A Faster RCNN Based Image Text Detection and Text to Speech Conversion

Abitha A, Lincy K (Assistant Professor)
Department of Electronics and Communication Engineering
JCET
Palakkad, India

Abstract—*The reading of text contained in images plays an important role in understanding the contents of images. Text found in images contain important contents for information indexing and retrieval, structuring and automatic annotation of images. Hence text detection is the crucial stage of analyzing the images and is a well-known problem in the computer vision research area. Text detection is a very challenging task due to the variations in text size, font, style, orientation, alignment and complex background. The goal of this system is to detect the text regions in images accurately and convert the detected text to speech. The text to speech conversion process is done after text recognition from the detected text regions. In this system, a technique based on faster region based convolution neural network is proposed for image text detection. Then the detected text is converted to speech using MATLAB.*

Keywords—*Faster RCNN (Region based Convolutional Neural Network); text recognition; text to speech conversion*

I. INTRODUCTION

The advancements in the multimedia technology have given a lot of attention to the captured images from the computer vision community. Understanding these images through its contents help in clear understanding the information present within. A text is one of the contents in images that carries semantic information, and thus it helps to provide the scene description of an image. Hence, the detection of graphics text has been widely used in content based image indexing and retrieval. Effective text detection from images can enhance the performance of numerous multimedia applications, e.g. mobile visual searches, automatic sign translation and content-based image retrieval. A series of international scene text detection competitions has been successfully organized to drive the research of scene text detection.

The character recognition or character recognition system converts image text or scanned text into a computer format text. That is, it is the technique in which characters are recognized from images digitally. The recognized character is saved as text file. The Speech Application Programming Interface is developed by Microsoft. The Speech Application Programming Interface or SAPI is an Application Programming Interface that permits the use of speech synthesis and speech recognition within Windows applications. With the help of SAPI, text to speech conversion is performed.

II. PROPOSED SYSTEM

The block diagram for the implementation of the proposed system is shown in fig 1. The input image given to the proposed system can be in colour or gray scale format.

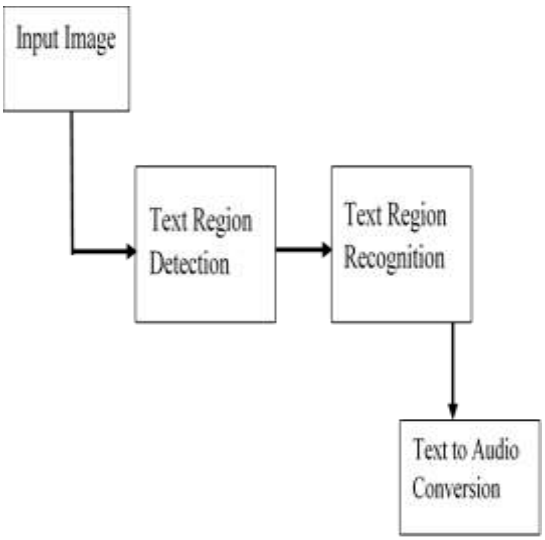

Fig 1: Block diagram of proposed system

The various steps involved in this faster RCNN based image text detection and text to speech conversion system are:

- Text region detection
- Text recognition
- Text to speech conversion

A. Text Region Detection

The text region detection is the stage at which the regions having the chance of containing the text in an image is proposed. Text found in images contain

important contents for information indexing and retrieval, structuring and automatic annotation of images. Hence text detection is the crucial stage in analysing the images. The text detection technique employed in this system is based on faster Region based Convolutional Neural Network or faster RCNN. The Faster R-CNN text detection system is composed of two parts. A Region Proposal Network or RPN [2] part and Fast RCNN part. The first part is a deep fully convolutional network that proposes regions, and the second part is the Fast R-CNN detector that uses the proposed regions. The entire system for text detection is thus, a single, unified network. The basic steps involved in the proposed text detection system are as follows:

1. Initially, the input image goes through a convolution network. The convolutional network outputs a set of convolutional feature maps on the last convolutional layer. The convolutional network [3] architecture used is VGG 16 [4].

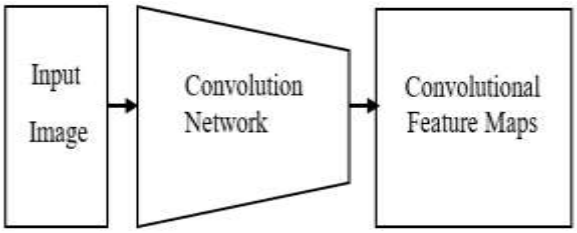

Fig 2: Convolutional feature map generation

2. Then, on the generated feature maps of the image, a sliding window is run spatially. The sliding window used is of the size3*3. For each sliding window, a set of 9 anchors are generated at an image position. The generated anchors have the same centre but with 3 different aspect ratios and 3 different scales. This is shown in fig 3. Here, coordinates of all anchors are computed with respect to the original image.

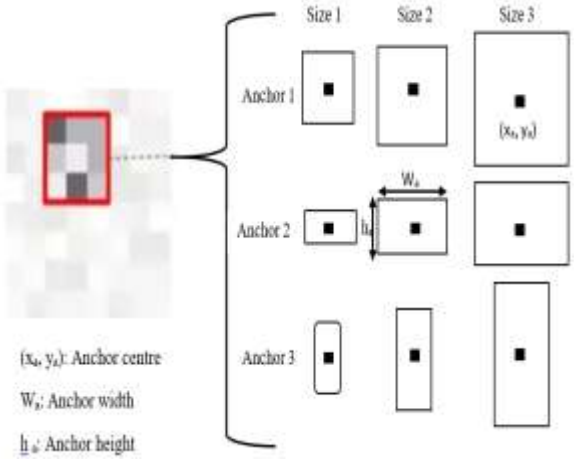

Fig 3: Anchor generation

Moreover, for each of these generated anchors, a value is computed which indicated how much these anchors overlap with the ground-truth bounding boxes. This overlapping probability is found as per the following equation:

$$P^* = \begin{cases} 1 \text{ if IoU} > 0.7 \\ -1 \text{ if IoU} < 0.3 \end{cases} \quad (1)$$

Where,

$$IoU = \frac{Anchor \cap GTBox}{Anchor \cup GTBox} \quad (2)$$

3. Finally, the spatial features extracted from the convolution feature maps (as shown in figure 3 within red box) are fed to a smaller network. This smaller network performs two tasks. The tasks that these network performs are regression and classification. The output of regressor determines a predicted bounding-box. The output of classification sub-network is a probability indicating whether the predicted box contains text or it is from background.

So at each location of the convolution layer, the bounding box regression provides the bounding box offsets for each anchor box and the classification layer provides the probability scores or confidence scores that represents whether the text region is present or not within each anchor box. Only those anchor boxes with a corresponding high probability of text being present are further processed. Thus the final proposals at each location are the anchor boxes and the box offsets with a high probability of containing text. The algorithm is as follows:

- Load the detection network
- Read the input image
- Find the bounding box values of text regions from the input image using detection network
- Expand the bounding boxes by a small amount
- Clip the bounding boxes to be within the image bounds
- Compute the overlap ratio between bounding boxes. Also, set the overlap ratio between a bounding box and itself to zero
- Find the connected text regions and thus draw the final text box

B. Text Recognition

Text recognition is the technique that is used in handwriting analysis programs on cellphones. It is also used as in the gigantic mail-sorting machines which ensures all those millions of letters reach their destinations. The text recognition stage is that performs the mechanical or electronic conversion of text in images into machine-encoded text. The text recognition from the detected input region is performed using image processing technique in MATLAB. The ocr function provided from the Computer Vision System Toolbox can be used to perform Optical Character Recognition. The various computer vision applications such as document

analysis, image search, and robot navigation use text recognition in images. Text recognition functionality can be added to a wide range of application by ocr function. The basic steps while performing OCR [5] are:

- Load an image
- Perform OCR
- Display the recognized words

But if the text is on a non-uniform background, it is actually challenging for the OCR. So an rgb to gray scale conversion is performed on the image. Then binarization of the image is performed using Otsu's method [6]. Certain pre-processing techniques are performed to remove the background variations and improve the text segmentation. These pre-processing steps include:

- Use morphological opening to estimate the background
- Subtract the background image from the original image
- Increase the image contrast

Then, on the pre-processed image, thresholding is performed to create a new binary image. The noise is removed from background with bwareaopen command. Even after removing the background variation, Additional pre-processing is performed to remove the artifacts. This additional pre-processing is performed using morphological reconstruction. Thus a cleaner image is produced for OCR.

C. Text To Speech Conversion

In this process, the recognized text is converted into speech using MATLAB. Using this approach text from a web page, e-Book or word document, can be read and can thereby generate synthesized speech output. The speech output is made through speakers of computer. A TTS [7] synthesizer consists of Natural Language Processing module (NLP) and a Digital Signal Processing module (DSP) as functional blocks. Natural Language Processing module produces phonetic transcription of the text that is read. This is performed with the desired intonation and rhythm. This is termed as prosody. DSP module transforms the symbolic information that it receives into speech. Speech enabled applications can be performed using Microsoft Win 32 SAPI library. Microsoft Win 32 SAPI library retrieves the voice and audio output information available in computer.

The steps involved in text to speech conversion system are as follows:

- Initially, check whether the Win 32 SAPI library is available in the computer. If Win 32 SAPI is not available in the computer, then an error will be generated and it should be loaded in the computer.
- This step will be executed only if the Win 32 SAPI file is available in the computer. Then in this step a new server for this file is made by actxserver command
- From the Win 32 SAPI library, the voice objects are received by invoke command
- Comparison of the input string with Win 32 SAPI string is made with strcmp command
- The voice is extracted by selecting the voice which are available in Win 32 SAPI library
- By using the commands, nargin and nargout, the pace of the voice is chosen
- The wave player is initialized for converting the text into speech by convert uint8 to double precision.
- Thus obtains the speech output

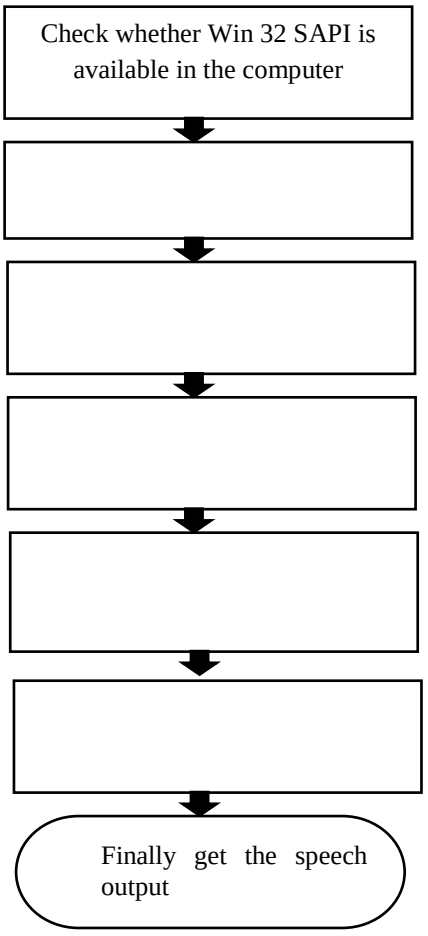

Fig 4: Flowchart

The flow chart for text to speech conversion system is shown in fig 4.

III. RESULT

The Faster RCNN based image text detection and text to speech conversion system is developed in MATLAB and evaluated for various images on computer. The proposed method have been done on Intel(R) Core(TM) i3-6006U CPU, 2.0 GHz with 4 GB RAM under Matlab R2017a. The training of text

detection network was performed using GT 1030M GPU environment running Windows 10 and several experiments are carried out to evaluate the proposed technique.

Fig 5: Input image

Fig 6 shows the output of text region detection for the input image in fig 5.

Fig 6: Text region detection

Fig 7 shows the recognized text. Finally the recognized text is obtained as speech output.

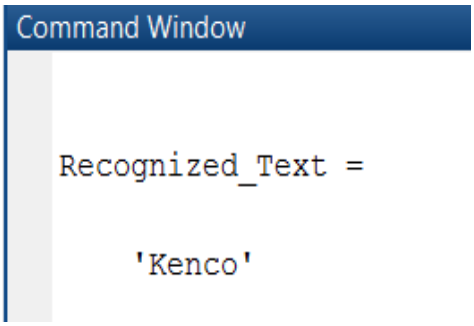

Fig 7: Recognized text

IV. CONCLUSION

This paper proposes a faster convolutional neural network based approach for image text detection and text to speech conversion. By this approach text regions can be detected from images. Also, the detected text can be converted to speech after performing text recognition.

Acknowledgment

At the outset, I thank God almighty for making my endeavour a success. I am very much grateful to Dr. V. P Sukumaran Nair, the Principal of our college for supporting all the way long. I also express my gratitude to Prof.C.Venugopal, Head of the Department of Electronics and Communication Engineering, JCET for providing me with adequate facilities, ways and means by which I was able to complete the project. I express my sincere gratitude to my guide Ms. Lincy. K, Assistant Professor, Department of Electronics and Communication Engineering, JCET for her constant support and valuable suggestions without which the successful completion of this project would not have been possible. I express my immense pleasure and thankfulness to all the teachers and staff of the Department of Electronics and Communication Engineering, JCET for their cooperation and support. Last but not the least I thank all others and especially my classmates and my family members who in one way or another helped us in the successful completion of the project.

References

[1] Jian Sun, Ross Girshick, Kaiming He, Shaoqing Ren, "Faster R-CNN: Towards Real- Time Object Detection with Region Proposal Networks" Microsoft Research {v-shren, kahe, rbg, jiansun}@microsoft.com

[2] Ross Girshick, Jeff Donahue, Trevor Darrell , Jitendra Malik "Region-based Convolutional Networks for Accurate Object Detection and Segmentation" DOI 10.1109/TPAMI.2015.2437384, IEEE Transactions on Pattern Analysis and Machine Intelligence

[3] Fei-Fei Li, Justin Johnson, Serena Yeung, "Convolutional Neural Networks", Lecture 5-1, April 18, 2017

[4] Karen Simonyan and Andrew Zisserman, Visual Geometry Group, Department of Engineering Science, University of Oxford, "Very Deep Convolutional Networks For Large-Scale Image Recognition", ICLR 2015

[5] Ray Smith. Hybrid Page Layout Analysis via Tab-Stop Detection. Proceedings of the 10th international conference on document analysis and recognition. 2009

[6] Miss Hetal J. Vala, Prof. Astha Baxi, "A Review on Otsu Image Segmentation Algorithm:", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 2, February 2013

[7] Thierry Dutoit, Milos Cernak, "TTSBOX: A Matlab Toolbox For Teaching Text-To-Speech Synthesis", IEEE, ICASSP 2005