A Short Survey on Real-Time Object Detection and its Challenges.

Naba Krushna Sabat*, Umesh Chandra Pati*, Santos Kumar Das*

Dept. of Electronics & Communication Engineering,
National Institute of Technology, Rourkela, India
Nabakrushna4u@gmail.com, ucpati@nitrkl.ac.in, dassk@nitrkl.ac.in.

Abstract: Object detection is an attractive research interest in recent years. It plays an significant role to understand the image in video analysis. Object detection technique has improved in computer vision when deep learning method came to the picture. From the last two decades, several algorithms have been developed and improved to detect the object in different conditions and still, it has number of challenges. The objective of this paper is to provide a short survey on object detection and its challenges not only in computer vision but also in the wireless sensor network.

Keywords: Deep learning, Object detection, Convolution Neuron Network, Sensor, IoT.

1.1. Introduction

The term detection and tracking are quite different from each other. The idiom detection describes the identification (the type of object say human, chair, bottle, toll gate [1,2], etc.) as well as the exact location of the object [3]. On the other hand, tracking means getting information about the object. It consists of different sub-tasks. For example, in human tacking, the sub-tasks include pedestrian detection [4], skeleton detection [5], and face detection [6] etc. Hence, the object tracking is an important key component in today's scenario especially in security and surveillance system. For example, suppose an unauthorized person has entered into the restricted area, then it requires to detect and track the person (say intruder). The direction of movement, position and behaviour information of intruder can be collected from the deployed sensors. There are several techniques evolved for object detection and tracking still, different challenges are present like low-intensity light, occlusion area, crowded area, long-distance tracking, Field of View (FoV) of sensors [7-9] etc.

Different sensors like Passive Infrared (PIR), Light Detection And Ranging (LiDAR) sensor etc. can also be deployed for detection and tracking of objects.

Presently, the most popular deep learning technique is being used for detection and tracking in computer vision. It has been found the advanced deep learning algorithm is giving higher accuracy and greater performance in terms of efficiency and execution time as compared to the previous algorithm with few constraints. In deep learning method, the complete human detection model is segregated to different stages such as informative region selection, feature extraction, and classification [10]. Once the model classifies the type of object, then tracking of the object can be possible. In the case of informative region selection, the image is scanned by the neural network which further extracts the features and classifies the objects. For detection of an object, it needs to extract visual features. The feature extraction techniques such as Scale Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), Harr-like, Histogram of Oriented Gradients (HOG), etc., are used to extract the feature from an image or video.

The organisation of this paper is as follows: Section 1.2 interpret the basics of deep learning using neural network and its architecture. An overview of different types of models used for detection and tracking of objects is described in Section 1.3. The application and uses of sensors for object detection are mentioned in Section 1.4. The research challenges are explained in Section 1.5 and Section 1.6 concludes the review.

1.2 Deep Learning and Its Architecture

Deep learning (DL) is a subpart of machine learning (ML), which is a part of Artificial Intelligence (AI). It is similar to machine learning, but the main difference between ML and DL is the learning process. DL has more learning capability as compared to ML. The basic blocks to design a DL architecture is Neuron Network (NN), where more than one layers are present. Each layer has number of neurons, all neurons of the previous layer are connected to each neuron of next layer. They are computed with some activation function like sigmoid, Rectified Linear Unit (ReLU) to produce the desired result [11]. The neurons are updated continuously with their weights and bias value using feed forward and backward propagation method with a learning rate termed as ‘α’(Alpha). It helps for smoothly and slowly updates the weights having a small value between 0 to 1.

A network has more than two networks it is treated as a deep neural network (DNN), and if the network has more than ten network, then it is called very deep neural network. A fully connected (FC) DNN architecture is depicted in Fig.1.1.

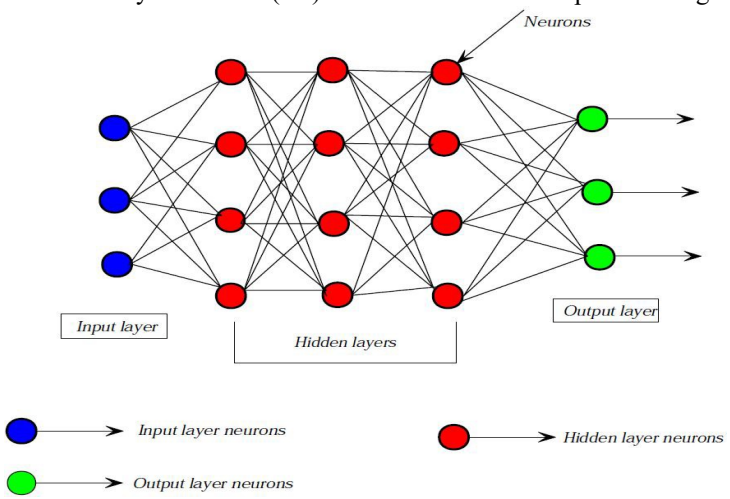

Fig.1.1. A fully connected DNN Architecture.

The DNN gives better performance when a convolution neural network (CNN) is used. The input frame processed with number of cascading convolution and pooling layer in CNN. It extracts the features in each layer, then the last one or two layers are fully connected where the output is taken. Fig.1.2 shows that the LeNet architecture has two layers of CNN for object detection.

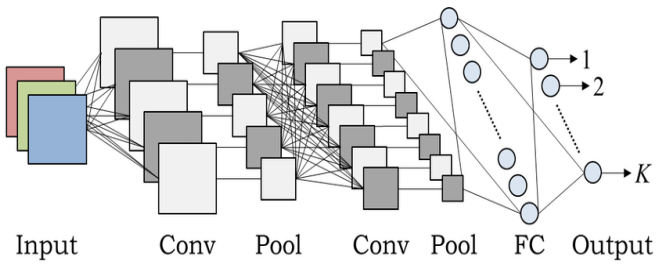

Fig.1.2. LeNet Architecture

1.3 Types of Model Used for Object Detection and its Comparison

Earlier in 2001, first object detection model Viola-Jones was designed and implemented. This model is treated as an object detector, but for the most case, it was used in facial detection. They used haar-like feature extraction method for detection [12]. The first deep learning object detection model OverFeat network [13] was designed which uses CNN and sliding window approach to detect objects from the image.

In present scenario, the generic object detection models have been classified based on region proposal and regression/ classification proposal. The classification of object detection model is illustrated in Fig.1.3.

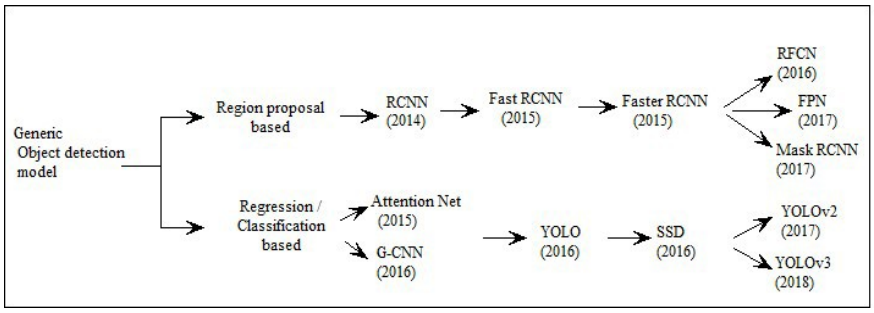

Fig.1.3. Classification of Generic Object detection model based on region proposal and classification methods.

In 2014, R. Girshick et al. [14] introduced Region Convolutional Network (RCNN), where only 2000 regions are selected from the complete input image using selective search algorithm. These regions are warped and fed to CNN, to extract the feature. Then applied to Support Vector Machine (SVM), which helps to classify the object. This model has some issues like it takes more time to training and also because of 2000 selective region, it requires long time to extract the features. Hence, it is not preferable to deploy for real-time application.

In 2015, the same author improved their network called Fast RCNN [15]. Here, the input image is given directly to the convolution layer instead of giving region proposal. The image is processed and generates feature map. Then, by using Region of Interest (RoI) and polling layer fixed feature vectors are extracted from each region which is then fed to the FC layer and at the output the classifier helps to classify the object with the bounding box. The drawback of RCNN overcomes training speed and accuracy improved using Fast RCNN.

Both RCNN and fast RCNN uses selective search algorithm to find the region proposal, which is a sluggish process and also affects the network performance. S. Ren et al. proposed a model called Faster RCNN, which has similar work to Fast RCNN, but instead of selective search, it uses Region Proposal Network (RPN), which has the capability of predicting region using the concept of Anchor [16]. Later regional fully convolution network (RFCN) introduced in 2016, and in 2017 feature pyramid network (FPN) and mask RCNN is developed. These all are an extension of faster RCNN with higher accuracy and network performance.

You Only Look Once (YOLO) and Single Shot multi-box Detection (SSD) are the two most popular object detection models in regression/classification based object detection. Hence, overviews of these two models are only described. J. Redmon et al. in 2016, introduces YOLO model. Here the input frame segregates to S X S grid; the grid cell has responsible to predicts the confidence as well as bounding box about the object. The confidence shows how accurate the object is; in other words, it reflects the accuracy and the bounding box states the object is present or not [17]. It is designed with 24 convolution layer and 2 fully connected layer. It gives a fast response and in real-time processing, it takes 45 Frames per Second (FPS), whereas Fast YOLO process 155 FPS and better response than other models [10]. YOLO has advantages of speed detection. But, it faces problem to detect very small object.

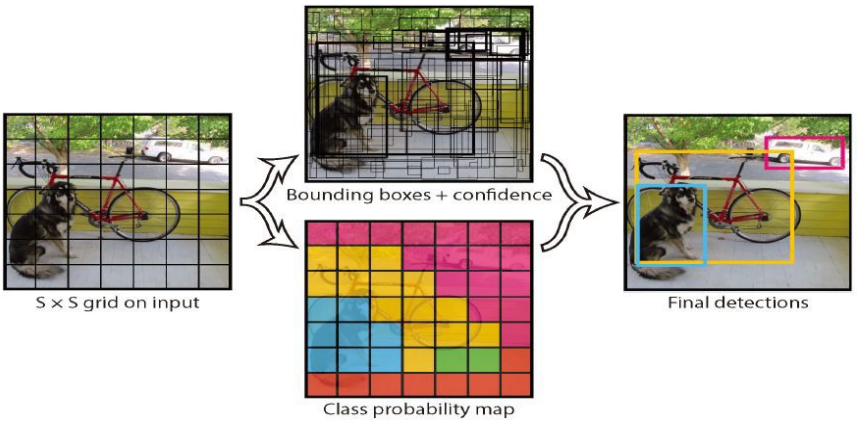

Fig.1.4. Basic YOLO concept [15].

In 2016 another model has been proposed by Liu et al. which is called Single Shot multi-box Detector (SSD) [18]. It uses anchor in the frame to define the number of regions. Generally, anchor predicts the bounding box co-ordinate and class score. The convolution network VGG16 is the backbone of the SSD model used to detect the object. It uses multiple feature layers in the network, which helps to detect exact bounding box for the object irrespective of their aspect ratio [10]. This model gives better accuracy as compared to YOLO. Later in [17] YOLOv2 and in [21] YOLOv3 was introduced improving the speed of detection and accuracy.

Basically, four popular data sets used for object detection are MS COCO, ImageNet, PASCAL VOC, and Open images. Each model performance can be judged over a dataset and calculating its mean Average Precision (mAP). Table 1.1. express the performance of different models with the COCO dataset.

Object Detector Type	Backbone	AP	AP_{50}	AP_{75}
Faster RCNN+++ [7]	ResNet-101-C4	34.9	55.7	37.4
Faster RCNN w FPN [7]	ResNet-101-FPN	36.2	59.1	39.0
Faster RCNN by G-RMI [7]	Inception ResNet-v2 [13]	34.7	55.5	36.7
Faster RCNN w TDM [7]	Inception ResNet-v2-TDM	36.8	57.7	39.2
YOLOv2 [17]	DarkNet-19 [17]	21.6	44.0	19.2
SSD513 [18, 19]	ResNet-101-SSD	31.2	50.4	33.3
DSSD513 [18]	ResNet-101-DSSD	33.2	53.3	35.2
RetinaNet [20]	ResNet-101-FPN	39.1	59.1	42.3
RetinaNet [20]	ResNeXt-101-FPN	40.8	61.1	44.1
YOLOv3 608 X 608	DarkNet53	33.0	57.9	34.4

Table 1.1. Performance comparison of different object detector model.

1.4 Deployment of Physical Sensor for Object Detection.

Sensors like PIR, LiDAR, and RADAR (Radio Detection and Ranging) are used for object detection and tracking. This paper discussed some moving object detection and tracking based on physical sensors deployment.

Unauthorized human detection and intimation are done using PIR sensor and GSM module in [22]. They used IoT concept for detection and intimation about the presence of an intruder. In [23], Luo. et al. proposed a technique of indoor human object localization, by mounting PIR sensor nodes on the ceiling of a room. J Yun et al. [24] designed and implemented a hardware using PIR sensor which detect the human (consider as object) movement direction. The movement data are collected from the array of PIR sensors and are applied to machine learning algorithm. The model gives an accuracy of result 89% - 95%. Here, the measure difficulty is to arrange the PIR sensor array.

Wu Jianqing et al. designed a model to detect roadside vehicle from LiDAR sensor data. They used background subtraction (3D-DSF) method to eliminate static objects and identify the lane. Then applied the clustering method to identify the object [25]. In [26] J. Chen et al. used a 3D LiDAR for detecting deer crossing on the highway. When a deer or group of deer are detected on the specific area, it activates a warning signal so that the driver will be alert before the accident.

1.5 Research Challenges

The sensors LiDAR or PIR can detect the object individually but not able to recognize the type of object. These sensor fails to detect object in bad weather conditions (like foggy and snow), group of objects, and occluded objects. To avoid such problems, multiple sensors with cameras can be deployed. The 3D sensor can be deployed to detect an object from which more depth information can be collected and processed for detection.

Reducing the training time period of deep learning architecture is also a challenging task because each model takes a long time to train. Using cyclic learning rate and super convergence technique, the training time can be reduced. It is difficult to find the object if the image is blurred or the video is defocused. The accuracy is also degraded in the network because it is not trained end to end. For this problem LSTM, optical flow, spatiotemporal tabulate can be used in consecutive frames. The other challenges include multimodal information fusion, network optimization, cascade network, unsupervised and weakly supervised learning etc.

1.6 Conclusion

This paper provides the short survey on different most useful deep learning object detection method. Every year, the performance of the network is improved and overcomes the demerit of older networks like RCNN to Fast RCNN, then Faster RCNN, and so on. It has been seen that each model has their own priority i.e. for more accurate detection SSD model gives better result but, for speed detection approach YOLO is preferred. It also provides some research challenges which will guide to improve the network model as well as understand the object landscape.

Acknowledgment

This research is supported by the Defence Research Development Organisation (DRDO), India, Sanction no. ERIP/ER/1506047/M/01/1710.

References:

[1] Sabat N. K., Pati U. C., Senapati B. R., and Das S. K.: An IoT Concept for Region Based Human Detection Using PIR Sensors and FRED Cloud., 2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP), Chennai, India, pp. 1-4, (2019).

[2] Senapati B. R., Khilar P. M., and Sabat N. K.: An Automated Toll Gate System Using VANET., 2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP), Chennai, India, pp. 1-5, (2019).

[3] Felzenszwalb P. F., Girshick R. B., McAllester D., and Ramanan D.: Object detection with discriminatively trained part-based models., IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp.1627–1645, (2009).

[4] Dollar P., Wojek C., Schiele B., and Perona P.: Pedestrian detection: An evaluation of the state of the art., IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 4, pp. 743–761, (2011).

[5] Kobatake H., and Yoshinaga Y.: Detection of spicules on mammogram based on skeleton analysis., IEEE Transactions on medical imaging, vol. 15, no. 3, pp. 235–245, (1996).

[6] Sung KK, and Poggio T.: Example-based learning for view-based human face detection., IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 1, pp. 39-51, (1998).

[7] Girshick R., Donahue J., Darrell T., and Malik J.: Rich feature hierarchies for accurate object detection and semantic segmentation., in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587, (2014).

[8] Redmon J., Divvala S., Girshick R., and Farhadi A.: You only look once: Unified, real-time object detection., in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788, (2016).

[9] Ren S., He K., Girshick R., and Sun J.: Faster r-cnn: Towards real-time object detection with region proposal networks., in Advances in neural information processing systems, pp. 91–99, (2015).

[10] Zhao ZQ., Zheng P., Xu S., and Wu X.: Object detection with deep learning: A review., IEEE transactions on neural networks and learning systems, vol. 30, no. 11, pp. 3212-3232, (2019).

[11] Schmidhuber J.: Deep learning in neural networks: An overview., Neural networks, vol. 61, pp. 85–117, (2015).

[12] Viola P., Jones M.: Rapid object detection using a boosted cascade of simple features., CVPR, pp 511-518, (2001).

[13] Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., and LeCun Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks., arXiv preprint arXiv:1312.6229 (2013).

[14] Girshick R., Donahue J., Darrell T., and Malik J.: Rich feature hierarchies for accurate object detection and semantic segmentation., In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587. (2014).

[15] Girshick R. : Fast r-cnn., In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448, (2015).

[16] Ren S., He K., Girshick R., and Sun J.: Faster R-CNN: Towards real-time object detection with region proposal networks., In Advances in neural information processing systems, pp. 91-99, (2015).

[17] Redmon J., Divvala S., Girshick R., and Farhadi A.: You only look once: Unified, real-time object detection., In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779-788, (2016).

[18] Liu W. et al.: Ssd: Single shot multibox detector., In European conference on computer vision, pp. 21-37. Springer, Cham, (2016).

[19] Redmon J. and Farhadi A.: YOLO9000: Better, faster, stronger., In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263-7271, (2017).

[20] Fu CY. et al.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659, (2017).

[21] Redmon J., and Farhadi A.: Yolov3: An incremental improvement., arXiv preprint arXiv:1804.02767 (2018).

[22] Sahoo KC. and Pati UC.: IoT based intrusion detection system using pir sensor., in 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, pp. 1641–1645, (2017).

[23] Luo X. et al.: Human indoor localization based on ceiling mounted PIR sensor nodes., in 2016 13th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, pp. 868–874, (2016).

[24] Yun J. and Song MH.: Detecting direction of movement using pyroelectric infrared sensors., IEEE Sensors Journal, vol. 14, no. 5, pp. 1482-1489, (2014).

[25] Jianqing Wu. et al.: Automatic Vehicle Classification using Roadside LiDAR data., Transportation Research Record, vol. 2673, no. 6, pp. 153-164, (2019).

[26] Chen J. et al.: Deer crossing road detection with roadside lidar sensor., IEEE Access, vol. 7, pp. 65944-65954, (2019).