Object Detection in Video Frames using Deep Learning

Krishna Kumar
Department of CSE,
BIET, Lucknow, UP

Krishan Kumar
Department of Computer Science,
Gurukula Kangri (Deemed to be University)

C.L.P. Gupta, PhD
Department of CSE, BIT,
Lucknow, UP

ABSTRACT

The object detection based on deep learning is an important application like scene understanding, video surveillance, robotics, self-driving systems etc. in deep learning method which is eminent by its strong capability of feature learning and feature representation compared with the traditional object detection methods. With the rapid development in deep learning, more powerful tools, able to learn semantic, high-level, deeper features, are introduced to address the problems existing in traditional architectures. These models perform differently in network architecture, training strategy and optimization function. This paper introduces the classical methods for object detection and illustrates the relation and difference between the classical methods and the deep learning methods for object detection. Moreover, it introduces the appearance of the object detection based on deep learning elaborates the most typical methods nowadays via deep learning. The paper focuses on the framework design and the working principle of the models and examines the model performance in the real-time environment and hence for the accuracy of object detection. Furthermore, a survey of several specific tasks including salient object detection features, face detection and pedestrian detection has also been briefly discussed. Finally, the main challenges in object detection using deep learning and some solutions for reference has been discussed.

General Terms

Object detection, deep learning, framework design, model performance.

Keywords

Object detection, deep learning, framework design, model performance.

1. INTRODUCTION

The purpose of object detection is to identify and locate one or more effective targets from still image or video data. It widely includes a variety of important techniques, such as image processing, pattern recognition, artificial intelligence and machine learning. It has broad application prospects in such areas like road traffic accident prevention [1], warnings of dangerous goods in factories, military restricted area monitoring and advanced human–computer interaction [2], [3]. While the application scenarios of multi-target detection in the real world are usually complex and variable, balancing the relationship between accuracy and computing costs is a difficult task.

Object detection from video frames has already been the notable research direction and focus in computer vision which can be applied in automatic car, robotics, video surveillance and pedestrian detection. The experience of deep learning technology has changed the traditional ways of object recognition and object detection. The deep neural network has the vital feature representation capacity in video frames processing and is usually used as the feature extraction module in object detection. The deep learning models don't require special hand engineering features and can be designed as the classifiers. Therefore, the deep learning is of significant prospect in object detection as well. The problem statement of object detection is to determine where objects are normally located in a given video frame (object localization in image) and hence detecting it. Therefore, the traditional object detection models mainly divided into three stages: 1) informative region selection; 2) feature extraction; and 3) detection.

1.1 Informative Region selection

Varieties of objects appears in any positions of the video frame and have different aspect ratios or sizes. It is a compulsory task to scan the whole frame with a multi-scale sliding window. Although this exhaustive strategy can find out all possible positions of the objects, its shortcomings are also obvious. Due to a huge number of candidate windows, it is computationally expensive and produces several redundant windows. However, if only a possibility of unsatisfactory regions in the given present video frame (image).

1.2 Feature Extraction

To classify different objects, we need to extract visual features which could provide a semantic and robust representation. This is due to the features which can produce representations associated with complex cells in human brain. However, due to the diversity of appearances, illumination conditions and backgrounds, it is difficult to design a robust feature descriptor for describing all kinds of objects manually.

1.3 Detection

A detector isolates a target object from all other categories and to make the presentations in more hierarchical way and informative for visual recognition. Generally, the Support Vector Machine (SVM) and Deformable Part-based Model (DPM) are good choices for it. Among these classifiers, the DPM is a flexible model by combining object parts with deformation cost to handle severe deformations. In DPM, with the support of a graphical model, carefully designed low-level features and kinematically inspired part decompositions are combined. And discriminative learning of graphical models permits for building high-precision part- based models for a variety of object classes.

2. OBJECT DETECTION

Object detection and tracking in wide research areas in computer vision and other applications in traffic detection, traffic detection, vehicle navigation, interpersonal connections. Object detection is a computer equipment which is related to computer visualization and image processing that deals with detecting examples of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. The wide area of applications in object detection field is face detection, face recognition and video object detection. Some of the applications are – tracking motion of the ball, tracking ball during the match, tracking person in a video. Normally, object detection has uses in many areas of computer vision which includes image fetching and video surveillance [4]. The object detection system recognizes the presence or the absence of objects in certain scenes and cameras' viewpoints. Various domains of the object detection based on the different objectives are classified on specific and conceptual categories. The object detection based on the various models can be either explicit or implicit. The components may vary as per the different approaches. The selection of the object is based on the hypothesis and matching. Moreover, it is an appropriate technique for the processing and searching of the objects where the images are found in real world applications.

2.1 Features of the Object Detection

In the object detection, tracking and the selection of the various characteristics features that can reduce the work accessibility of the computer. When the tracking is done using various algorithms the combination of the different features determined in various steps [5] given below.

a) ***Color:*** This is the feature of the computer system that is used for the histogram appearance representations. The widest features of the color representations are the features of the color representations for the tracking. The features of the color are tracking of serious problem which recognize the illumination variation.

b) ***Histogram of gradients:*** It is the most popular feature used for the detection of the human body. The operations of the histogram feature based on the local grid unit of the image. So, the geometric variations influence the optical deformations. Moreover, the sampling orientation and local optimization maintain the upright posture and body movements. These movements do not influence the detection phase which is the main reason of HOG feature in detection of humans [6].

c) ***Edges:*** The boundaries of the image intensities may change during the identification of the object detection. The feature of the object detection is different from the color features technique [7].

d) ***Optical Flow:*** The feature based on the motion segmentation and the applications of the tracking. The displacement vector recognise about the every pixel of the region. The displacement vector is that which determines the transactions of each pixel of each image. Optical flow is usually used as a feature in motion-based segmentation and tracking applications. It is a dense field of the displacement vectors which defines the translation of each pixel in a region. It is computed using the brightness constraint, which assumes brightness constancy of consistent pixels in consecutive frames. With the development of technology, there are many popular techniques for computing dense optical flow, such as Horn-Schunck Algorithm [8].

2.2 Challenges of object detection

a) ***Position*** of the image can be changes at any time. In the template matching the system will handle the images uniformly in the system [9].

b) ***Lighting*** conditions may change during course of the system. The changes in the weather may affect the lighting of an image. In such case, the lighting condition may vary with the time. The shadow of the image affects the image lighting system. The detection of object from an image can be done during any condition of the lighting. [10]

c) ***Rotation*** images may be capable of handling such type of the difficulty. For instance, character may appear in any form, but the orientations of an image are not affected by the detection of the character.

d) When objects are not visible then image and that condition is referred as ***occlusion.***

e) ***Scaling method*** is the process of the recognition of the scaling of the images in the object detection [11]. The object detection systems are not affected by the change in the size of the object. The challenges may occur due to the object detection.

3. LITERATURE REVIEW

In this section, the literature on various topics considered in this proposal is being discussed.

CeLi et. al [12] proposed an object detector based on deeplearning of smallsamples. The proposed nodel uses the semantic relevance of objects to improve the accuracy of weak feature objects in complex scenarios.

Cong Tang et. al [13] discuss on the framework design and the working principle of the models and analyzes the model performance in the real-time and the accuracy of detection.

Christian Szegedy et.al [14] presents a simple and yet powerful formulation of object detection as a regression problem to object bounding box masks. It defines a multi-scale inference procedure that produces high- resolution object detections at a low cost by a few network applications.

XiaogangWang et.al [15] provides an overview of deep learning and focus on the applications in object recognition, detection, and segmentation which are the key challenges for computer vision and have numerous applications to images and videos.

ShuaiZhang et.al [16] proposes a framework for achieving tasks in a nonoverlapping multiple camera network. A new object detection algorithm using mean shift (MS) segmentation is introduced and objects are further separated with. The help of depth information derived from stereo fixed number of sliding window templates are applied there is vision. It is also possible for supervised learning in implementing the problem using Decision trees or more likely SVM in deep learning which is implemented in

XinyiZhou et. al [17] deals with the field of computer vision mainly for the deep learning in object detection task. There is a simple summary of the datasets and deep learning algorithms used in computer vision.

Girshick et al. [18] proposed a multi-stage pipeline called Regions with Convolutional Neural Networks (R-CNN) for training deep CNN to classify region proposals for object detection. It decomposes the detection problem into several stages including bounding-box proposal, CNN pre-training, CNN fine-tuning, SVM training, and bounding box regression. Such framework has shown good performance and was adopted by other methods.

Zhong-QiuZhao et.al [19] discuss a detailed review of deep learning-based object detection frameworks which handle different sub-problems, such as clutter and low resolution, with different degrees of modifications on R-CNN.

Sandeep Kumar et.al [20] deals with the Easynet model where the detection predictions with a Single network is possible. The Easynet model looks at the whole image at test time so the predictions are informed by global context.

4. FUNDAMENTAL OF DEEP LEARNING

The deep learning technology has become a buzzword. The reason behind popularity of deep are two folded, first is large availability of dataset and second is powerful graphics processing unit. The deep learning is an artificial intelligence function that imitates the workings of the human brain in processing data and creating patterns for decision making. It is a subset of machine learning in artificial intelligence which has networks capable of learning unsupervised from data.

We aim to access deep learning techniques based on Convolutional Neural Network (CNN) for object Detection. The beauty of Convolutional Neural Networks is that do not rely on manually created feature extractor of filters. Rather, they train per se from raw pixel level up to final object category.

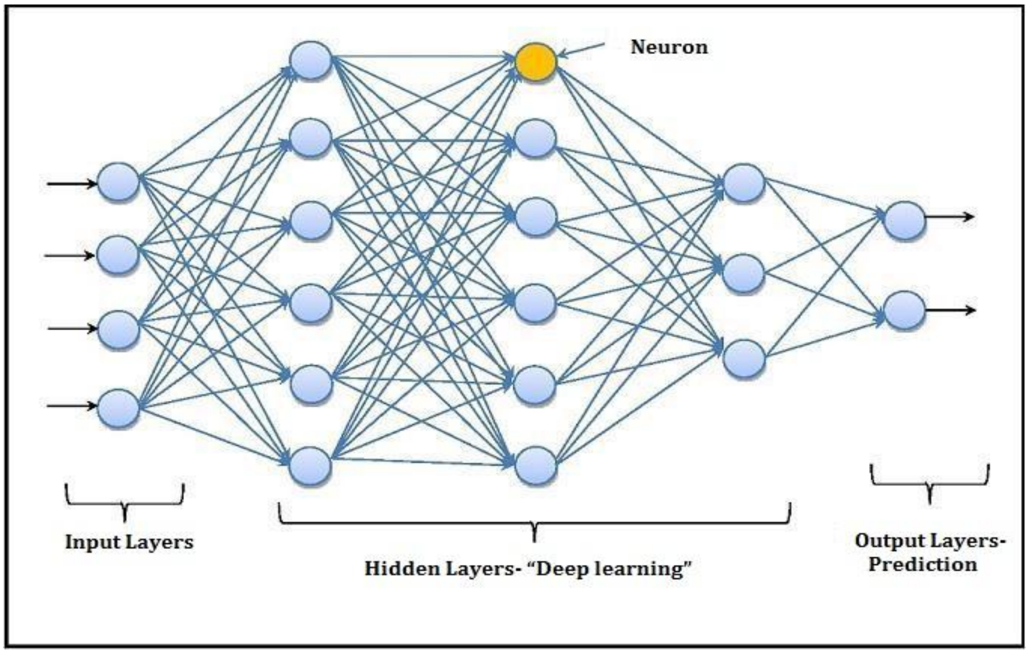


Fig 1: Architecture of Deep Learning

4.1 A Convolution Neural Network

It includes an input, an output layer and multiple hidden layers. The hidden layers of a CNN have a series of convolutional layers which convolve with a multiplication or other dot product. The activation function is generally a RELU layer which is followed by additional convolutions such as pooling layers, fully connected layers and normalization layers, referred to as hidden layers as their inputs and outputs are masked by the activation function and final convolution. It is generally a sliding dot product or cross-correlation. These CNN layers convolve the input and pass the result to the next layer.

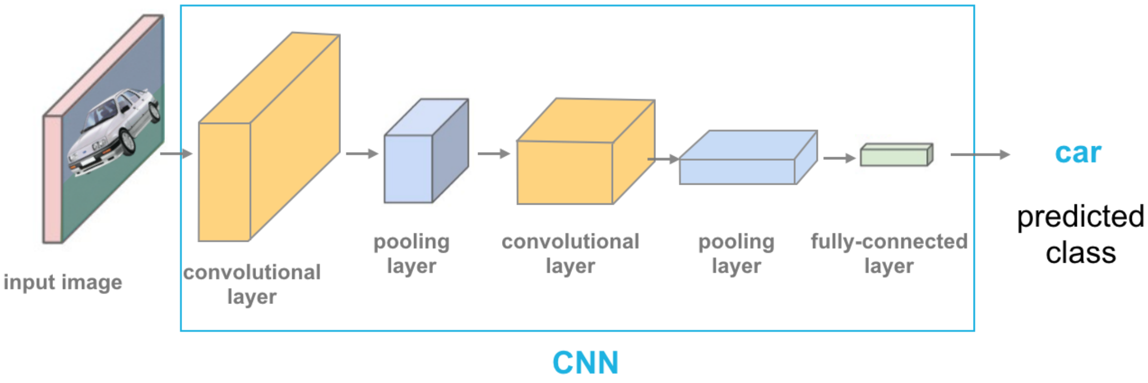
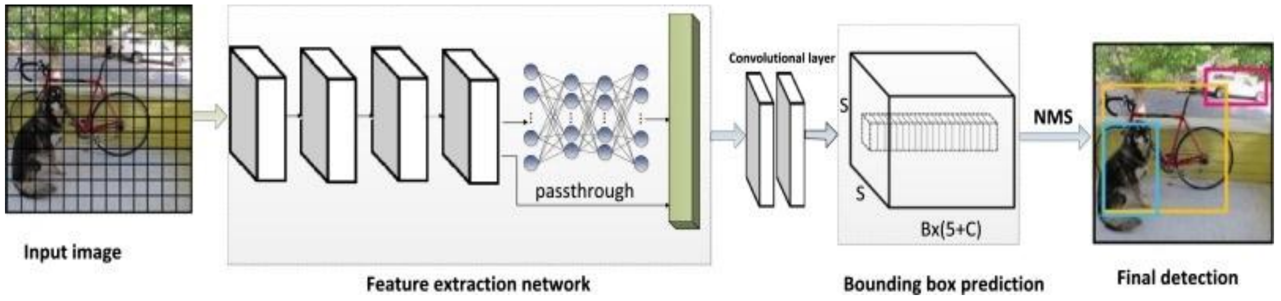


Fig 2: Working process of Convolution Neural Network

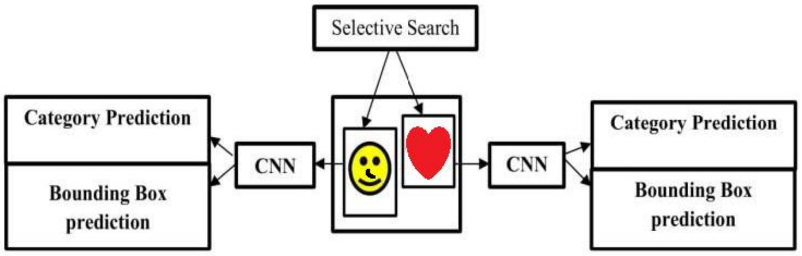
4.2 The YOLO (You Only Look Once)

It uses deep learning and convolutional neural networks (CNN) for object detection. It only needs to "see" each image once. It allows YOLO to be one of the fastest detection algorithms. It can detect objects in real time up to 30 FPS. For the detection, the image is divided in a grid of S*S (left image). Each cell will predict N possible bounding boxes and the level of probability of each one of them. This means S*S*N boxes are calculated.



4.3 R-CNN- Region based CNNs

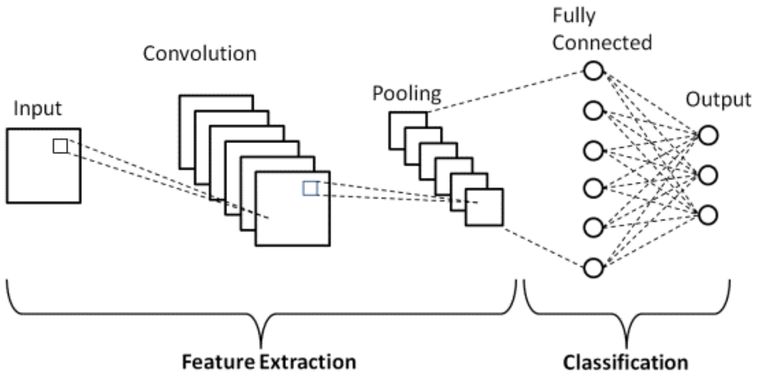
The R-CNN model first selects several proposed regions from given image and then label their categories and bounding boxes. It uses a CNN to perform forward computation to extract features from each proposed area. Then we use the features of each proposed region to predict their categories and bounding boxes.



4.4 Methodology

Basic architecture of CNN model consists of

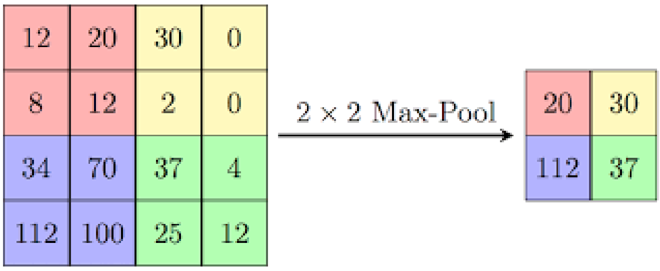
a) Input Image
b) Convolution Layer
c) Pooling Layer
d) Fully Connected Layer
e) Output Layer



a) ***Input Image:*** The input image is the image given to the model to check the output by performing various functions on it. It is given to the block named as convolution layer.

b) ***Convolution Layer:*** Convolution is the first layer of this method. This layer to extract features from the given input image. Convolution restores the relationship between pixels by learning image features using small squares of input data. It is a mathematical operation which takes two inputs such as image matrix and a filter.

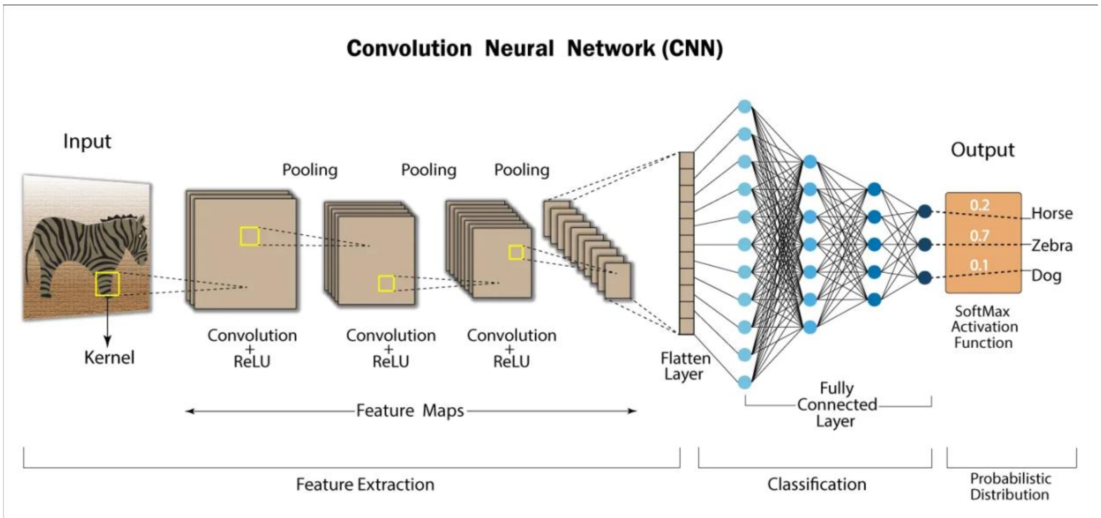
c) ***Pooling Layer:*** The pooling layers reduce the number of parameters when the images are too large. Spatial pooling also called subsampling or down sampling that reduces the dimensionality of each map without changing the important information. Spatial pooling can be of different types: Max Pooling, Average Pooling, Sum Pooling. Max pooling take the largest element from the rectified feature map. Taking the largest element could also take the average pooling. Sum of all elements in the feature map called sum pooling.



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

20	30
112	37

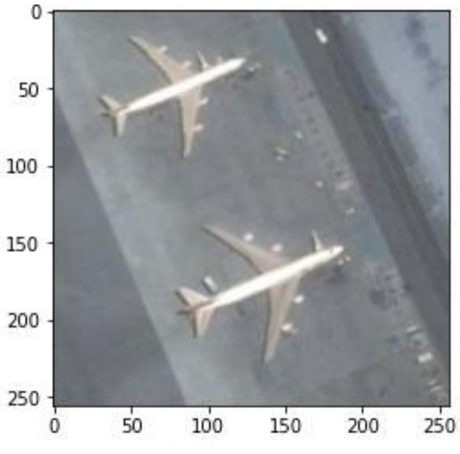
d) ***Fully Connected Layer:*** In this layer, the given matrix into vector and feed it into a fully connected layer like neural network. The fully connected layer combines all the features together to create a model. Finally an activation function such as softmax or sigmoid is used to classify the outputs as cat, dog, car, truck etc.



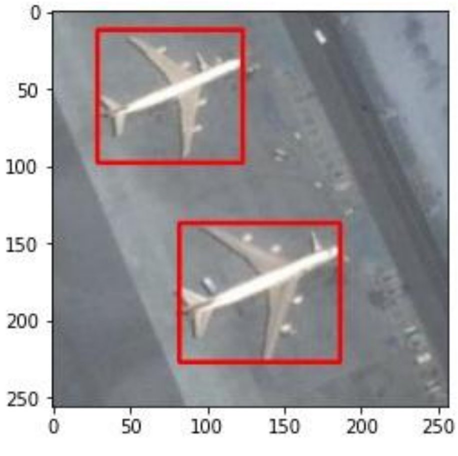
e) ***Output layer:*** The result of output image provides by the model. On the basis of the fully connected layer of the output is accepted.

5. EXPERIMENTAL RESULTS

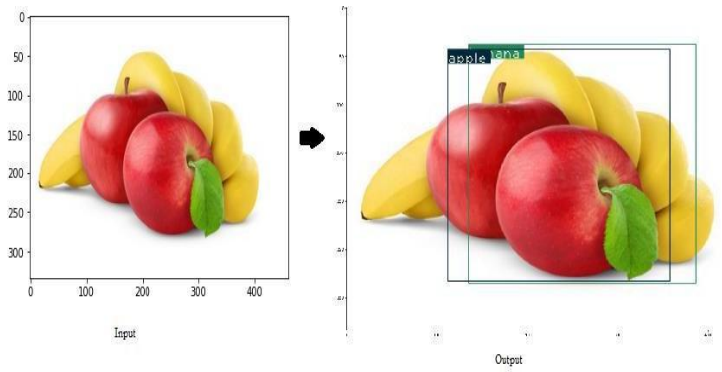
With this method, more powerful classification networks can be adopted to accomplish object detection in a fully-convolutional architecture by sharing nearly all the layers, and the results are obtained on both PASCAL VOC and Microsoft COCO datasets at a test speed of 170ms per image. The performance of a model for object detection is evaluated using the precision and recall across each of the best matching bounding boxes for the known objects in the image.

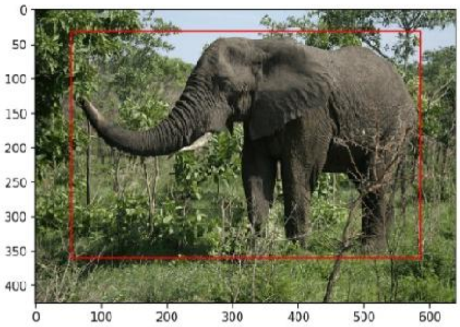


Input



Output





Input

Output

6. CONCLUSION

Firstly, it introduces the classical methodologies of object detection in deep learning. Its powerful learning ability and advantages in dealing with occlusion, scale transformation and background switches, deep learning-based object detection has been a research hotspot in recent years. We it elaborates on the common object detection model based on deep learning which handle different sub-problems, such as occlusion, clutter and low resolution, with different degrees of modifications on CNN. Afterwards, it elaborates on the common object detection model based on deep learning. Finally, this paper makes a further analysis of the challenges in object detection based on deep learning, and provides valuable insights and guidelines for future progress.

7. REFERENCES

[1] Shine L, Jiji CV (2020) Automated detection of helmet on motorcyclists from traffic surveillance videos: a comparative analysis using hand-crafted features and CNN. Multimed Tools Appl. https ://doi.org/10.1007/s1104 2-020-08627 -w

[2] Liu J, Yang Y, Lv S, Wang J, Chen H et al (2019) Attention-based BiGRU-CNN for Chinese question classification. J Ambient Intell Humaniz Comput. https ://doi.org/10.1007/s1265 2-019-01344 -9

[3] Cao D, Zhu M, Gao L et al (2019) An image caption method based on object detection. Multimed Tools Appl 78(24):35329–35350.

[4] Patel, D., & Gautam, P. K. (2015). A Review Paper on Object Detection for Improve the Classification Accuracy and Robustness using different Techniques. *International Journal of Computer Applications*, vol. 112, no 11, pp- 975, 8887.

[5] Papageorgiou, C. P., Oren, M., &Poggio, T. (1998). A general framework for object detection. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), IEEE, vol 5, no.2,* pp. 555-562).

[6] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Nonlinear estimation and classification, Springer, New York, NY, vol 2, no.3, pp. 149-171.

[7] Parekh, H. S., Thakore, D. G., &Jaliya, U. K. (2014). A survey on object detection and tracking methods. International Journal of Innovative Research in Computer and Communication Engineering, vol2, no. 2, pp. 2970-2979.

[8] Awan, S. (2014). Object Class Recognition Using Global Shape Descriptors in 3D (Doctoral dissertation).,vol 2, no.1, pp. 345- 389.

[9] Mashak, S. V., Hosseini, B., Mokji, M., & Abu-Bakar, S. A. R. (2010). Background subtraction for object detection under varying environments. In 2010 International Conference of Soft Computing and Pattern Recognition IEEE, vol 3, no.6, pp. 123- 126.

[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, vol2, no. 4, pp. 580-587.

[11] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, vol 3, no.2, pp. 580-587.

[12] Ce Li, Yachao Zhang and YanyunQu, "Object Detection Based on Deep Learning of Small Samples,"

International Conference, pp.1-6, March 2018.

[13] Cong Tang, YunsongFeng, Xing Yang, Chao Zheng and

Yuanpu Zhou, "The Object Detection Based on Deep Learning," International Conference, pp.1-6, 2017.

[14] Christian Szegedy, Alexander Toshev and DumitruErhan, "Deep Neural Networks for Object Detection," IEEE, pp.1-9, 2007.

[15] Xiaogang Wang, "Deep Learning in Object Recognition, Detection, and Segmentation," IEEE, pp.1-40, Apr. 2014.

[16] Shuai Zhang, Chong Wang and Shing-Chow Chan, "New Object Detection, Tracking, and Recognition Approaches for Video Surveillance Over Camera Network," IEEE SENSORS JOURNAL, vol. 15, no.69, pp. 1-13, May 2015.

[17] Xiao Ma, Ke Zhou and JiangfengZheng, "Photo-Realistic Face Age Progression/Regression Using a

Single Generative Adversarial Network," Neurocomputing, Elsevier B.V., pp.1-16, July 2019.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR, 2014. 1, 2

[19] Zhong-Qiu Zhao, PengZheng, Shou-taoXu and Xindong Wu, "Object Detection with Deep Learning: A Review," IEEE, pp.1-21, 2019.

[20] Sandeep Kumar, AmanBalyan and ManviChawla, "Object Detection and Recognition in Images," IJEDR, pp.1-6, 201