IMAGE TO TEXT AND AUDIO CONVERSION USING TESSERACT

[1]Nisha Pawar,[2]Zainab Shaikh,[3]Poonam Shinde

[1,2,3]Department of Computer Engineering, Marathwada Mitra Mandal's Institute of Technology, Pune, India.

Abstract: In this growing age of technology, people search for methods that will make their work easy. For example, an individual wants to extract text from images and save the text to modify it for their own purpose. It can be very challenging to detect text from images due to its complex background and noise. This can be done using Optical Character Recognition with the help of Tesseract OCR Engine. The objective of this technique is to design a text detection system which will be able to detect text from images in efficient manner. OCR is well known for its use in applications to recognize text from the images. The features of this proposed system doesn't end here, the extracted text can also be converted into audio format to help visually impaired people and also normal people who prefer to hear the information instead of reading it.

***IndexTerms* - Optical Character Recognition, Text Detection, Text Recognition, Tesseract, Text-to-Speech.**

I. INTRODUCTION

In today's world, everything is digital. The things that were done in traditional ways - which used to take lot of time - are done faster and in less time with the use of technology that operates everything digitally. Many resources are available to collect textual information from documents, newspapers, faxes and much more. People usually scan documents to store the required data in their computers. Other example of storing text data can be college students, who use digital devices to scan or capture photos of notes and important information. When the document is scanned with scanner or captured by digital devices' cameras, it is stored in image format. However, the text in these images cannot be edited and it is difficult to extract the required information by an individual from the scanned image or digital photo. Also, images take more space than word files in the storage of system. Thus, there is demand for applications that can recognize characters from scanned or captured images and make them editable according to user's requirements.

In the field of Artificial Intelligence, machines are trained to learn human behaviour and act like a human brain will act in various situations. In our proposed system, an artificial intelligence technique known as Optical Character Recognition (OCR) is used. This technique is well known for text detection and to extract characters from scanned and captured documents; and use this information for editing purposes. Text detection from image is one of the crucial tasks for image processing and computer vision. Earlier methods of OCR used for telegraphy and convolutional neural networks, but they were complicated and mostly suitable for single characters. These methods had high error rate. To overcome these limitations, Tesseract OCR Engine is used. Tesseract OCR Engine makes use of Long Short Term Memory (LSTM) which is a part of Recurrent Neural Networks. It is open source, which makes it easily available and free of cost. It can recognize larger portion of text data instead of single characters. Tesseract OCR has reduced the error rate in the process of character recognition.

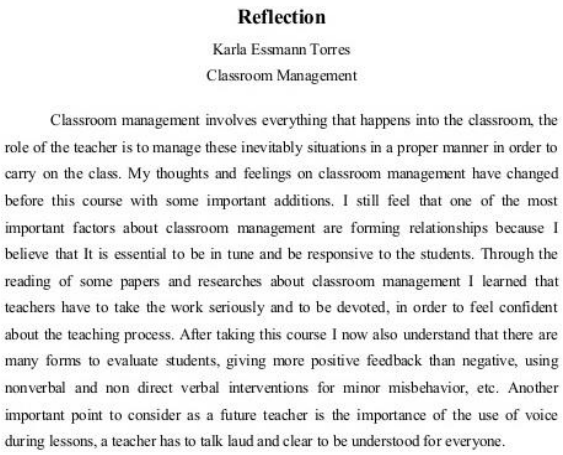
In this system, the scanned and captured images are taken as input. The input image is binary image and processing takes place step-by-step. The information can be retrieved in text file format and can be used for editing according to users requirements. Also, it is difficult for blind and visually impaired people to read the text from the document. Blind people have to use Braille to read, but it is easier for them to simply listen to audio of the data. This system can convert text data into audio format which makes it simpler for people who prefer listening rather than reading.

1.1 Different types of image

There are various types of images available. Some of them are:

1.1.1 Document Images

These images are nothing but image-format of the document. It is created by scanners or cameras. In this, image is transformed from paper based documents into image-format for electric read.



Reflection

Karla Essmann Torres

Classroom Management

Classroom management involves everything that happens into the classroom, the role of the teacher is to manage these inevitably situations in a proper manner in order to carry on the class. My thoughts and feelings on classroom management have changed before this course with some important additions. I still feel that one of the most important factors about classroom management are forming relationships because I believe that It is essential to be in tune and be responsive to the students. Through the reading of some papers and researches about classroom management I learned that teachers have to take the work seriously and to be devoted, in order to feel confident about the teaching process. After taking this course I now also understand that there are many forms to evaluate students, giving more positive feedback than negative, using nonverbal and non direct verbal interventions for minor misbehavior, etc. Another important point to consider as a future teacher is the importance of the use of voice during lessons, a teacher has to talk laud and clear to be understood for everyone.

Fig 1. Document Image

1.1.2 Scene Images

These are the images which contains text, for example, such as advertising boards, banners, etc. The scene text appears with the background part of the scene. It is very challenging to recognize and detect in this type of images, because the backgrounds are complex, contains texts in different size, styles and alignments.

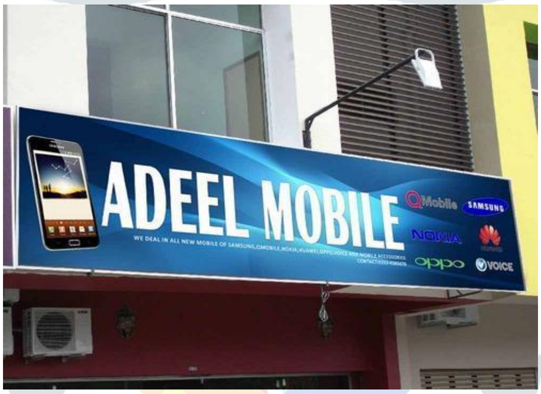

ADEEL MOBILE
QMobile
SAMSUNG
NOKIA
HUAWEI
oppo
VOICE
WE DEAL IN ALL NEW MOBILE OF SAMSUNG,QMOBILE,NOKIA,HUAWEI,OPPO,VOICE AND MOBILE ACCESSORIES

Fig 2. Scene Image

1.1.3 Born-Digital Images

This type of images are generated by computer software and saved as digital images. Compared to document images and scene images, there are more defects in born-digital images. Images are more complex, low resolutions, compression loss. Therefore, during text extraction, it is difficult to detect text from the background.

Fig 3. Born-Digital Image

1.1.4 Heterogeneous Images

This type of images contains the combination of all given images above. It can contain digital images with scene text and document text.

1.2 Existing System

There are different numbers of technologies available for Optical Character Recognition. They are mentioned as follows:

1.2.1 Connected Components Based Method

It is the technique used for detecting text from images. Connected components are extracted with the help of algorithm. Then, resulting components are partitioned into clusters. It detects pixel differences between the text and the background of the image. It can recognize and extract the characters, too.

1.2.2 Sliding Window Based Method

It is also known as text binarization process. It classifies single pixel as text or background in the textual information. This technique acts as bridge between localization and recognition.

1.2.3 Hybrid Method

This method is used for text classification. It detects and recognizes texts in CAPTCHA images. The strength of CAPTCHA can be checked. This method detects and recognizes text with a low false positive.

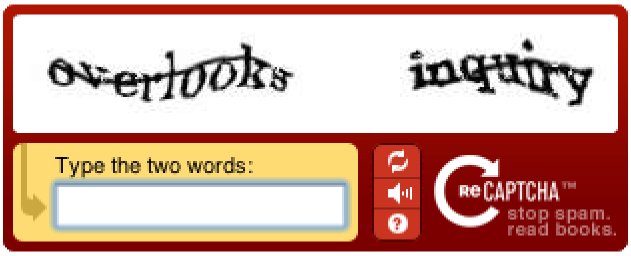


Fig 4. Captcha

1.2.4 Edge Based method

This technique is also known as image processing technique. It finds boundaries of the images or other objects in the images. It detects discontinuities in brightness. This method is used for image segmentation and data extraction in areas such as image processing and machine vision.

1.2.5 Color Based method

This method is used for clustering. It consists of two phases: text detection phase and text extraction phase. Text detection phase considers two features - homogeneous color and sharp edges, and color based clustering used for decomposing color edge map of image into several edge maps. This makes text detection more accurate. In text extraction phase, difference between text and background in image is considered.

1.2.6 Texture Based method

Another method used for detection of texts in images. It uses Support Vector Machine (SVM) to analyze the textural properties of texts. It also uses continuously adaptive mean sift algorithm (CAMSHIFT) for texture analysis. It combines both above methods to provide robust and efficient text detection.

1.2.7 Corner Based method

It is used for text extraction method. It consists of three stages: a) computes corner response in multi-scale space and thresholds it to get candidates region of text, b) verifies candidate region by combining color and size range features and, c) locates the text line using bounding box.

1.2.8 Strokes Based method

This technique is used to detect and recognize text from the video. In this method, components are extracted and grouped into text lines based on text confidence maps. It can detect multilingual texts in video with high accuracy.

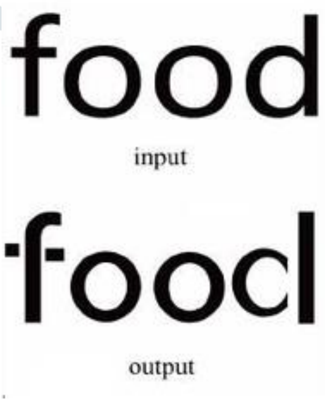


Fig 5. Stroke-Based Method

1.2.9 Semi-Automatic Ground Truth Generation Method

This technique is also used in detecting and recognizing texts from the videos. It detects English and Chinese characters of different orientations. It contains attributes like: line index, word index, script type, area and many more. It detects text from videos in an efficient way.

II. Proposed System

The proposed system is a web-based application developed from Django and Python.

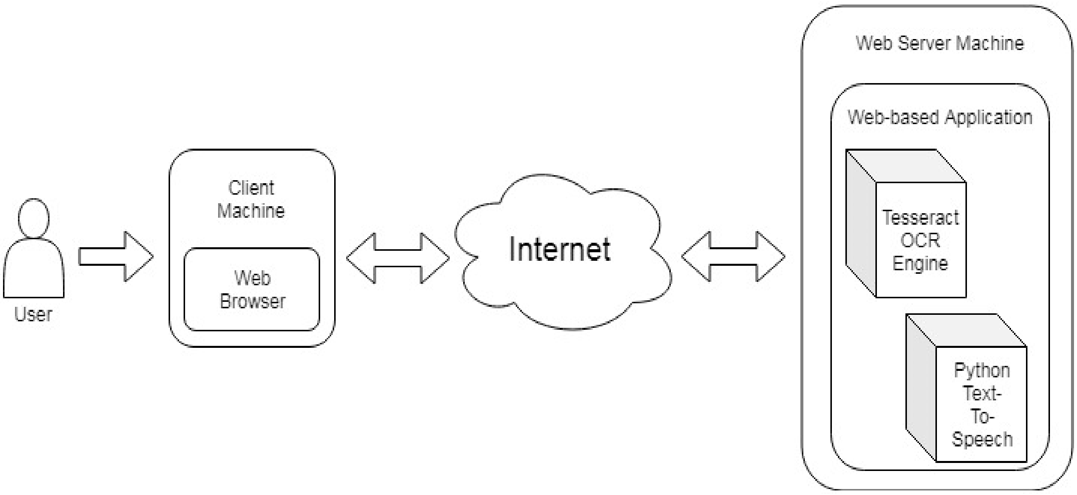


Fig 6. System Architecture

Tesseract OCR Engine is used for text detection and recognition. The system can recognize three languages, i.e., English, French and Spanish. The first step is to select a language. Then, the user has to choose an image from their device and submits it to the application. The application accepts the image and passes it to the Tesseract OCR Engine which performs the following steps:

2.1 Line Finding

Blob filtering and line construction are the main goals of this step. A simple height filter is used to remove drop-caps and characters that touch vertically. The text size in a region can be approximated with median height. With the help of median height, blobs that are smaller than it can be filtered out. To estimate baselines, a least median of square fits is used and the filtered-out blobs are assigned to approximate lines.

2.2 Baseline Fitting

A quadratic spline is used to fit the baselines more precisely. An advantage of quadratic spline is that its calculations are more stable.

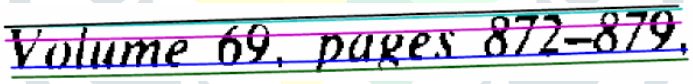


Fig 7. Baseline Fitting

2.3 Pitch Detection and Chopping

The text lines are tested to check if they are in a fixed pitch. If yes, the words are chopped into characters with the help of the pitch. In case of non-fixed pitch words, Tesseract measures the gap between the baseline and mean line. It then uses fuzzy spaces to help identify characters.

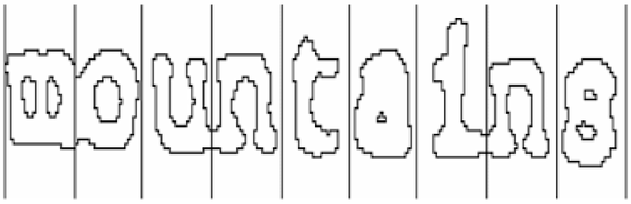


Fig 8. Chopping

2.4 Word Recognition

This is a two-pass process. An adaptive classifier is used. At first, each word is recognized in turn. These recognized words are given as input to the adaptive classifier. Then, the classifier is used for the second pass is used to recognize words that were not recognized in the first pass.

After this, the recognized text is displayed on the application. If the user wishes to modify the text, they can download it in .txt format and perform the required changed according to their needs. Then, the Python Text-To-Speech module is used to convert the recognized text to audio format that the user can listen to. It takes the output of Tesseract in the form of a string as input and uses the say() method to read the text out loud. The user can also download this audio file in .mp3 format so that they can hear it, whenever required.

III. MATHEMATICAL MODEL

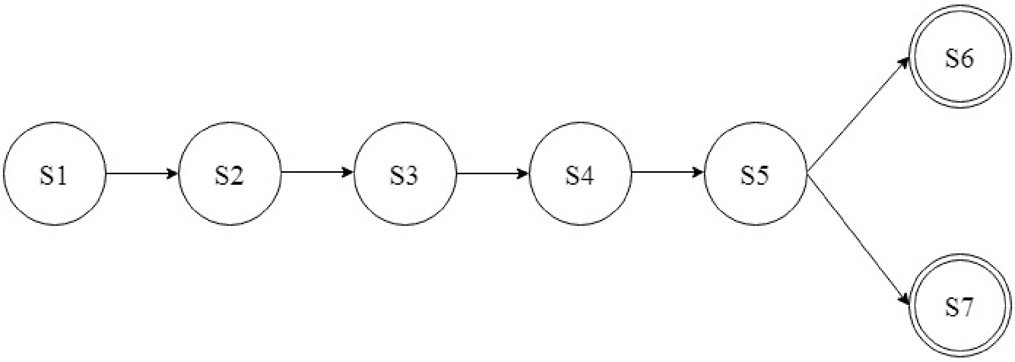


Fig 9. Mathematical Model

M = {Q, ∑, δ, q0, F}
Q = {S1, S2, S3, S4, S5, S6, S7}
∑ = {Image}
q0 = S1
F = S6, S7
δ(S1, Upload Image) = S2
δ(S2, Submit Image) = S3
δ(S3, Text Detection and Recognition) = S4
δ(S4, Listen Audio) = S5
δ(S5, Download Text File) = S6
δ(S5, Download Audio File) = S7
S6 = Downloaded data as .txt
S7 = Downloaded data as .mp3

IV. RESULTS

The main page of the web-based application is as follows:

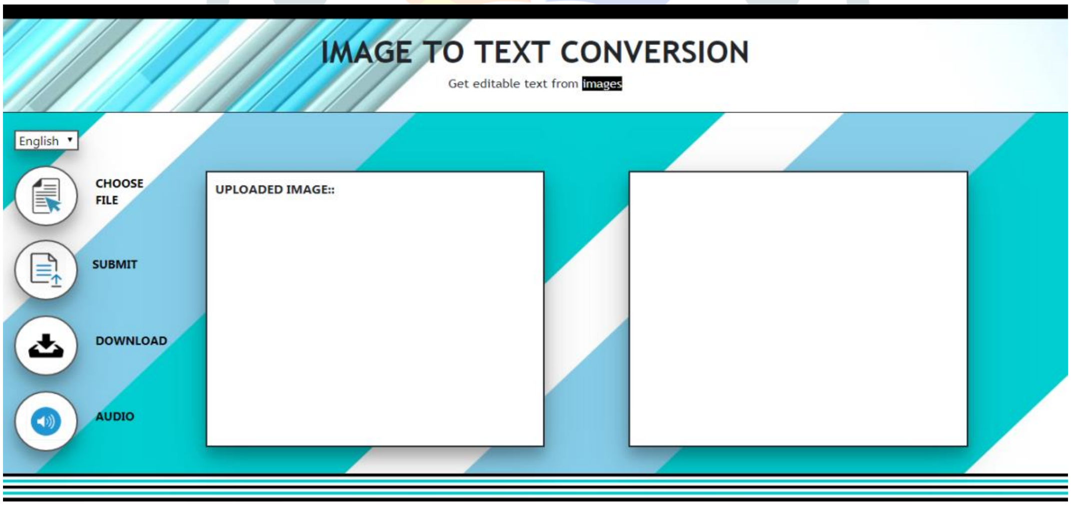
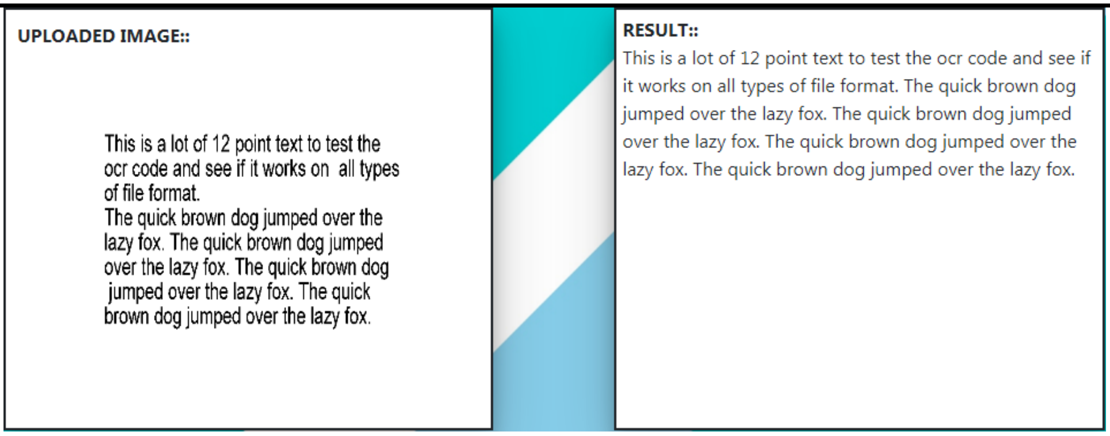


Fig 10. Home Page

The user selects a language from the drop-down list. Then, they choose an image from their computer that they wish to extract text from and click on the Submit button. The image is taken as input and the text recognized from that image is displayed as output in the right box.



UPLOADED IMAGE::

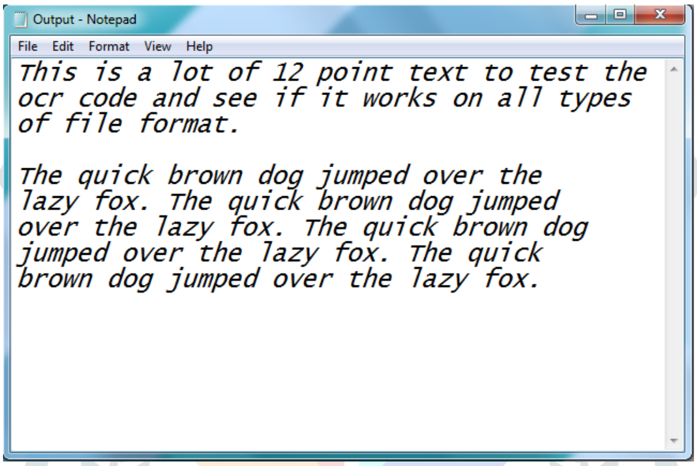
This is a lot of 12 point text to test the ocr code and see if it works on all types of file format.
The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox.

RESULT::

This is a lot of 12 point text to test the ocr code and see if it works on all types of file format. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox.

Fig 11. English Text Output

The user can click on the Download button to download the text in .txt format so as to edit it according to their requirements.



This is a lot of 12 point text to test the ocr code and see if it works on all types of file format.

The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox. The quick brown dog jumped over the lazy fox.

Fig 12. .txt File

The user can click on the Audio button to listen to the text in audio format. They can also download the .mp3 file so as to use it whenever required.

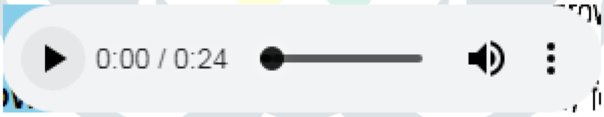


Fig 13. Audio Format

The user can select Spanish language. The steps are the same. An image with Spanish text will be given as input. Spanish text will be displayed as output. The user can download this data in text format. The user can listen to the Spanish audio file and download it.

Similarly, the system can also work with French language. Suppose, the user chooses French and uploads an image of a sign board that has French text on it. The output is displayed as French text. The user can download the text file. The user can listen to the French audio and download the same.

In this way, the system provides accurate result for all three languages.

V. Conclusion

In today's world, most of the information is in digital format such as images. It is necessary to provide the users a way that can help them process this information in a simple and easy manner. The proposed system can detect and recognize text from images. It can recognize three languages. The system uses Tesseract for the purpose of Optical Character Recognition. Tesseract has a very high accuracy rate as compared to other OCR engines. The system can convert the text to audio format by using Python Text-to-Speech. It also allows the user to download the data in text and audio formats. In this manner, the user can easily use and manage the information that they have in image format. The future enhancement to this system is to add more languages and enable it to accept more than one image for processing at a time. It can also be experimented on to make it more application specific.

ACKNOWLEDGEMENT

This research was supported by Marathwada Mitra Mandal's Institute of Technology. We express our gratitude to our colleagues who provided insight that greatly assisted the research and made our work easier, although they may not agree with all of the interpretations provided in this paper.

We are also grateful to Mr. Vivek Deshpande for assistance with Tesseract OCR Engine, and Prof. Yamini Warke, who moderated this paper and in that line improved the manuscript significantly. We have to express our appreciation to everyone who helped us, for sharing their pearls of wisdom with us during the course of this research. We are immensely grateful for their comments on earlier versions of the manuscript, although any errors are our own and should not tarnish the reputations of these esteemed professionals.

REFERENCES

[1] Pratik Madhukar Manwatkar, Dr.Kavita R. Singh, "A technical review on text recognition from images," IEEE Sponsored 9th International Conference on Intelligent Systems and Control (ISCO), 2015.
[2] Akhilesh A. Panchal, Shrugal Varde, M.S. Panse, "Character detection and recognition system for visually impaired people," IEEE International Conference On Recent Trends In Electronics Information Communication Technology, May 20-21, 2016.
[3] Nada Farhani, Naim Terbeh, Mounir Zrigui, "Image to text conversion: state of the art and extended work," IEEE/ACS 14th International Conference on Computer Systems and Applications, 2017.
[4] Azmi Can Özgen, Mandana Fasounaki, Hazım Kemal Ekenel, "Text detection in natural and computer-generated images," 2017.
[5] Sandeep Musale, Vikram Ghiye, "Smart reader for visually impaired," Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018) IEEE Xplore Compliant - Part Number:CFP18J06-ART, ISBN:978-1-5386-0807-4; DVD Part Number:CFP18J06DVD, ISBN:978-1-5386-0806-7.
[6] Christian Reul, Uwe Springmann, Christoph Wick, Frank Puppe, "Improving OCR accuracy on early printed books by utilizing cross fold training and voting," 13th IAPR International Workshop on Document Analysis Systems, 2018.
[7] U. Springmann and A. Ludeling, "OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus," Digital Humanities Quarterly, vol. 11, no. 2, 2017.
[8] Ray Smith, "An overview of the tesseract OCR engine," 2005.
[9] J. C. Handley, "Improving OCR accuracy through combination: A survey," in Systems, Man, and Cybernetics, IEEE, 1998.
[10] F. Boschetti, M. Romanello, A. Babeu, D. Bamman, and G. Crane, "Improving OCR accuracy for classical critical editions," Research and Advanced Technology for Digital Libraries, pp. 156–167, 2009.
[11] Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru, "Show and tell: A neural image caption generator", In CVPR, 2015.
[12] Rafeal C. Ginzalez and Richard E. Woods, "Digital Image Processing", Pearson Education, Second Edtition, 2005.
[13] Shirly Edward A.,Jothimani A.,JayaprakashV.,JoeBenhur Xavier F.,"Text-to-Speech Device for Visually Impaired People", International Journal of Pure and Applied Mathematics, 2018.
[14] K. Lakshmi, T. Chandra Sekhar Rao, "Design And Implementation Of Text To Speech Conversion Using Raspberry PI", International Journal of Innovative Technology and Research, 2016.
[15] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao, "Multi-Orientation Scene Text Detection with Adaptive Clustering", IEEE Transactions On Pattern Analysis And Machine Intelligence, VOL. 37, NO. 9, 2015, pp. 1930-1937.
[16] K. Wang, B. Babenko, and S. Belongie."End-to- end scene text recognition", International conf. on Computer Vision, pp.1457-1464, Vol. 10, 2011.