ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/359393141

# REAL TIME OBJECT DETECTION WITH SPEECH RECOGNITION USING TENSORFLOW LITE

**Article** · February 2022

CITATIONS
2

READS
388

**2 authors:**

Ganesh Khekare
Vellore Institute of Technology
**32** PUBLICATIONS **289** CITATIONS
SEE PROFILE

Kalpeshkumar Solanki
Parul Universiy
**1** PUBLICATION **2** CITATIONS
SEE PROFILE

**Some of the authors of this publication are also working on these related projects:**




Transportation in Smart City View project

All content following this page was uploaded by Ganesh Khekare on 22 March 2022.

The user has requested enhancement of the downloaded file.

# REAL TIME OBJECT DETECTION WITH SPEECH RECOGNITION USING TENSORFLOW LITE

**Dr. Ganesh Khekare[1], Kalpeshkumar Solanki[2]**

*[1,2]Department of Computer Science and Engineering, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India*

**Abstract:** This is an android-based system for visually challenged people. It provides object detection in the near area for visually challenged users. This system helps visually impaired people to identify surrounding things like a chair, table, phone, etc. used in their daily life. This system aims to obtain the image information of the surrounding environment through an RGB color camera and uses the deep learning method to identify the type and location of the object in front of the blind person, which assists the blind person to shop in the supermarket. This system builds an object detection system based on an SSD network as well as an object classification system based on a MobileNets network and builds two sets of image databases containing 9 categories in total for training a neural network. This system is buildup of an Android - TensorFlow interface to deploy an object classification network on the Android phone. A voice announcement function is added to the android terminal to feedback the recognized object to the blind person in real-time. This system proven to be more efficient for visually challenged people.

***Keywords:*** **Blind, Deep Learning, Convolutional Neural Network, Object Detection, Object Classification**

## 1. Introduction

The field of computer vision has always been an active area of research, and in this paper, we have a device integrated with its application (object detection) and running in real-time. Due to the limited processing power of the project equipment, we need a lightweight object detection model for real-time detection. Available light models are TinyYOLO and SSD MobileNet. In this paper, we use SSD MobileNet. Research background and significance.This topic is a subdivision of the big topic of blind visual aids, aiming to use deep learning-based. The related network and algorithm in the field of object detection are used to detect objects in their daily life, and the detection results are Feedback to blind people to help blind people improve their day-to-day life. According to statistics on people with visual impairment and blindness released by the World Health Organization in October 2017 Look, about 253 million people worldwide are currently visually impaired. 36 million of them are blind, 2.17 million people have moderate to severe visual impairment. The number of blind people is huge. Vision, hearing, touch, and smell constitute the most important perception systems of human beings. vision is human experience and the primary source of knowledge. The impact of picture information on human perception is far greater than that of other media information. Therefore, to solve the visual impairment of the visually impaired the issue of access to information became a pressing need. Computer vision is one of the emerging fields of current deep learning research, and object detection in computer vision an important branch of digital word images has become an indispensable information medium, and massive amounts of image data are being generated every moment. with this at the same time, it is becoming more and more important to accurately identify objects in images [1]. The object classification is responsible for judging the input Whether there are objects of the category of interest in the image, output a series of labels with scores indicating the category of interest the likelihood of an object appearing in the input image [2]. The position and range of the object, the bounding box of the output object, the center of the object, or the closed boundary of the object, etc. The shape bounding box is the most used choice [3]. Its purpose is to help blind people obtain the classification information of images in a specific environment and process them. android application cameras can obtain environmental information, and identify the type of objects, such as computers, bags, toothbrushes, and other daily necessities. In this way, blind people recognize objects surrounding them in a way and enhance the adaptability and independence of blind people in complex environments.

### .2 Research status at home and abroad

Object classification, detection, and segmentation are three major tasks in the field of computer vision. This topic involves the classification and detection of the first two types of tasks. The object classification is to return a single image The category with the highest confidence in the film is usually the most obvious object in the corresponding image; object detection is in addition to returning all detected object categories in a single image, it is also necessary to return the position of the object.

### 1.2.1 Object Detection Algorithm

The current mainstream object detection algorithms are based on deep learning models, which can be divided into two categories: The first type is the two-stage detection algorithm, that is, the detection process is divided into two steps, and the first step is to generate a candidate area. The second step is to classify and localize the candidate regions. The main representative algorithm is the R-CNN series network proposed by Ross Girshick et al., including R-CNN, Fast R-CNN, Faster R-CNN, etc. [4, 5]. The network first uses the selective search algorithm to extract Take the areas of the input image that may contain the inspected object, and then compress these areas to a uniform size (227*227) and input the convolutional neural network for feature extraction and input the feature vector into the SVM (Support Vector Machines) classifier to get the category of the candidate region. The algorithm uses convolutional neural networks only in the second step, while the first step uses traditional computer vision algorithms, which directly leads to the complexity of the algorithm. high, the detection speed is slow. Because traditional algorithms are usually calculated within the CPU, this consumes a lot of computational time. The second type is the one-stage detection algorithm, that is, it does not need to perform the candidate region stage and directly returns to the object algorithms of categories and locations this type of algorithm achieves the purpose of improving detection speed by sacrificing detection accuracy, representing the algorithm is the YOLO series network proposed by Joseph Redmon et al. [6] (including YOLO, YOLO9000, YOLOv3, etc.) and the SSD (Single Shot Detector) series of networks proposed by Wei Liu et al. Including SSD, DSSD, etc.) [7]. The core idea is to return to the object directly from the picture. The position and type of the object are abandoned, the traditional machine vision algorithm is abandoned, and the end-to-end model detection is realized(end-to-end) feature. YOLO The biggest contribution of the algorithm is to introduce the object detection task into the real-time detection stage, which greatly expands deep learning the scope of application in the field of object detection. The original intention of its design is to maintain the conditions of real-time detection. to further improve the detection accuracy. Considering that YOLO simply meshes and processes the image It will lead to problems such as inaccurate positioning, and the method based on candidate regions has the advantage of more accurate positioning, so SSD combines YOLO and anchor box methods for detection, and the result is more accurate than yolo Much improved, it can reach 72% mAP, and the detection speed is 58fps.

### 1.2.2 Object Classification Algorithm

The earliest entry of deep learning into the field of computer vision is to complete the task of object classification, and the development of such networks Gradient-Based Learning applied to Yann LeCun et al. Document Recognition" [8], this article proposes the world's first real convolutional deity after the network LeNet. This network sparrow is small but complete, including the basic components of modern CNN networks such as convolutional layers, pooling layers, and fully connected layers are introduced. In 2012, Alex Krizhevsky et al. proposed the AlexNet network [9]. The network's top-5 accuracy in the ImageNet image classification competition was higher than last 10 years. The product neural network has officially become the core algorithm model in the object classification problem. The reason for the success of this algorithm is: 1) using the nonlinear activation function ReLU; 2) using the method of preventing overfitting Dropout, Data Argument; 3) Using a million-level database ImageNet for training; 4) GPU implementation. In 2014, the famous research group VGG (Visual Geometry Group) of Oxford University proposed the VGG-Nets [10], the network is in the 2014 ImageNet competition localization task (Localization task) base network used by the first and second place. Due to the good generalization performance of VGG-Nets, it is pre-trained models on the ImageNet dataset are widely used in addition to the most commonly used there are many problems other

than feature selection: such as generating object candidate frames, small object positioning and retrieval, and image co-determination. bit etc. VGG-Nets are optimized from AlexNet and mainly modified the following two aspects: 1) In the first convolutional layer uses a smaller size filter; 2) supports multi-scale training and testing pictures. The basic network in the SSD network used in this project is also VGG-Nets. In 2015, Kaiming He et al. proposed Deep Residual Networks (ResNet) [11]. The mobile net architecture is as shown in figure 1.1.

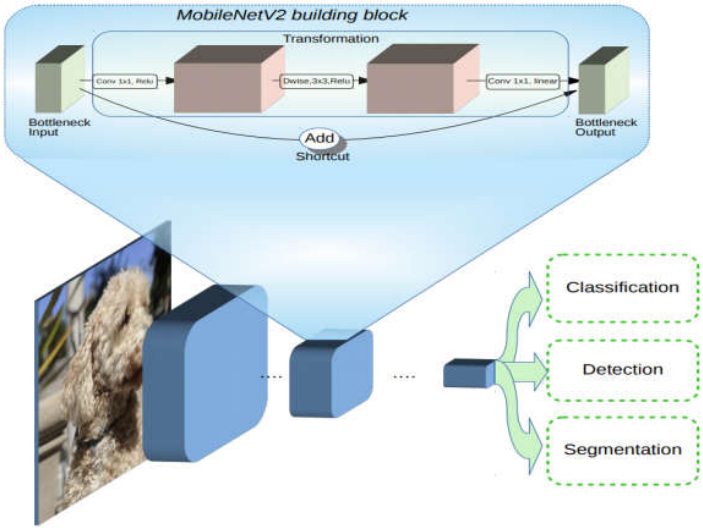



Figure 1.1. Mobile net Architecture

It is a good solution to solve the problems of gradient explosion or gradient disappearance in the training process of traditional network depth, its network performance far outperforms the traditional network model, which makes it a good candidate in the ILSVRC 2015 and COCO 2015 competitions. It won first place in detection, localization, and segmentation tasks. Compared with traditional VGG-Nets, residual networks It is deeper VGG-Nets. In 2017, Andrew G. Howard et al proposed MobileNets [12], the network is mainly for mobile devices and embedded device design, greatly improving the speed of the network [13], making it more suitable for mobile devices and embedded devices [14].

### 1.3 Main design tasks and expected goals

The main work of this subject is to build a system based on the deep learning network framework to assist blind people in the surrounding area. The MobileNets network and the classification algorithm respectively designs two systems for object recognition. At the same time, construct Build a dataset of daily necessities things and use an object detection system based on SSD network to carry the category and location of items are returned in real-time on the GPU-based system and using MobileNets-based object classification. The classification system returns the object category in real-time on the android phone and makes voice feedback.

## 2. System Design Scheme

### 2.1 Overall design scheme

This project uses two kinds of neural networks to build two sets of the system implementing object detection and classification respectively. In the object detection module, the RGB camera collects color images to obtain a three-dimensional image matrix. Then input the obtained image into the SSD network, and the parameters and hyperparameters of the network have been trained and achieved over 90% accuracy in test evaluation. The design scheme is as shown in figure 2.1.

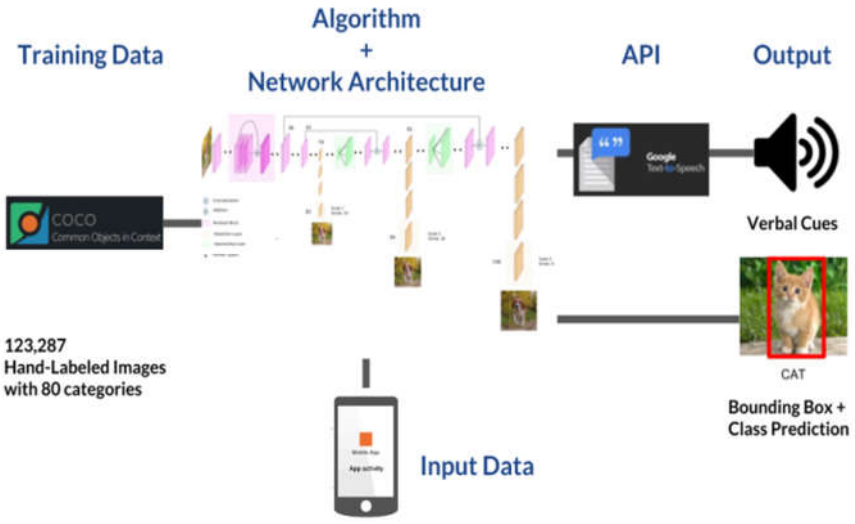



Figure 2.1. Design Scheme

The result is a border (bounding box) and the label of the object, the result will be visually displayed on the input image, and the network is ready to accept the next picture. After the camera captures the picture, the picture is directly sent to the network through the Android-TensorFlow interface. The android app will display the detected category and the confidence of the category in real-time according to the returned result. If the confidence level is greater than 0.8, the corresponding category will be announced by voice. The average forward propagation time of each network is 200ms, so the system can realize real-time detection on the android side.

## 2.2 TensorFlow Deep Learning Framework

TensorFlow is a software library developed by Google for use in machine learning .As the popularity of deep learning research continues to increase, various open-source deep learning frameworks emerge in an endless stream, among which include TensorFlow, Caffe, Keras, CNTK, Torch7, MXNet, etc. Figure 2.2 shows the TensorFlow architecture.

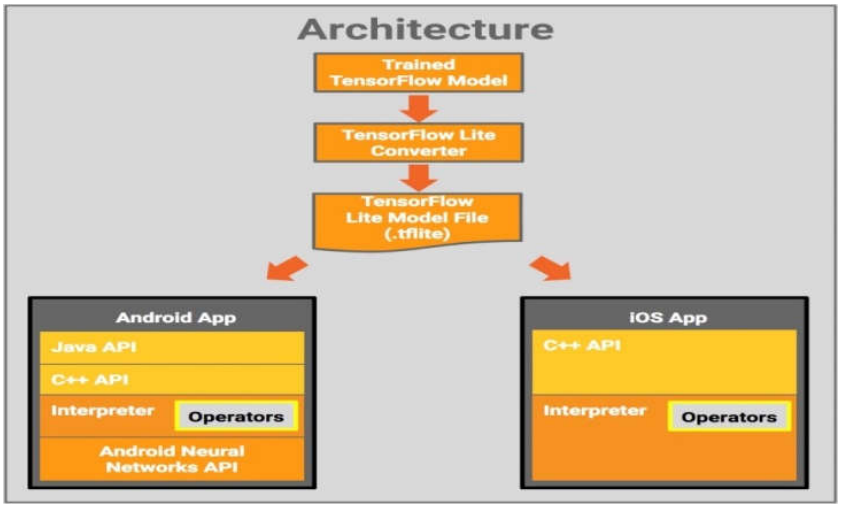



Figure 2.2. TensorFlow Architecture

There are also two installed versions of TensorFlow, one is CPU-only TensorFlow, the other is TensorFlow with GPU support. TensorFlow is currently available on Windows The client only supports 3.5.x and 3.6.x. Based on the above analysis, the CPU version of TensorFlow was installed on the notebook using the pip command. On the server, in order not to interfere with other python-based

installation packages in the system, an independent virtual machine is created. environment and install the GPU version of TensorFlow.

### 2.3 Object Detection Network Design

The SSD object detection network used in this topic is divided into two parts, the former part is based on VGG-Nets. The base net of the latter part is a multi-scale single-shot detector (Single Shot MultiBox Detector, referred to as SSD). VGG-Nets are a basic classification network with regular network characteristics and a simple structure.

The SSD detector is based on the VGG basic network. The whole system takes the VGG16 network the first five layers are used as the basic classification network, and the two fully connected layers fc6 and fc7 are converted by the astrous algorithm It is a convolutional layer, and finally 3 additional convolutional layers with different channel numbers with a receptive field of 1*1 and a global average pooling layer. The feature maps of each layer starting from the sixth convolutional layer are used for the prediction of the size and position of the default box and the prediction of different categories of objects, the final network passes Non-Maximum Suppression yields the result. The core of the SSD algorithm is that the feature map of each layer after the sixth layer of the network will pass through a detector. For the feature maps of each layer, according to different sizes (scale) and aspect ratios (ratio), the algorithm automatically generates k default boxes.

When the intersection ratio of the default box and the real box is greater than the threshold (set to 0.5), the two match, and the setting is not the default boxes for matching are negative samples, and the default boxes for matching are positive samples.

### 2.4 Object Detection Network Deployment

This topic adopts the SSD object detection network based on the TensorFlow deep learning framework.

### 2.5 Design of object classification system

Compared with traditional classification networks, the main contribution of MobileNets is the use of deeply separable volumes products to build lightweight deep neural networks, which enables the deployment of neural networks in mobile and embedded systems to become possible.

### 2.6 Object Classification System Deployment

The object classification system is deployed using TensorFlow slim, which is a high-level library based on TensorFlow like Keras, TensorLayer, tfLearn, etc. Made easy via the web. Its two major advantages are: First, it can eliminate many heavy burdens in native TensorFlow. The complex template code makes the code more compact and more readable; the second is to provide a lot of computer vision twenty-two well-known models in vision (VGG, AlexNet, etc.) and can be extended in various ways. The first two steps of the object classification system deployment are like the first two steps of the object detection system deployment, including the dataset conversion and training. After training, the obtained network parameters and network framework need to be encapsulated into .pb format(tensorflow model file format).

## 3. System testing, verification, and result analysis

### 3.1 Analysis of the results of the object detection system

In this subject, a class of data sets is made for the object detection system, the class is clock, the number of pictures in the data set is A total of 2635 pictures, including 487 pictures of the imagenet dataset. All images were manually marked with labeling software. There are several marked images of different categories, the labeling marking interface, and the corresponding label. A series of hyperparameters need to be set before training. Figure 3.1 shows the analysis of the results of the object detection system.

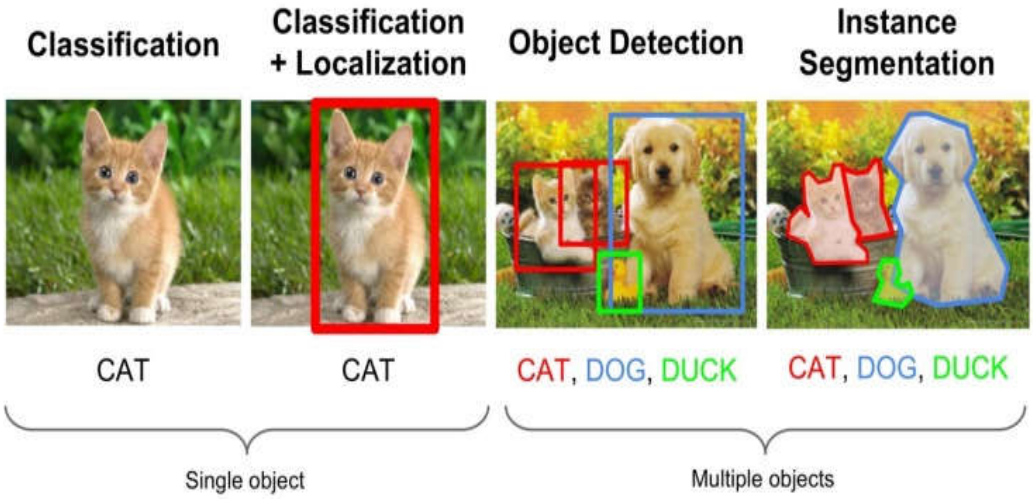
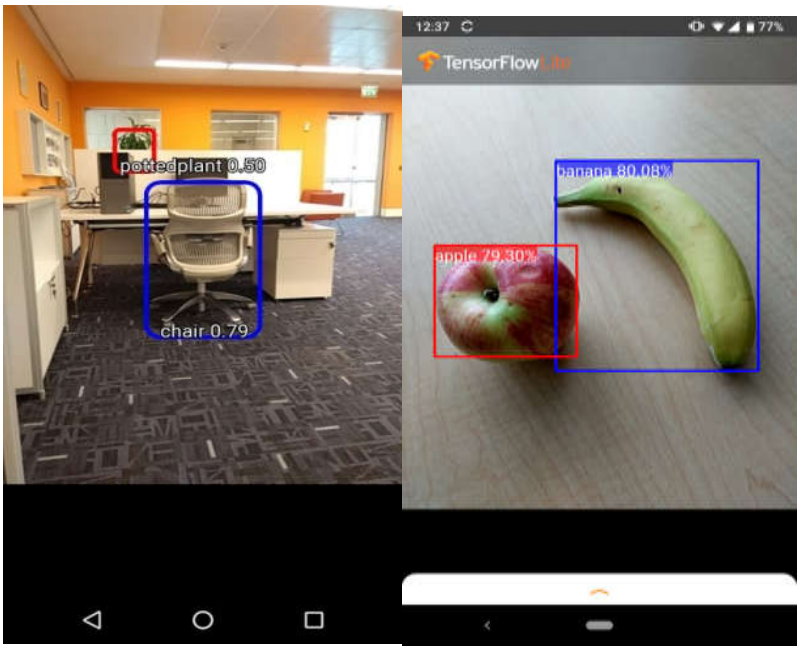



Figure 3.1. Analysis of the results of the object detection system




(a) (b)

Figure 3.2. Results of Object Detection System

Figure 3.2 shows the results of the object detection system. The training is deployed on the server's the increase of training times during the training process: After training, test a single image. In terms of single object detection, both can achieve higher detection the confidence level of the three images is above 0.99. After completing the single image evaluation, use the network for real-time detection. Since real-time detection and video detection have the same method and principle in-network calls, for convenience, this project only uses the network to test the input video stream without calling the computer camera in the program. Video Both stream reading and output utilize OpenCV.

### 3.2 Analysis of the results of the object classification system

The object classification system data set of this subject includes eight categories, namely books, bags, clocks, computers, pens, drink bottles, spoons, and toothbrushes. Each type of training data set is 1200 images, and the evaluation data set is 200 pictures. The types of hyperparameters used in training are the same as those used in the object detection system. In the real-time detection of this subject, the detection result of the detected object will be within a few seconds of the first detection. Drift, that is, the detection result when the camera is just aimed at the object to be detected is unreliable. When the

software is used for blind assistance, the confidence level displayed on the screen is useless. The real information people get is from the voice feedback. Therefore, the primary task of this system is to increase voice feedback accuracy. The method adopted by the author is that the feedback task is triggered once every 1000 times of detection, that is, for each detection, The result is sampled, and the sampling frequency is 1/1000 of the picture (the reason why the second is not used here is that the value of each picture is the detection time is different and cannot be measured by time). Then, voice feedback is performed on the sampled graph, if Only feedback if the confidence is greater than 0.8, otherwise enter the next cycle. This setting greatly increases the accuracy of voice feedback and makes the whole system more robust.

## 4. Summary

### 4.1 Features and summary of this design

This project mainly develops an image recognition system for the blind, this topic draws from the existing deep learning in the field of computer vision Starting from the application of the domain, this paper summarizes the mainstream image detection since deep learning entered the field of computer vision. and image classification algorithms, and according to the actual needs of blind people, two sets of helping them to object surrounding and Image recognition system. The specific results obtained in this study are as follows: 1. This subject has produced two sets of data sets, one for image detection, including 1 type of object, 2,635 color images, 2,635 .xml default box files; another set for image classification contains 8 Class objects, 11,200 color pictures; 2. This topic trains two convolutional neural networks based on the TensorFlow deep learning framework, respectively. is an SSD object detection network with 10 convolutional layers and 1 global pooling layer and an SSD object detection network with 27 MobileNets object classification net with 2 convolutional layers, 1 average pooling layer, and 1 fully connected layer network; 3. This topic deploys the framework on both the PC and Android sides, and the PC side is used to import the SSD object The detection network is used for real-time detection of video, and the Android side is used to import MobileNets objects The classification network is used for real-time detection of surrounding objects through the mobile phone camera; 4. In this project, a simple audio library containing 8 types of object voices is made and deployed On the Android side, real-time feedback of test results is realized.

### 4.2 Problems and thinking in design

During the completion of this project, the following problems were mainly solved:

1. In the initial training, the loss has not been able to converge. The solution is to first increase the learning rate and the number of learned layers is reduced, allowing the network to perform gradient descent for the first 1000 steps. When the loss oscillation amplitude changes after it is too large, stop training, increase the number of layers that can be learned, reduce the learning rate, and re-read to stop training at the checkpoint, continue training. After repeating this several times, the loss will converge to a value that can be the accepted value, at which point training stops.
2. In the early stage of training, it has always been reported that the dimensions do not match. The dimension mismatch is mainly due to the number of input categories and the network. The original number of classes in the network does not match, even if num_class in the program is changed to the class to be trained. Don't still report errors. After that, the solution is to change the original network structure parameters during fine-tuning the last layer excludes so that the number of network classes matches the number of input classes.
3. The network's image detection accuracy for multiple objects is not high. Found after trying various data augmentation methods It does not improve the accuracy of network detection, so this can only be removed by increasing the multi-object dataset. problem.

## 5. Conclusion

The real-time object detection system using Tensorflow Lite is discussed in a research article. The purpose of this system is to aid the blind people surrounding them, so there are still many areas for improvement. First, it can increase the number of categories for training, and collect more data in to improve the system's ability to adaptability and robustness of object detection; Second, the system of this subject is not sensitive to small objects, which requires blind people can identify goods only when they are closer to the camera. The solution can be to realize the camera can be used in the process of building the system.

## REFERENCES


[1] *Szegedy C, Toshev A, Erhan D., "Deep Neural Networks for object detection", Advances in Neural Information Processing Systems, **(2013)**, 26:2553-2561.*

[2] *Felzenszwalb P F, Girshick R B, Mcallester D, et al., "Object Detection with Discriminatively Trained Part-Based Models", IEEE Transactions on Pattern Analysis & Machine Intelligence, **(2014)**, 47(2):6-7.*

[3] *Li Xudong, Ye Mao, Li Tao., "A review of object detection research based on a convolutional neural network", Computer Application Research, **(2017)**, 34(10):2881-2886.*

[4] *Khekare G., Wankhade K., Dhanre U., Vidhale B. **(2022)** Internet of Things Based Best Fruit Segregation and Taxonomy System for Smart Agriculture. In: Verma J.K., Saxena D., González-Prida V. (eds) IoT and Cloud Computing for Societal Good. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-73885-3_4*

[5] *Ren S, He K, Girshick R, et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Trans Pattern Anal Mach Intell, **(2015)**, 39(6):1137-1149.*

[6] *Redmon J, Divvala S, Girshick R, et al., "You Only Look Once: Unified, Real-Time Object Detection", **(2015)**:779-788.*

[7] *Liu W, Anguelov D, Erhan D, et al., "SSD: Single Shot MultiBox Detector", 2015:21-37*

[8] *Lecun Y, Bottou L, Bengio Y, et al., "Gradient-based learning applied to document recognition", Proceedings of the IEEE, **(1998)**, 86(11):2278-2324.*

[9] *Khekare, Ganesh and Shahrukh Sheikh. "Autonomous Navigation Using Deep Reinforcement Learning in ROS." IJAIML vol.11, no.2 **(2021)**: pp.63-70. http://doi.org/10.4018/IJAIML.20210701.oa4*

[10] *Simonyan K, Zisserman A., "Very Deep Convolutional Networks for Large-Scale Image Recognition", Computer Science, **(2014)**.*

[11] *He K, Zhang X, Ren S, et al., "Deep Residual Learning for Image Recognition", **(2015)**:770- 778.*

[12] *Howard A G, Zhu M, Chen B, et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications". **(2017)**.*

[13] *Temurnikar, A., Verma, P., & Dhiman, G. **(2022)**. A PSO Enable Multi-Hop Clustering Algorithm for VANET. International Journal of Swarm Intelligence Research (IJSIR), 13(2), 1-14. http://doi.org/10.4018/IJSIR.20220401.oa7*

[14] *G. K. Yenurkar, R. K. Nasare and S. S. Chavhan, "RFID based transaction and searching of library books,"* 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), ***(2017)**, pp. 1870-1874, doi: 10.1109/ICPCSI.2017.8392040.*




View publication stats