A Survey on Object Detection with Voice

Prinsi Patel[1], Prof. Barkha Bhavsar[2]

[1]MTECH Student, Dept. of computer Engineering, LDRP Institute of Technology and Research, Gandhinagar, Gujarat, India 382015
[2]Professor, Dept. of computer Engineering, LDRP Institute of Technology and Research, Gandhinagar, Gujarat, India 382015

---***---

Abstract - *object detection systems have been growing in the last few years for various applications. Since the hardware can not detect the smallest objects. Many algorithms are used for object detection like Yolo, R-CNN, fast R-CNN, faster R-CNN ,etc. object detection using YOLO is faster than other algorithms and the YOLO scans the whole image completely at one time. Object detection, which is based on Convolutional Neural Networks (CNNs) and it's based on classification and localization. The main motive of this survey is that the smallest amount of objects can be detected object and labeling the object with voice for real time object detection. An object is detected by extracting the features of an object like color of the object, texture of the object or shape or some other features. Then based on these features, objects are classified into many classes and each class is assigned a label. When we subsequently provide an image to the model, it will output many objects it detects, the location of a bounding box that contains every object with their label and score indicates the confidence. Text-To-Speech (TTS) conversion is a computer- based system which require for the label are converted text-to-speech.*

***Key Words*: Object Detection, Object Recognition, Text-to-Speech Convert, You Only Look Once(YOLO),CNN, R-CNN.**

1. INTRODUCTION

In previous research, there are various algorithm to the detected object with their label. Object detection is the combination of image classification and object localization. In image classification is used for classify or predict the class of specifying the object in an image. In image classification main goal is accurately to identify the feature of an image. In object localization is locate the object on an image with the boundary box. Object detection is highly capable to deal with multi-class classification and localization.

object detection can be broken down into machine learning-based approaches and deep learning-based approaches for object detection and recognition,[1] such as Support vector machine (SVM), Convolutional Neural Networks (CNNs), Regional Convolutional Neural Networks (R-CNNs), You Only Look Once (YOLO) model etc., Since machines cannot detect the objects in an image instantly like humans, it is really necessary for the algorithms to be fast and accurate and to detect the objects in real-time object detection and recognition.

In real world there are many object detection systems available and they are providing such an accurate result too. By this survey, we are trying to detect the smallest amount of object of accuracy and label with voice. Text-To-Speech (TTS) conversion is a computer- based system that divide the two module image processing and voice processing module. In image processing module, optimal character recognition(OCR) has convert the .jpg to .txt format. OCR has recognize the character automatically. In vice processing module has convert .txt to speech.

This paper will give a literature survey on object detection and detected object convert the text-to speech. Different types of object detection algorithm like CNN, R-CNN, fast-CNN, faster-CNN, YOLO.

2. ALGORITHMS

2.1. YOLO

You look only once(YOLO) is an object detection algorithm that targets the real time application. In YOLO is the predict the boundary boxes and confidence score. YOLO algorithm is based on regression which predicts the boundary boxes and probabilities for each region of the whole image at one time. YOLO trains on full images and directly optimizes detection performance. In YOLO, first we take the input image and the image is resized according to the algorithm. The image is passed through the 24 convolution network layer followed by two fully connected layers. After that the non-maximum suppression apply to the image. The output becomes a detected object that shows the boundary box and confidence.

2.2. Faster R-CNN

The faster region convolutional neural network(R-CNN) is another state of CNN-based deep learning object detection approach. It main part of the training time of the whole architecture. Faster R-CNN is updated version of fast R-CNN. The main difference is that faster R-CNN used a region proposal network(RPN) and fast R-CNN used a region of interest(ROI). In faster R-CNN, firstly we take a input image pass through a convNet which returns the featured map. Region proposal network applied on featured map which returns the object proposal along with objectness score. A RoI pooling layer is applied on these proposals and all the proposals to the same size. Proposal are passed through fully connected layer which contain softmax layer and linear regression layer at the top.

2.3. Fast R-CNN

In fast R-CNN, first we take the input image forward to convNet which returns the region of interest from the proposed method. And then the apply the Rol pooling layer to the extracted regions of interest to check all the regions are of the same size. The regions are passed through a fully connected layer which returns the boundary box using a softmax classifier. A softmax layer is used on the top layer which has fully connected to the network to outputclasses. Along with the softmax layer, a linear regression layer is also used parallelly to output bounding box coordinates for predicted classes.[3]

2.4. R-CNN

In R-CNN, check the number of boxes that contain any object. R-CNN uses selective search to extract these boxes from an image.[3] R-CNN algorithm to detect the object, first we take an input image and then we get regions of interest (ROI) from a proposed method. All the regions are reshaped as per CNN input and forward to convNet. Each region is extracted through features using CNN. Regions are divided into different classes with SVMs. And apply the boundary-box regression to classify the region with SVMs.

3. DIFFERENT APPROACHES TO OBJECT DETECTION

3.1. Real time object detection with YOLO[4]

Authors: Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam

Publication: International Journal of Engineering and Advanced Technology (IJEAT) 2019

Technique/Method: CNN, F-CNN, YOLO

Description: - In this paper, the main objective is to detect objects using You Only Look Once (YOLO) approach. This approach has various advantages as compared to other object detection algorithms. In other algorithms like Convolutional Neural Network, Fast- Convolutional Neural Network the algorithm will not scan the whole image completely but in YOLO the algorithm scan the whole image completely by predicting the bounding boxes using convolutional network and the class probabilities for these boxes and detects the image faster as compared to other algorithms.[4] To find the objects in the image we use Object Detection and have to find more than one object in real-time systems.

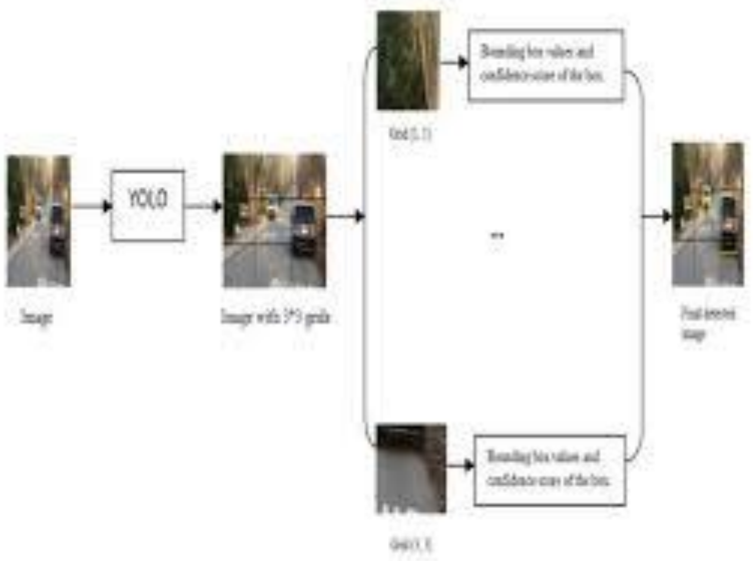

Bounding box values and confidence score of the box.
Grid (1,1)
Image
YOLO
Image with 3*3 grids
...
Final detected image
Bounding box values and confidence score of the box.
Grid (3,3)

Figure 1: Working Of YOLO[4]

In Figure shows the image is divided into grids of 3x3 matrixes. Depending on the complexity of the image, We can divide the image into any number of grids. Once the image is divided, each grid goes through to the classification and localization of the object. The confidence score or objectness of each grid is found. If there are not proper objects found in the grid, then the objectness and bounding box value of the grid will be zero otherwise the objectness will be 1 and the bounding box value will be its corresponding bounding values of the found object.

To YOLO algorithm, the main motive is the predict a class of object with their boundary box specifying the object. All the bounding box can be described using four coordinates:

1. center of a bounding box (**bxby**)
2. width (**bw**)
3. height (**bh**)
4. value **c**is corresponding to a class of an object (such as: bicycle, traffic lights, etc.).

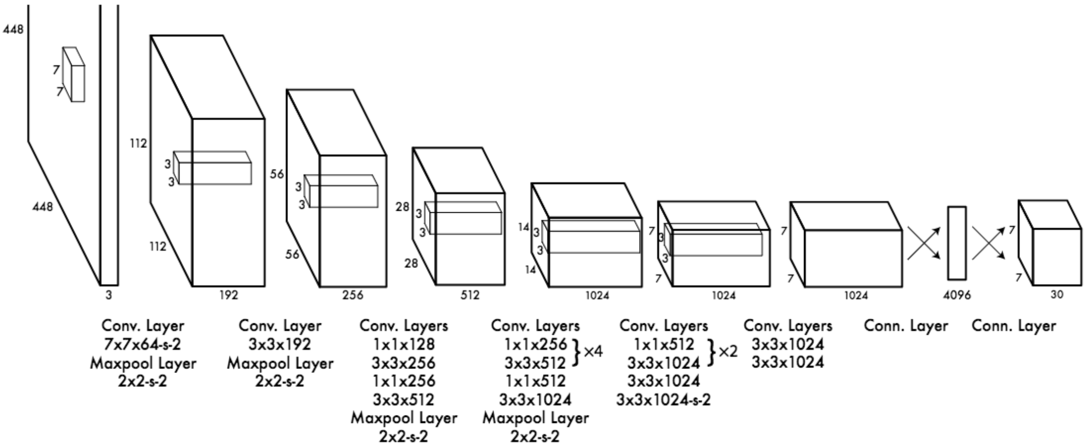


Figure 2: CNN Network Design[4]

In this Figure shows, The network has 24 convolutional layers followed by 2 fully-connected layers. Reduction layers with 1x1 filters followed by 3x3 convolutional layers replace the initial inception modules. The final layer outputs a S*S*(C+B*5)[4]tensor corresponding to the predictions for each cell of the grid. C is the number of assume probabilities for each class. B is the fixed number of anchor boxes each cell, each of the boundary boxes is related to 4 coordinates (coordinates of the center of the box, width and height) and a confidence value.

RESULTS

The YOLO is to make a Convolutional neural network to predict a (7, 7, 30) tensor. It uses a Convolutional neural network to scale back the spatial dimension to 7x7 with 1024 output channels at every location. image classification and object techniques are applied for each grid of the image and each grid is assigned with a label. Then the algorithm checks each grid separately and marks the label which has an object in it and also marks its bounding boxes. The labels of the gird without object are marked as zero.[4]

3.2. Implementation of Text to Speech Conversion[5]

Authors: Chaw Su Thu Thu1, Theingi Zin

Publication: International Journal of Engineering Research & Technology (IJERT) 2014

Technique/Method: Text-to-speech **(**TTS), Optical Character Recognition(OCR)

Description: -

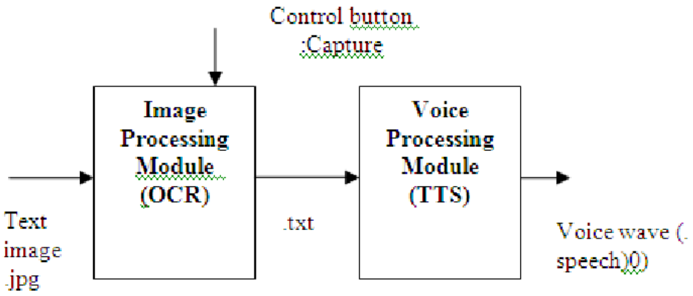


Figure 3: Block Diagram of TTS[5]

In above image shows the block diagram of Text-To-Speech device, first is image processing module, where OCR converts .jpg to .txt form. Second is voice processing module which converts .txt to speech.

There are mainly two parts: 1) Optical Character Recognition System for Paper Text 2) Text to Speech Conversion

1. Optimal Character Recognition System

In this part, there are three portions as described in the follow:

- Template file Creation
- Creating the Neural Network
- Character Recognition

In Template file creation Letter A to Z and number 0 to 9 images are collected. By using step 1 to 5 which described in the character recognition section, particular image is changed into 5 x 7 character representation in single vector and that data are saved as data file for training in neural network.

In Creating the neural network, A feed forward neural network is used to set up for pattern recognition with 25 hidden neurons. The weights and biases of the network to be ready for training after creating the network. The goal is assigned between the range of 0.01 and to 0.05. The created Neural Network is trained by using data file and target file. until the performance reaches to goal the neural network has to be trained by adjusting weight and bias of network.

The following steps are implemented for character recognition.[5]

- Firstly acquire the image and then analyze
- Second step is preprocessing step. In this step the image is transform into gray scale. Then this gray image is converted into black and white image (binary image).
- Find character image of the boundary and Crop the image to the edge.
- In this step Character is extracted and resized and the letters are resized according to templates size.
- The resized binary image is convert according to the 5 x 7 character representation in single vector.
- Load templates can be matched the letters with the templates.
- Text.txt file open for write.
- Write combine the letters in text.txt file.

2. Text to Speech conversion

The character image is converted into text and then text into speech. The algorithm is followed.[5]

- First of all check the condition Win 32 SAPI library loaded in the computer or not. If it is not available then error will be generated.
- Win 32 SAPI is provide the voice of the object.
- Compares the input string with Win 32 SAPI string.
- Firstly select the voice which are available in library and then extract the voice.
- Choose the speed of voice.
- The wave player has convert the text into speech.
- Finally, the output get the speech for given image.

3.3. Object Detection System for Blind with Voice Command and Guidance[6]

Authors: Moonsik kang

Publication: IEIE Transactions on Smart Processing and Computing

Technique/Method: SSD, YOLO, TTS-based technology

Description: - According to this paper, Object recognition algorithms are designed based on the Single Shot MultiBox Detector (SSD) structure, an object recognition deep learning model, to detect objects using a camera. SSD to predict all at once time the bounding boxes and the class probabilities with a end-to-end CNN architecture.

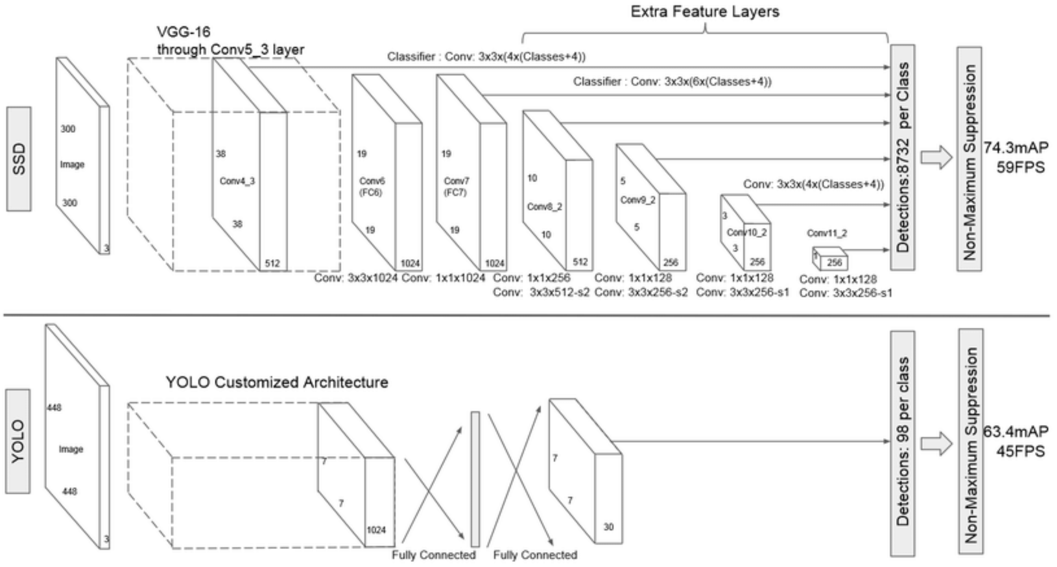


Figure 4: comparison between SSD and YOLO network Architecture[6]

The image as input which passes through multiple convolutional layers with different sizes of filter (10x10, 5x5 and 3x3). Feature maps from convolutional layers at different position of the network are used to predict the bounding boxes. They are processed by a specific convolutional layers with 3x3 filters called extra feature layers to produce a set of bounding boxes similar to the anchor boxes of the Fast R-CNN.

3.4. You Only Look Once: Unified, Real-Time Object Detection[7]

Authors: Joseph Redmon∗, Santosh Divvala∗†, Ross Girshick¶, Ali Farhadi∗†

Publication: arXiv:1506.02640v5(9 May 2016)

Technique/Method: R-CNN, Deformable parts models (DPM), YOLO

Description: - According to this paper, Unified architecture is faster and divides the components of object detection into a single neural network. Our network uses features from the entire image to predict each bounding box and also predicts all bounding boxes across all classes for an image simultaneously. It means our network reasons globally about the full image and all the objects in the image. We reframe object detection as a single regression problem, straight from image pixels to bounding box coordinates and class probabilities.by looking at the image you can predict, what objects are present and where they are using this system.

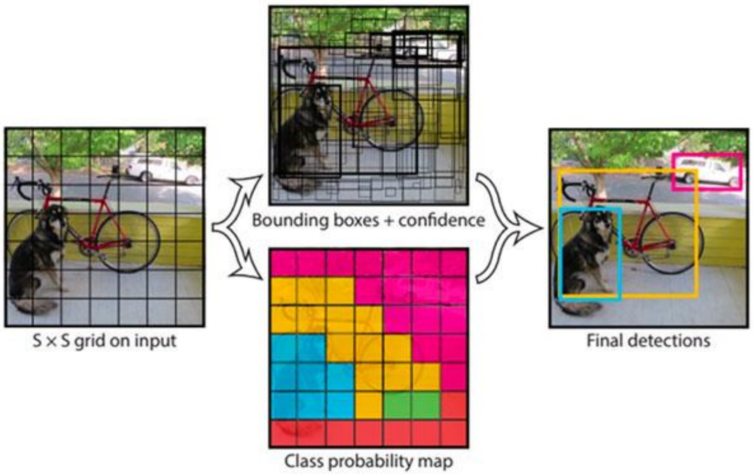
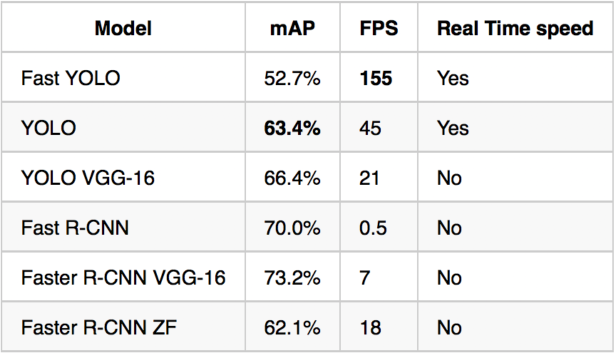


Figure 6: YOLO Model[7]

First of all the model takes an image as input then divides it into an SxS grid. Each cell of this grid predicts confidence score and B bounding boxes. This confidence score probability to detect the object are multiply by the IoU between the predicted and ground truth boxes. The predicted bounding boxes contained an object. The end of the network applied non-maximum suppression method. The highly-overlapping boundary boxes can be merging of a same object into a single one.

Comparison to other Real system:



Model	mAP	FPS	Real Time speed
Fast YOLO	52.7%	**155**	Yes
YOLO	**63.4%**	45	Yes
YOLO VGG-16	66.4%	21	No
Fast R-CNN	70.0%	0.5	No
Faster R-CNN VGG-16	73.2%	7	No
Faster R-CNN ZF	62.1%	18	No

Figure 7: comparison on different model

RESULT:

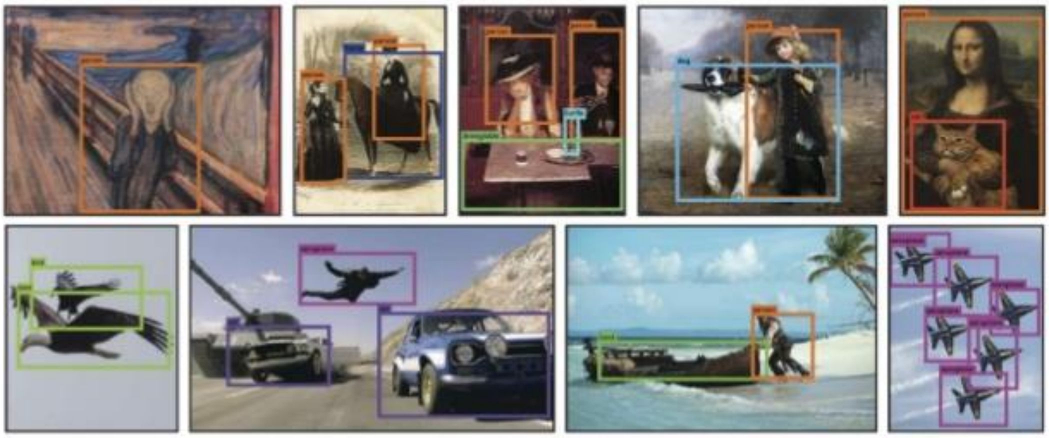


Figure 8: Qualitative Results[8]

In Figure shows the YOLO running on sample artwork and natural images from the internet. We connect YOLO to a webcam and verify that it maintains real-time performance, including the time to fetch images from the camera and display the detections. The resulting system is interactive.

4. RESULTS

Object Detection Or Recognition by using different approaches is studied in this survey paper. The below table show the overview of comparison between different approaches.

Sr.	Papers	Datasets	Method/Architecture	Limitation	Output
1.	Real-Time Object Detection with Yolo[4]	-	You only look once(YOLO)	The algorithm is simple to build and can be trained directly on a complete image. Region proposal strategies limit the classifier to a particular region.	YOLO accesses to the entire image in predicting boundaries. And also it predicts fewer false positives in background areas. Comparing to other classifier algorithms this algorithm is much more efficient and the fastest algorithm to use in real time.
2.	You Only Look Once: Unified, Real-Time Object Detection[7]	PASCAL VOC	You only look once(YOLO), Convolution neural network(CNN)	YOLO imposes strong spatial constraints on bounding box predict the each grid cell only predicts two boxes and can only have one class. This spatial constraint limits the number of objects that our model can predict.	The single neural network predicts bounding boxes and class probabilities directly detect whole images in one evaluation. The whole detection pipeline is a single network, it can be optimized end-to-end detection performance directly.
3.	Object Detection and Tracking using Tensor Flow[8]	Common object in context (COCO)	openCV, TensorFLOW, CNN	This method we will be using Tensor Flow and OpenCV library and CNN algorithm will be used and we will be labelling the detected layers with accuracy being checked at the same time.	It utilize tensorflow to join information from different sources and our joint improvement strategy to prepare all the while on COCO. This is a solid advance towards shutting the dataset measure hole between recognition also characterization.
4.	Implementation of Text to Speech Conversion[5]	-	Text-To-Speech (TTS), Optical Character Recognition(OCR)	OCR system is implemented for the recognition of capital English character A to Z and number 0 to 9. Each character is recognized at one time.[5]	The recognized character is saved as text file. In this work a text-to-speech conversion system that can get the text through image and directly input in the computer then speech through that text using MATLAB.[5] It is cost effective user friendly image to speech conversion system.

Table 1: overview of comparative study

5. CONCLUSION

The main motive of this Survey and compare the different algorithm for object detection with voice. In this survey paper has details of object detection approaches in section 2. After a complete survey and comparative study of different algorithm, it is conclude the accurate results of object detection using YOLO are high compare to others. Object detection with YOLO library which take less time for object detection and highly accurate.

REFERENCES

[1] https://www.researchgate.net/publication/337464355_OBJECT_DETECTION_AND_IDENTIFICATION_A_Project_Report

[2] https://www.researchgate.net/publication/310769942_Object_Detection_using_Image_Processing

[3] https://www.analyticsvidhya.com/blog/2018/10/a-step-by-step-introduction-to-the-basic-object-detection-algorithms-part-1/

[4] Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam "Real-Time Object Detection with Yolo", International Journal of Engineering and Advanced Technology (IJEAT)ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019

[5] Chaw Su Thu Thu, Theingi Zin "Implementation of Text to Speech Conversion", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 3 Issue 3, March – 2014

[6] Moonsik Kang "Object Detection System for the Blind with Voice Command and Guidance", IEIE Transactions on Smart Processing and Computing, vol. 8, no. 5, October 2019

[7] Joseph Redmon, Santosh Divvala, Ross Girshick, "You Only Look Once: Unified, Real-Time Object Detection",The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[8] R. Sujeetha, Vaibhav Mishra "Object Detection and Tracking using Tensor Flow", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-1, May 2019

[9] YOLO Juan Du1,"Understanding of Object Detection Based on CNN Family",New Research, and Development Center of Hisense, Qingdao 266071, China.

[10] S. Venkateswarlu , D. B. K. Kamesh , J. K. R. Sastry and Radhika Rani "Text to Speech Conversion", Indian Journal of Science and Technology, Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/102967, October 2016

[11] https://www-geeksforgeeks-org.cdn.ampproject.org/v/s/www.geeksforgeeks.org/r-cnn-vs-fast-r-cnn-vs-faster-r-cnn-ml/amp/?usqp=mq331AQFKAGwASA=&amp_js_v=0.1

[12] https://www.fritz.ai/object-detection/

[13] https://www-geeksforgeeks-org.cdn.ampproject.org/v/s/www.geeksforgeeks.org/r-cnn-vs-fast-r-cnn-vs-faster-r-cnn-ml/amp/?usqp=mq331AQFKAGwASA%3D&amp_js_v=0.1#aoh=16012635795556&referrer=https%3A%2F%2Fwww.google.com&amp_tf=From%20%251%24s

[14] https://www.javatpoint.com/tensorflow-object-detection

[15] https://automaticaddison.com/real-time-object-recognition-using-a-webcam-and-deep-learning/

[16] https://www.kdnuggets.com/2018/09/object-detection-image-classification-yolo.html

[17] https://towardsdatascience.com/getting-started-with-coco-dataset-82def99fa0b8

[18] https://medium.com/zylapp/review-of-deep-learning-algorithms-for-object-detection-c1f3d437b852