

Implementation to Extract Text from Different Images by Using Tesseract Algorithm

Akshay Parwar¹, Akansha Goverdhan², Apurva Gajbhiye³, Prajkta Deshbhratar⁴, Roshan Zamare⁵,
 Prasanna Lohe⁶

^{1,2,3,4 & 5}Final Year Student of Department of Computer Science and Engineering,
 J D College of Engineering and Management, Maharashtra, India

parwar.akshay25@gmail.com, akansha.goverdhan@gmail.com, apurva.gajbhiye@gmail.com,
prajkta.deshbhratar@gmail.com, roshanzamare@gmail.com

⁶Assistant Professor, Computer Science and Engineering,
 J D College of Engineering and Management, Maharashtra, India
prasanna.lohe@gmail.com

Abstract: Today's life there is a huge demand of storing the information which is available in newspaper as well as in a paper documents format into a computer. So there is a one simple way to store the information of the newspaper into computer system for that first we have to scan the paper then we can store as well as we can also do the some changes as per the requirement. It is based on the basically mobile quality, paper quality and quality of images. Image processing is one of the most growing field in research and technology in today's world. Detection of text from the images and identification of characters is a challenging problem. Image content can be divided into two types: Which contain attributes like texture, shape and other is in the form of text. It is a little bit challenging task to detect the text from the captured images.

Keywords: Android SDK, Tesseract2, Text recognition, Optical character recognition.

1. Introduction

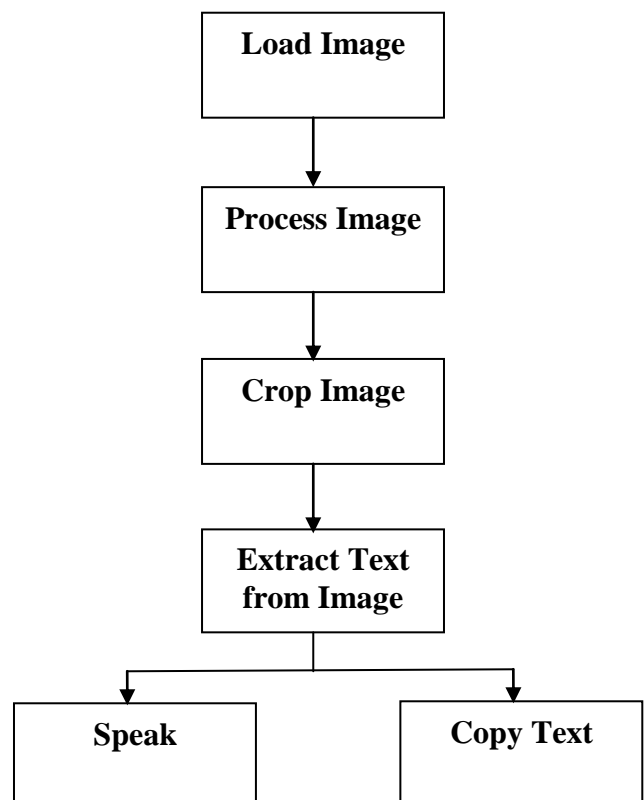
In today's generation every individual people carry the mobile phone rather than the computer. These new environments lead to new application. Optical character recognition is designed for scanned documents. It is a very powerful tool. This application is for all type of people not only for the travelers. Android mobile phones are more widely used in today's generation. Image text scanner is for those people who cannot understand other languages. In this application we can translate the text as well as we can copy, edit, and speak the text. Basically this project is designed to extract text from different images so the user can use this application anywhere. The application can run without any internet connection. In this application first we load the image from camera or gallery after that we can proceed for as per the user requirement which portion of the image they want to crop or simply edit. After editing the image it convert the extracted text into the other language. There is one other option that if the user in not able to understand the language then the user can also use the speak feature to listen to the extracted text.

2. Method

Tesseract is an open-source OCR engine that was developed at HP between 1984 and 1994. Tesseract began as a PhD research project in HP Labs, Bristol, and gained momentum as a possible software and/or hardware add-on for HP's line of flatbed scanners. The line finding algorithm is designed so that a skewed page can be recognized without having to de-skew, thus saving loss of image quality. Once the text lines have been found, the baselines are fitted more precisely using a quadratic spline. Tesseract tests the text lines to determine whether they are fixed pitch. Where it finds fixed pitch text, Tesseract chops the words into characters using the pitch, and disables the

chopper and associator on these words for the word recognition step. Part of the recognition process for any character recognition engine is to identify how a word should be segmented into characters.

3. Proposed System



3.1 Load Image:-

To begin with the process the user must load an image first. The user can choose the image either from the camera or from the gallery. If the image captured from the camera is blurry the user can take a new image again. The quality of the output depends on the quality of the image captured. The user will use the fields to indicate which image he would like to choose the process accordingly.

3.2 Process Image:-

Next it will process the image loaded by the user. It will check whether the image is clear or not. If the image is blurry or not taken properly there might not be a proper output. So the user has to be careful about that.

3.3 Crop Image:-

After the processing of image is done it will move on to the next part i.e. cropping the image. In this part the user will be able to select the part from the according to his need. The user is free to crop the required size of the image. The region selected by the user will be the region from where the text will be extracted.

3.4 Extract Text from Image:-

After cropping the image the text will be extracted from the region selected by the user. Each character in a particular language is stored in the database. The characters from the image will be processed and compared with the characters in the database. If a match is found then the character will be displayed on the output. The same process will be done for all the characters present in that region. In case if the character is not matched then that character will not be considered and the pointer will move to the next character.

3.5 Speak:-

The text extracted from the previous stage will be shown as the output of the image. And the user will be allowed to choose the speak option provided in the app. The speak option is provided in case if the user is not able to read then the user can listen to the extracted text.

4. Result



Fig: The image capture from the news paper and shows the selected image to be cropped



Fig: Shows the extracted text user can listen to the extracted text



Fig: Shows the image selected from gallery and shows the selected image to be cropped

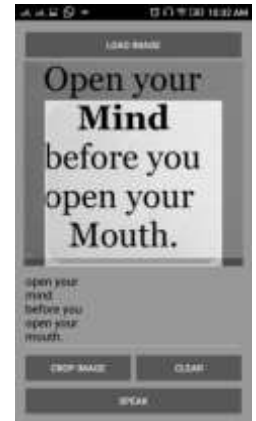


Fig: Shows the extracted text user can listen to the extracted text



Fig: The image capture from the booklet and shows the selected image to be cropped

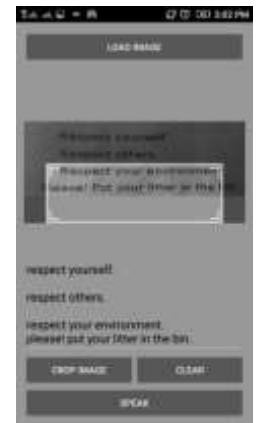


Fig: Shows the extracted text user can listen to the extracted text.

5. Conclusion

With this approach the text from the captured image or saved image are recognizable easily with less effort and more accuracy. In this way we implement many techniques which are necessary to convert image to text as well as speech. It helps to make people more intractable with the real world. It is useful for the traveller people because it is offline application so there is no need of internet connection. Old documents can also be scanned and corrected with the help of this application.

Acknowledgement

We would like to express our intense honor and deep regard to our project guide Prof. Prasanna Lohe for his guidance and constant encouragement for carrying out the project. Working under him was a great experience for us. He kept us going, and this would not have been possible without him

References

- [1] A.J.Jadhav , Vaibhav Kolhe and Sagar Peshwe “Text Extraction from Images” in International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3,pp 3 March 2013.
- [2] Pratik Madhukar Manwatkar, Dr. Kavita R.Singh “Text Recognition from Images: A Review”in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4,pp 11, November 2014.
- [3] Ray Smith “An Overview of the Tesseract OCR Engine”.
- [4] Mrunmayee Patil and Ramesh Kagalkar “A Review on Conversion of Image to text as Well as Speech using Edge Detection and Image Segmentation ” in International Journal of Science and Research,Volume 3,pp 11,November 2014.
- [5] A S N Chakravarthy , Penmetsa V Krishna Raja and Prof P S Avadhani “Hand written Text *Image Authentication Using Back Propagation*”in International Journal of Network Security and its Application,Volume 3,pp 5 September 2011.
- [6] F. Chassaing , C. Wolf, J. M. Jolion, and, “Text localization, enhancement and binarization in multimedia documents,” in Pattern Recognition, Aug. 2002, vol. 2 of Proceedings. 16th International Conference on, pp. 1037–1040
- [7] Shrey Dutta, Naveen Sankaran, PramodSankar K., C.V. Jawahar, “Robust Recognition of Degraded Documents Using Character N-Grams”, IEEE, 2012.
- [8] Chucai Yi, Student Member, IEEE, and Yingli Tian, Senior Member “Localizing Text in Scene Images by Boundary Clustering, Stroke Segmentation, and String Fragment Classification”, IEEE, Vol. 21, No. 9, September 2012.