preprocessing, segmentation technique, text extraction detects the bills successfully

V. Conclusion:

The objective of the undertaking is to find an extractive precis from a given image the use of unsupervised extractive summarization algorithms and the research is efficaciously produce extractive precis of Image the usage of a set of rules referred to as EasyOCR. Equally, EasyOCR with OpenCV and PIL perform thoroughly to extract textual content from Image. Future work of the studies is to enforce this program as an application of text summarization which can be a time saver also, we want to work at the dimension of the accuracy of output . The implementation of the proposed answer can be in addition delicate to obtain better consequences in the OCR processing. Moreover, optimization and efficient implementation of a greater range of preprocessing methods have to be taken into consideration for in addition enhancements whilst also designing and implementing a parallel processing structure considering the big time necessities for preprocessing (acquiring voting applicants) as well as for the desired intermediate steps (skew correction, template construction) observed by using the execution of the OCR engine.

We have provided the text recognition in formatted bills using some methods and Libraries .The experiment proves that the high in accuracy and efficiency compared to previous method . There is also a possibility in finding solutions for extracting text from unformatted written text and updating it automatically into the output and in the form of json format. The aim of this project is to explore the task of classifying handwritten text and images into an json format output.

References:

1.Smith, R. (2007). An overview of the EasyOCR engine, *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2, IEEE, pp. 629–633.

2.Pawar, N., Shaikh, Z., Shinde, P. and Warke, Y. (2019). Image to text conversion using EasyOCR, *Image* **6**(02)

3.Chawla, M., Jain, R. and Nagrath, P. (2020). Implementation of EasyOCR algorithm to extract text from different images, *Available at SSRN 3589972* .

4.Palekar, R. R., Parab, S. U., Parikh, D. P. and Kamble, V. N. (2017). Real time li cense plate detection using opencv and EasyOCR, *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 2111–2115.

5.Joshi, N. (2019). Text image extraction and summarization, *Asian Journal For Convergence In Technology (AJCT)* .

6.Samani, Z.R.; Guntuku, S.C.; Moghaddam, M.E.; Preo¸tiuc-Pietro, D.; Ungar, L.H. Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr. PLoS ONE 2018, 13, e0198660.

7.Singh, A.; Virmani, L.; Subramanyam, A. Image Corpus Representative Summarization. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 21–29.

8. Ozkose, Y.E.; Celikkale, B.; Erdem, E.; Erdem, A. Diverse Neural Photo Album Summarization. In Proceedings of the 2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA), Istanbul, Turkey, 6–9 November 2019; pp. 1–6.

9.Chen, J.; Zhuge, H. Extractive summarization of documents with images based on multi-modal RNN. Futur. Gener. Comput. Syst. 2019, 99, 186–196.

10.Samani, Z.R.; Moghaddam, M.E. A knowledge-based semantic approach for image collection summarization. Multimed. Tools Appl. 2017, 76, 11917–11939.

11.S.Purnamawati, D. Rachmawati,Lumanauw, R.F. Rahmat, R. Taqyuddin, "Korean letter hand writte recognition using deep convolutional neural network on android platform," J. of Physics: , Conf. Series, 2018.

12.Kumuda, T., &Basavaraj, L. (2015). Detection and localization of text from natural scene images using texture 2015 IEEE International Conference on Computational Intelligence and ComputingResearch(ICCIC). doi:10.1109/iccic.2015.7435688

13.Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). EAST: An Efficient and Accurate Scene Text Detector. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2017.283

14.Cesarini, F., Gori, M., Marinai, S., & Soda, G. (1998). INFORMys: a flexible invoice-like form-reader system. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(7), 730–745. doi:10.1109/34.689303

15.Islam, M. Z., &Mondal, A. K. (2014). Towards a standard Bangla PhotoOCR: Text detection and localization, 2014 17th International Conference on Computer and InformationTechnology. (ICCIT). doi:10.1109/iccitechn.2014.7073084

16.Bhattacharya, U., Parui, S. K., &Mondal, S. (2009). Devanagari and Bangla Text Extraction from Natural Scene Images. 2009 10th International Conference on Document Analysis and Recognition.