EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Object detection in real time based on improved single shot multi-box detector algorithm

Ashwani Kumar[1], Zuopeng Justin Zhang[2] and Hongbo Lyu[3*]

* Correspondence: lvhongbo@zwu.edu.cn
[3]Logistics and E-Commerce School, Zhejiang Wanli University, Ningbo, Zhejiang 315100, China
Full list of author information is available at the end of the article

Abstract

In today's scenario, the fastest algorithm which uses a single layer of convolutional network to detect the objects from the image is single shot multi-box detector (SSD) algorithm. This paper studies object detection techniques to detect objects in real time on any device running the proposed model in any environment. In this paper, we have increased the classification accuracy of detecting objects by improving the SSD algorithm while keeping the speed constant. These improvements have been done in their convolutional layers, by using depth-wise separable convolution along with spatial separable convolutions generally called multilayer convolutional neural networks. The proposed method uses these multilayer convolutional neural networks to develop a system model which consists of multilayers to classify the given objects into any of the defined classes. The schemes then use multiple images and detect the objects from these images, labeling them with their respective class label. To speed up the computational performance, the proposed algorithm is applied along with the multilayer convolutional neural network which uses a larger number of default boxes and results in more accurate detection. The accuracy in detecting the objects is checked by different parameters such as loss function, frames per second (FPS), mean average precision (mAP), and aspect ratio. Experimental results confirm that our proposed improved SSD algorithm has high accuracy.

Keywords: Single shot multi-box detector (SSMB), Faster region convolutional neural networks (F-CNN), Loss function, Aspect ratio

1 Introduction

The information age has witnessed the rapid development of wireless network technology, which has attracted the attention of researchers and practitioners due to its unique characteristics such as flexible structure and efficiency. As wireless network technology continues to evolve, it has brought great convenience to people's life and work with its powerful technical capabilities. Wireless networks have gradually facilitated the main stream of people's online life. At the same time, the advent of 5G network will further enable the greater development and more advanced applications of wireless network technology. The future generations of wireless networks will provide strong support for related applications such as Internet of Things (IoT) and virtual reality (VR). Many of these applications connect to each other and transmit information within networks based on the detection of specific target objects. In order to achieve a comprehensive network connection between people and people, things and people, and things and things, one of the key tasks of future applications is to identify the target in a real-time manner in the wireless networks [1].

Identifying each object in a picture or scene with the help of computer/software is called object detection. Object detection is one of the most important problems in the area of wireless network computer vision. It is the basis of complex vision tasks such as target tracking and scene understanding and is widely used in wireless networks. The task of object detection is to determine whether there are objects belonging to the specified category in the image. If it exists, then the subsequent task is to identify its category and location information. Traditional object detection algorithms are mainly devoted to the detection of a few types of targets, such as pedestrian detection [2] and infrared target detection [3]. Due to the recent advance of deep learning technology [4], especially after the appearance of the deep convolution neural network (CNN) technology, object detection algorithms have made a breakthrough development. Within these algorithms, three major methods widely adopted in this field are You Only Look Once (YOLO), single shot multi-box detector (SSD), and faster region CNN (F-RCNN) [5].

However, with the upcoming of 5G, the characteristics of wireless network, such as massive data, service evolution, data diversification, and uneven spatial-temporal distribution of data, have posed severe challenges to object detection under a real-time environment. Besides, real-time object detection also needs to be completed on any device and in any environment. To address the challenges, this paper proposes object detection technique to detect objects in real time with a model that can be executed on any device in any environment. Specifically, our proposed method applies convolutional neural networks to develop a model that consists of multiple layers to classify the given objects into several defined classes. Based on the recent advancement in deep learning with image processing, the proposed schemes then use multiple images and detect the objects from these images, labeling them with their respective class label. These images can be from videos which are fed into the model we prepared, and the training of the model takes place until the error rate is reduced to an acceptable level. To speed up the computational performance of the object detection technique, we have used improved single shot multi-box detector (SSD) algorithm along with the faster region convolutional neural network. We also conduct experiments to check the accuracy of our proposed method in detecting the objects with different parameters including loss function, mean average precision (mAP), and frames per second. The experiment results demonstrate that the proposed model has a high performance in detect accurate objects for real-time applications.

Specifically, this research makes contributions to the existing literature by improving the accuracy of SSD algorithm for detecting smaller objects. SSD algorithm works well in detecting large objects but is less accurate in detecting smaller objects. Hence, we modify the SSD algorithm to achieve acceptable accuracy for detecting smaller objects. The images or scenes are taken from web cameras and we have used Pascal visual object class (VOC) and common objects in context (COCO) datasets to carry out experiments. We capture object detection (OD) datasets from our center for image processing lab. We make use of different libraries to form a network and use tensorflow-GPU 1.5. For experimental setup, tensorflow directory, SSD MobilenetV1 FPN Feature Extractor, tensorflow object detection API, and anaconda virtual environment are used. This entire setup enables us to produce real-time object detection in a better way.

The rest of this paper is organized as follows. The next section summarizes related work with a focus on the existing techniques of object detection. The third section discusses about the improved SSD algorithm. The fourth section represents the experimental results. The fifth section describes discussion and analysis, limitations, and future research directions. The final section concludes the paper.

2 Related work

2.1 Computer vision detection

In 2012, Alex [6] used the deep CNN Alex Net to win the championship in the task of ILSVRC 2012 image classification, which was superior to the traditional algorithms. Then scholars began to study the application of deep CNN in object detection. They used Alex Netto construct algorithms, such as R-CNN [7–9], YOLO [5], SSD [10], and others, which resulted in a surging research stream of computer vision detection.

Girshick et al. [8] proposed a method R-CNN by successfully combining region proposals with CNNs, which improves mean average precision (mAP) by more than 30%. The next year Girshick [11] named a new algorithm faster R-CNN, which employs spatial pyramid pooling networks. But it had a bottleneck in region proposal computation. In order to overcome this disadvantage, Ren et al. [12] successfully introduced a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network. Cao et al. [13] proposed a rotation invariant faster R-CNN target detection algorithm. By adding regularization constraints to the target function of the model, the invariance of the target CNN feature rotation is enhanced. The model improved the accuracy by an average of 2.4%. Dai et al. [14] proposed position-sensitive score maps to address a dilemma between translation-invariance in image classification and translation variance in object detection and successfully executed them 2.5–20 times faster than the F-RCNN counterpart. Lin et al. [15] developed a top-down architecture with lateral connections, called Feature Pyramid Network (FPN), by building high-level semantic feature maps at all scales. All these algorithms successfully solved the problem in object detection. However, there are still defects in accuracy and speed for wireless network object detection applications.

In order to get a better computing speed, Redmon et al. [5] proposed a new YOLO algorithm to object detection. It achieved double mAP in real-time detectors. Then Redmon et al. [16] put forward an improved algorithm YOLO V2. On the basis of YOLO, batch normalization [17] was added in the algorithm to speed up the training and the algorithm adds anchor boxes and high resolution classifier to improve accuracy. Result shows that it runs significantly faster than F-RCNN with ResNet [18] and SSD [19]. To achieve high speed and accuracy rate, scholars further optimized the YOLO V2. For instance, Wei et al. [20] used the dimension clustering of object box, classified network pre-training, conducted multi-scale detection training, changed the candidate box filtering rules and other methods, and made the algorithm better adaptive to the location task and object detection. It increased the average accuracy rate of the detection network to 79.5%. Redmon et al. [21] proposed the third version of the YOLO series, YOLO V3, which improved the algorithm at the accuracy of detection.

For SSD algorithm, researchers also made many other improvements, such as DSSD [22], F-SSD [23], and R-SSD [24]. They all improved the fusion method of different features. However, it has some difficulties at expressing the shallow features of the prediction layer. To address some of the concerns, Fu et al. [22] offered a Deconvolutional Single Shot Detector (DSSD) by combining a classifier (Residual-101 [25]) with an SSD [26]. Wang et al. [27] proposed an improved SSD algorithm based on target proposal and one-stage target detection to improve the target detection performance. It improved the mAP in the small target detection by 14.46% and 13.92% compared to F-RCNN [13] and R-FCN [14] algorithms. Lin et al. [28] designed a detector RetinaNet to address extreme foreground-background class imbalance by reshaping the standard cross-entropy loss.

2.2 Relevant systems

A deep neural network is basically consisted of two different models: the first is convoluted and the other is non-linear relationships. In both models, an object is considered as a layered configuration of primitives. Numerous architectures and algorithms have implemented the concept of deep learning neural networks including belief network, stacked network, and gated recurrent unit. The first CNN was constructed by LeCun et al. [29]. The different application domains of CNN now include image-processing, handwriting character recognition, etc. Object detection is performed by estimating the coordinates and class of particular objects in the picture. The presence of these objects in a picture may be in random positions. We next summarize the details of faster RCNN and YOLO v3 architecture as they are directly relevant to our proposed method.

2.2.1 Faster RCNN

Region Proposal Network for generating regions and detecting objects uses two methods of fast RCNN. The first method proposes regions and uses the proposed regions respectively. In fast RCNN, Ren et al. [12] has used 16 architectures in convolution layers to achieve detection and classification accuracy on datasets. Kumar et al. [30–32] proposed a method to detect the objects with audio device in real time for blind people using deep neural network. Figure 1 demonstrates the architecture of Faster RCNN. There is a limitation in Faster R-CNN that it has a complex training process and slow processing speed.

2.2.2 YOLO V3

YOLO V3 is a detector of objects which makes use of features learned by a deep convolutional neural network for detecting object in real time [21]. It consists of 75 convolutional layers with up-sampling layers and skips connections for the complete image one neural network being applied. Regions of the image are made. Later bounding boxes are displayed along with probabilities. The most noticeable feature of YOLO V3 is that the detections at three different scales can be done with the help of it. But the speed has been traded off for boosts in accuracy in YOLO v3, and it does not perform well

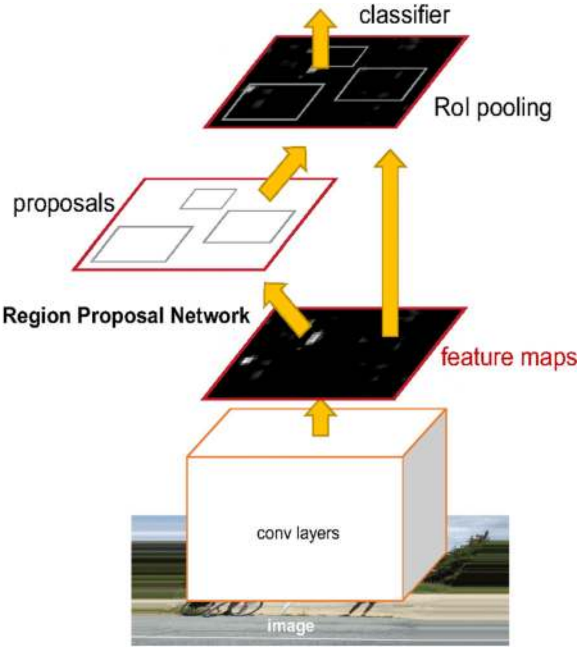


Fig. 1 Architecture of F-RCNN [12]. Demonstrates the architecture of Faster RCNN. There is a limitation in faster R-CNN that it has a complex training process and slow processing speed

with small objects that appear in groups. Figure 2 represents the working mechanism of the YOLO model.

2.2.3 Our contribution

To highlight our contribution to the existing literature, we next summarize some of the key points of our proposed object detection technique based on the improved SSD algorithm.

1. The improved SSD algorithm uses depth-wise separable convolution and spatial separable convolutions in their convolutional layers. The depth-wise separable convolution performs operations such that it maps each number of input channel with

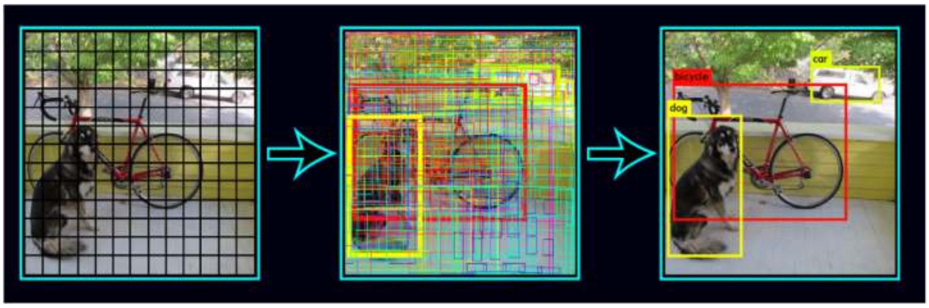


Fig. 2 A YOLO model [21]. Represents the working of YOLO model for detecting the objects from the image. But the speed has been traded off for boosts in accuracy in YOLO v3 and it does not perform well with small objects that appear in groups. The fast YOLO model has lower scores but good real-time performances

its corresponding number of output channel separately. Spatial separable convolution is the same as depth-wise convolution along the x- and y-axis.

2. This architecture reduces the number of operations to execute the algorithm in fast speed through ways used by depth-wise separable convolution to reduce the number of channels with the help of width multiplier and those used by spatial separable convolution to reduce the feature maps of spatial dimensions by applying resolution multiplier.
3. We use mAP and FPS as standard parameters for object detection. The major objective during the training is to get a high-class confidence score.
4. The proposed approach enables us to produce real-time object detection by using optimal values of aspect ratio.
5. Our improved SSD algorithm uses many default boxes, which results in more accurate detection of objects.

3 Methodology

This section presents our proposed approach for detecting the objects in real-time from images by using convolutional neural network deep learning process. The previous algorithms such as CNN, faster CNN, faster RCNN, YOLO, and SSD are only suitable for highly powerful computing machines and they require a large amount of time to train. In this paper, we have tried to overcome the limitations of the SSD algorithm by introducing an improved SSD algorithm with some improvement. The proposed scheme uses improved SSD algorithm for higher detection precision with real-time speed. However, SSD algorithm is not appropriate to detect tiny objects, since it overlooks the context from the outside of the boxes. To address this issue, the proposed algorithm uses depth-wise separable convolution and spatial separable convolutions in their convolutional layers. Specifically, our proposed approach uses a new architecture as a combination of multilayer of convolutional neural network. The algorithm comprises of two phases. First, it reduces the feature maps extraction of spatial dimensions by using resolution multiplier. Second, it is designed with the application of small convolutional filters for detecting objects by using the best aspect ratio values. The major objective during the training is to get a high-class confidence score by matching the default boxes with the ground truth boxes. The advantage of having multi-box on multiple layers leads to significant results in detection. Single shot multi-box detector was discharged at the tip of Gregorian calendar month 2016 and thus arrived at a new set of records on customary knowledge sets like Pascal VOC and COCO. The major problem with the previous methods was how to recover the fall in precision, for which SSD applies some improvements that include multi-scale feature map and default boxes. For detecting a small object with higher resolutions, feature maps are used. The training set of improved SSD algorithm depends upon three main sections, i.e., selecting the size of box, matching of boxes, and loss function. The proposed scheme can be understood by the system model given in Fig. 3.

3.1 SSMBD algorithm

In order to interpret the role of SSD algorithm, we first formally denote the following concepts.

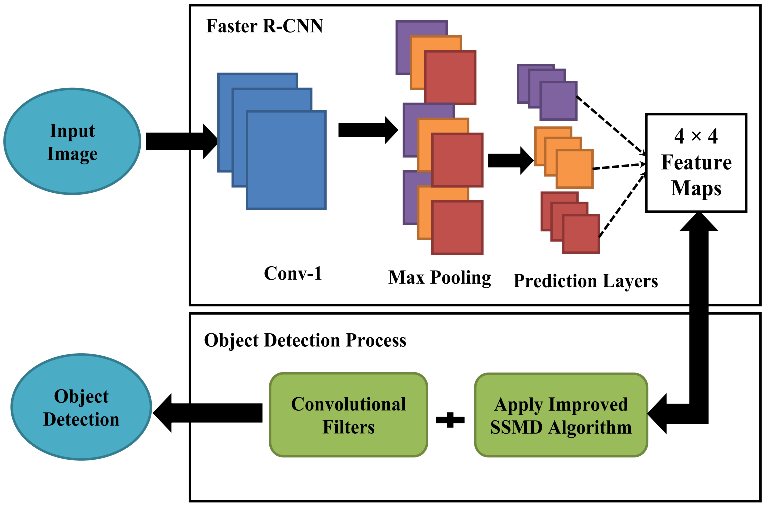


Fig. 3 The proposed system model. The training set of improved SSD algorithm depends upon three main sections, i.e., selecting the size of box, matching of boxes, and loss function. The proposed scheme can be understood by the system model given in Fig. 3

Single shot: This means that the tasks of the thing localization and classification are exhausted one passing play of the network.

Multi-box: Ground truth box and predicted box are the boxes in multi-box. This is introduced by Szegedy [33].

Detector: The network is an associate degree object detector that conjointly classifies those detected objects.

Default the size of the boxes: The selection of boxes is based on the minimum value of convolution layer and maximum values of change in intensity [34]. The first algorithm represents the procedure of producing specified feature maps F(m).

Truth boxes: After finding the size of boxes, the next phase is matching of the boxes with the corresponding truth boxes. A specific given picture to identify the truth boxes is explained in the second algorithm.

Loss function: The loss function is unbelievably simple, and it is a methodology of evaluating how well your role models your dataset. If your predictions are entirely of your loss function, it can operate next range. If the output range is less, it means that the model is good. The main objective is to minimize loss function. The loss function is also depending upon the sum of weighted localization and classification loss functions [35].

When a color image is fed into the input layer, SSD does the following.

Step 1: Image is passed through large number of convolutional layers extracting feature maps at different points.

Step 2: Every location in each of those feature maps uses a 4x4 filter to judge a tiny low default box.

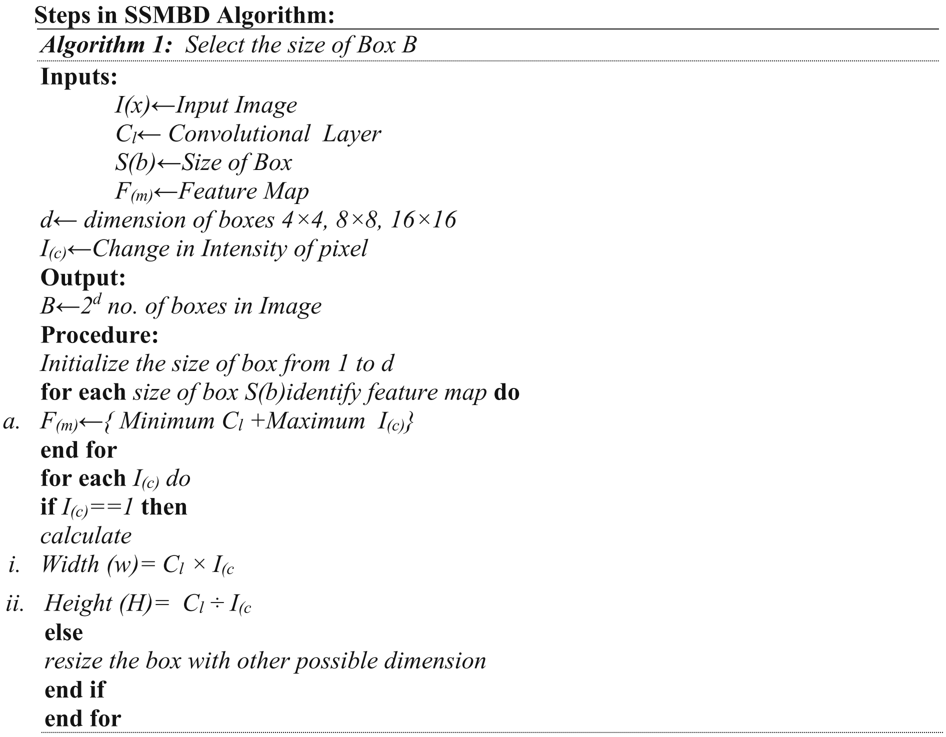
Step 3: Predict the bounding box offset for each box.

Step 4: Predict the class probabilities for each box.

Step 5: Based on IOU, the truth boxes are matched with the predicted boxes.

Step6: Instead of exploiting all the negative examples, the result exploits the best-assured loss for every default box.

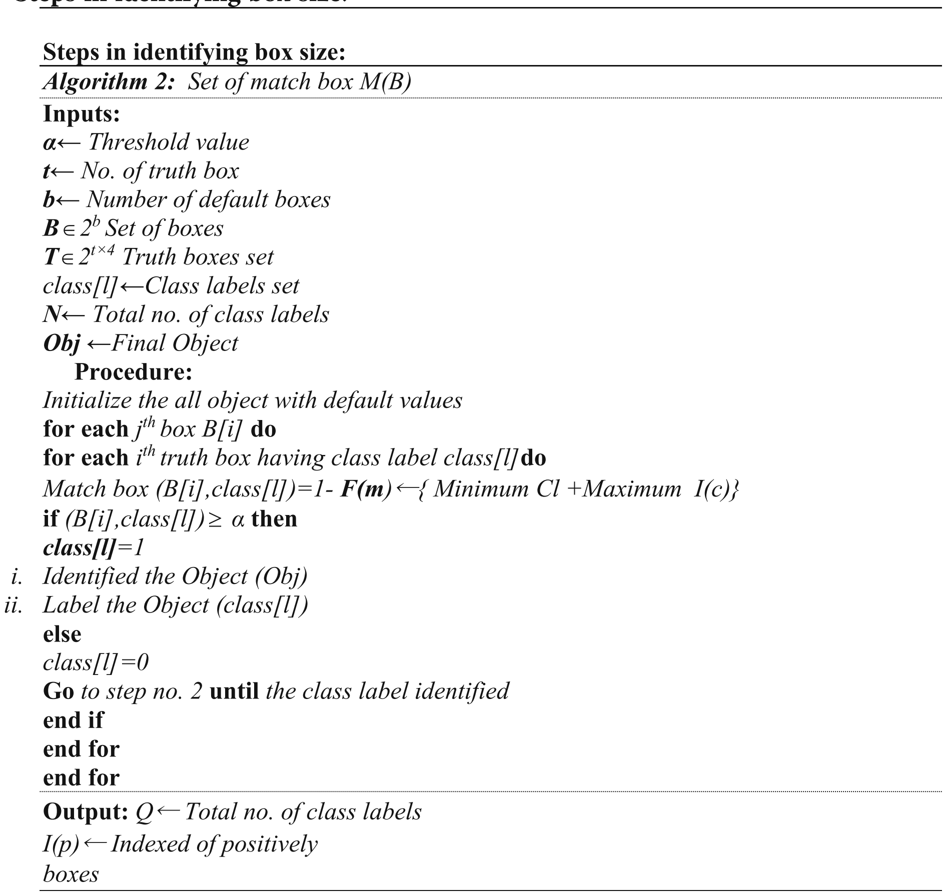
Steps in SSMBD Algorithm:



Steps in SSMBD Algorithm:

```
Algorithm 1: Select the size of Box B
Inputs:
        $I(x)$ ← Input Image
        $C_l$ ← Convolutional Layer
        $S(b)$ ← Size of Box
        $F_{(m)}$ ← Feature Map
$d$ ← dimension of boxes 4×4, 8×8, 16×16
$I_{(c)}$ ← Change in Intensity of pixel
Output:
$B \leftarrow 2^d$ no. of boxes in Image
Procedure:
Initialize the size of box from 1 to d
for each size of box S(b)identify feature map do
a.  $F_{(m)} \leftarrow \{$ Minimum $C_l$ +Maximum $I_{(c)}\}$
end for
for each $I_{(c)}$ do
if $I_{(c)}==1$ then
calculate
i.  Width $(w)= C_l \times I_{(c}$
ii. Height $(H)= C_l \div I_{(c}$
else
resize the box with other possible dimension
end if
end for
```

Steps in identifying box size:



Steps in identifying box size:

```
Algorithm 2: Set of match box M(B)
Inputs:
$\alpha$ ← Threshold value
$t$ ← No. of truth box
$b$ ← Number of default boxes
$B \in 2^b$ Set of boxes
$T \in 2^{t\times 4}$ Truth boxes set
class[l] ← Class labels set
$N$ ← Total no. of class labels
Obj ← Final Object
    Procedure:
Initialize the all object with default values
for each $j^{th}$ box B[i] do
for each $i^{th}$ truth box having class label class[l] do
Match box (B[i],class[l])=1- F(m) ←{ Minimum Cl +Maximum I(c)}
if (B[i],class[l]) ≥ $\alpha$ then
class[l]=1
i.  Identified the Object (Obj)
ii. Label the Object (class[l])
else
class[l]=0
Go to step no. 2 until the class label identified
end if
end for
end for
Output: Q ← Total no. of class labels
I(p) ← Indexed of positively
boxes
```

Figure 4 shows the process of identifying total number of default boxes, and Fig. 5 demonstrates the process of detecting objects with different color boxes.

4 Experimental results

This study proposes object detection technique to detect objects in real time on any device running the proposed model in any environment. We use python programming language and OpenCV 2.4 library to execute the proposed system. Python libraries are the open source framework for the construction, training, and identification of object detection. The chosen datasets taken into consideration for this research were bound to a group of people. Multi-scale feature extraction may improve the accuracy for detecting big object but does not exhibit a good precision of speed to detect small objects. Therefore, we have used depth-wise separable convolution along with spatial separable convolutions to achieve this. For conducting the experiments and producing the results, we use Pascal VOC[1] and COCO[2] object detection (OD) datasets from our center for image processing lab.

4.1 Experimental setup

We make use of different libraries to form a network and also use tensorflow-GPU 1.5. Once the training is done our next objective is to test the model for accuracy. The next objective is to optimize the model as a tensorflow serving and deploy that to the environment which we want to use. For experimental setup tensorflow directory, SSD MobilenetV1 FPN feature extractor, tensorflow object detection API, and anaconda virtual environment are used. This entire setup enables us to produce real-time object detection in a better way. To achieve great precision, we have increased the number of default boxes with less confidence and focus on the boxes having high confidence.

4.2 Performance etrics

The performance metrics which are used to evaluate the performance of improved SSD algorithm to predict the boundary boxes and truth boxes for classification of object are discussed here. These metrics include mAP, FPS, aspect ratio, logistic regression, and Intersection over Union (IoU). The box regression technique of SSD is used to identify the bounding box coordinate.

- The accuracy is calculated using Equation (1) below, which could be improved over the original dataset.

$$A_{ccuracy} = \frac{Object(O_{\mathrm{correct}})}{TotalObject(T_{\mathrm{obj}})} \tag{1}$$

In the equation, O_{correct} represents the number of correctly detected object and T_{obj} the total number of images.

[1] http://host.robots.ox.ac.uk/pascal/VOC/

[2] http://cocodataset.org/#home

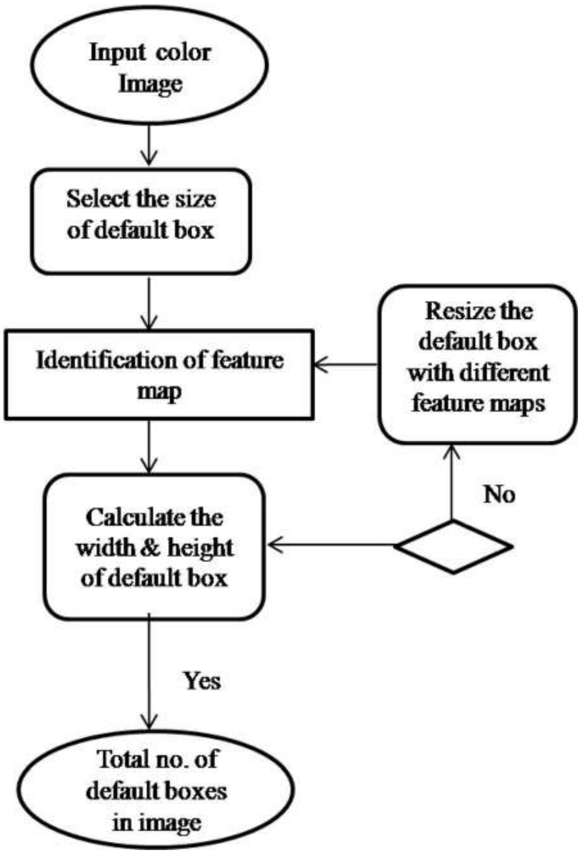

Input color Image
Select the size of default box
Identification of feature map
Resize the default box with different feature maps
No
Calculate the width & height of default box
Yes
Total no. of default boxes in image

Fig. 4 Process of identifying total number of default boxes. Shows the process of identifying total number of default boxes

- IoU is calculated by the Jaccard index to find out the overlap between two bounding boxes [35]. Equation (2) shows the formula for IoU.

$$IoU = \frac{area\ of\ overlap}{area\ of\ union} \tag{2}$$

- Logistic regression is a model which identifies the probability of a result being obtained. We have to segregate our problem dataset into different class labels. Logistic regression model usually gives one of the highest accuracies of all classification models [36]. Equation (3) indicates the logistic regression function.

$$Ln\left(\frac{p}{1-p}\right) = b_0 + b_1{*}x \tag{3}$$

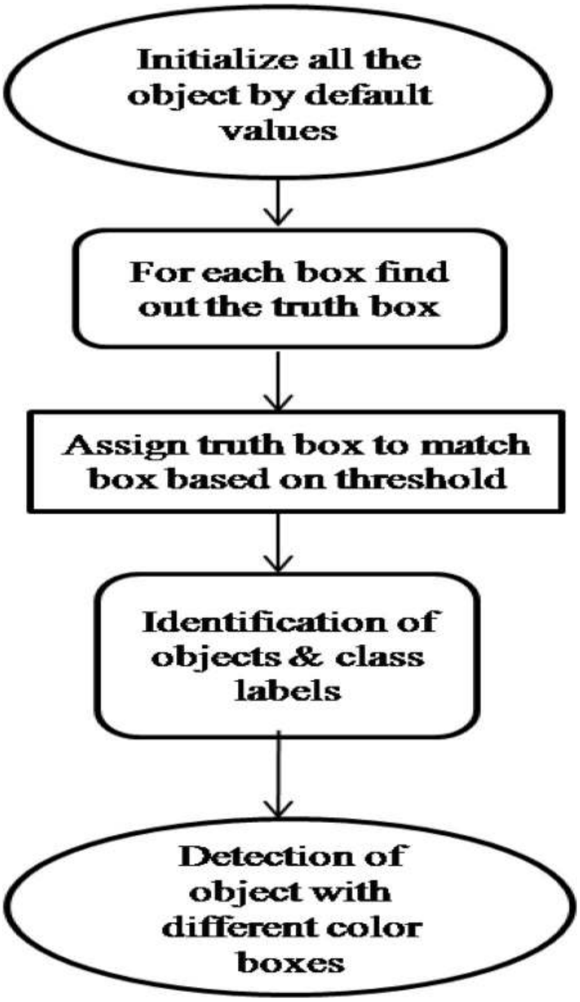

Initialize all the object by default values
For each box find out the truth box
Assign truth box to match box based on threshold
Identification of objects & class labels
Detection of object with different color boxes

Fig. 5 Process of detecting objects with different color boxes. Demonstrates the process of detecting objects with different color boxes

- Aspect ratio is used to find out the relationship between the width and height of the image. Basically, it represents a shape of an image. We have used $A_R$to represent the aspect ratio.

$$A_R = \frac{Width\ of\ image}{Height\ of\ image} \tag{4}$$

Figure 6 demonstrates the different sample images taken from object detection (OD) datasets from our center for image processing lab used in the experimental setup.

Figure 7 represents the various objects detected by the proposed algorithm. In this research work, we have used different colors of boxes to show different class labels. Our scheme correctly detects and recognizes bottle, laptop, mouse, cup, teddy bear, umbrella, person, keyboard, TV, zebra, toy car, bowl, chair, bird, vassal, and suitcase.

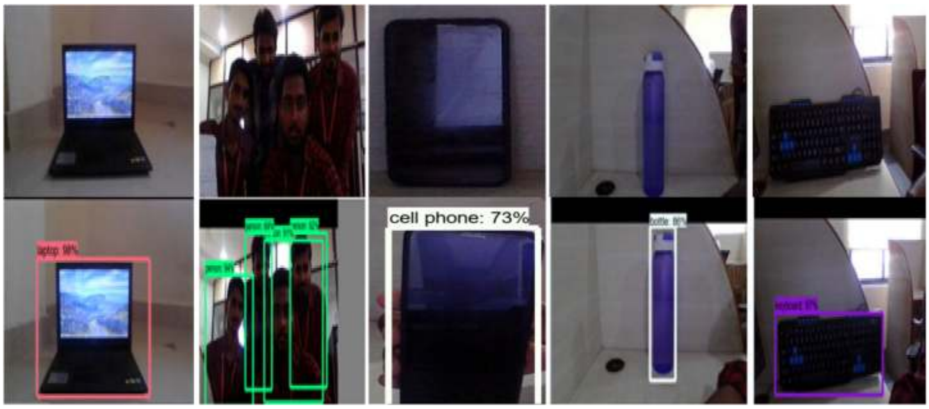


Fig. 6 Some sample images and object detection using improved SSD model. Demonstrates the different sample images taken from object detection (OD) dataset from our Centre for Image processing Lab used in the experimental setup

5 Discussion and analysis

We have analyzed the correctness of our improved SSD algorithm which uses depthwise separable convolution along with spatial separable convolutions generally called multilayer to increase the classification accuracy of detecting small objects without affecting the speed. These multilayer convolutional neural networks use the confidence

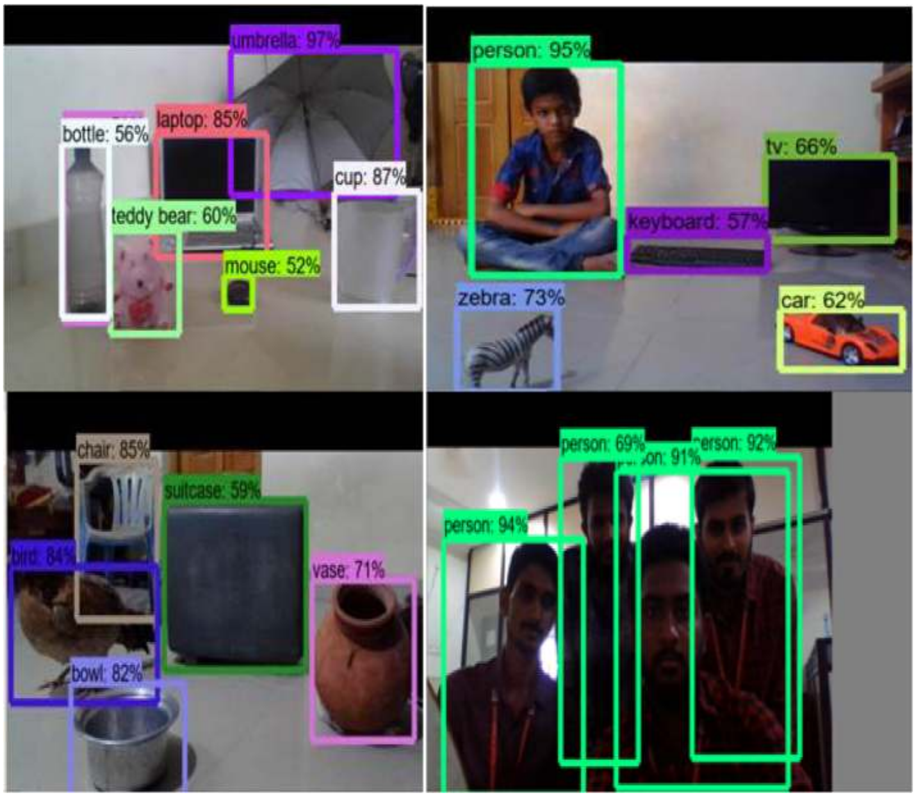


Fig. 7 Detection of objects with different boxes using proposed approach. Represents the various objects detected by the proposed algorithm. We have used different colors of boxes to show different class labels. Our scheme correctly detects and recognizes bottle, laptop, mouse, cup, teddy bear, umbrella, person, keyboard, TV, zebra, toy car, bowl, chair, bird, vassal, and suitcase

value for improving the process of detecting accurate boxes. For experimental setup, tensorflow directory, SSD MobilenetV1 FPN feature extractor, tensorflow object detection API, and anaconda virtual environment are used. The algorithm includes width multiplier and resolution multiplier to minimize the channels, and feature maps as well. The proposed approach produces real-time object detection by using aspect ratio. Our improved SSD algorithm consists of large amounts of data, easy trained model, and faster GPUs, which allows to detect and classify multiple objects within an image with high accuracy. The key functions of the proposed algorithm are object detection, object localization, loss function, default boundary box, truth box, feature map, and localization. In the object detection techniques, the selection convolutional layer plays a vital role to improve from 65.5 to 74.3% with respect to mAP. In the case of default box shapes, it improves from 71.6 to 74.3% with respect to mAP. Our improved SSD algorithm uses the 4 × 4 feature maps along with a greater number of default boxes, resulting in a more accurate detection. While comparing with the other previous models, the testing speed of our proposed model is still faster because our approach gives 79.8% of mAP and 89 FPS. We also compare with other feature extraction model such as YOLO, SSD512, SSD300, and F-CNN to obtain the results. Table 1 demonstrates the comparison between F-CNN, YOLO, SSD512, SSD300, and our proposed model. We have combined faster R-CNN with SSD together to achieve high accuracy and FPS with good speed to detect objects in real time as well. Table 1 represents the different parameter of the improved SSD algorithm by using VOC and COCO test datasets.

Table 2 shows the performance of different machine learning algorithm as image classifiers namely convolution neural network, faster R-CNN, R-CNN, and faster R-CNN VGG, ZF.

Table 3 represents the different values of mAP on Pascal VOC and COCO datasets. It is obvious that improved SSD with multi-scale contexts meets our demand as the best solution.

Although we have improved the SSD algorithm, there are certain limitations in our research, such as blockage, deformable objects, corrupt objects, and interlaced objects. One more limitation of our object detection algorithm is its inability to deal with new object classes. Although we have trained our model for every possible object class, this problem can occur when an anonymous object is present in the image.

For detecting the object, we have used different deep learning algorithms as object classifiers namely convolution neural network and logistic regression. We have applied four different object detection algorithms like SSD512, SSD300, YOLO, and F-CNN to obtain the various small objects from the images with respect to Intersection over

Table 1 represents the results on Pascal VOC and COCO test

System model	mAP	FPS	No. of boxes	Resolution
F-CNN	73.2	7	6000	1000 × 600
YOLO	66.4	155	98	448 × 448
SSD512	76.8	19	24564	512 × 512
SSD300	74.3	46	8732	300 × 300
Proposed approach	79.8	89	5988	1024 × 1024

Table 2 Speed and performances for trained model with the 2007 and 2012 VOC datasets

Model	mAP	FPS	Real-time speed
Faster 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

Union (IoU). The IoU curves and results demonstrate that our proposed approach gives the highest accuracy of 96.7%. Figure 8 a, b, and c show the different curves implemented on SSD512, SSD300, amd YOLO V3-based detector.

The proposed improved SSD approach has a higher recall value, i.e., 0.9, compared with that from YOLO, faster RCNN, NASNet, and R-FCN. Figure 9 demonstrates the graph of recall percentage versus threshold IoU compared with other object detection techniques such as YOLO, Faster-RCNN, NASNet, and R-FCN on object detection (OD) datasets from our center for image processing lab. The recall value of improved SSD algorithm is 79.8% when the different value of IoU is applied.

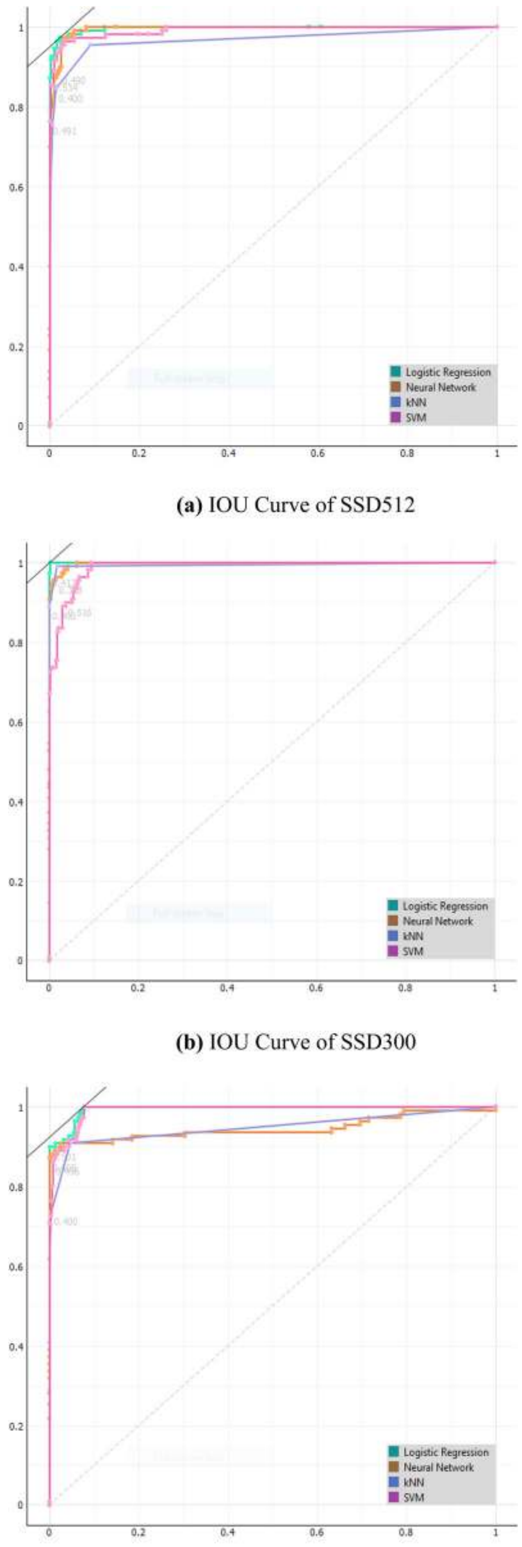
6 Conclusion

This study develops an object detector algorithm using deep learning neural networks for detecting the objects from the images. The research uses am improved SSD algorithm along with multilayer convolutional network to achieve high accuracy in real time for the detection of the objects. The performance of our algorithm is good in still images and videos. The accuracy of the proposed model is more than 79.8%. The training time for this model is about 5–6 h. These convolutional neural networks extract feature information from the image and then perform feature mapping to classify the class label. The prime objective of our algorithm is to use the best aspect ratios values for selecting the default boxes so that we can improve SSD algorithm for detecting objects.

For checking the effectiveness of the scheme, we have used Pascal VOC and COCO datasets. We have compared the values of different metrics such as mAP, loss function,

Table 3 The value of mAP on the 2007, 2010, 2012 VOC dataset, and 2015, 2016 COCO datasets

Model	Pascal VOC 2007	Pascal VOC 2010	Pascal VOC 2012	COCO 2015 (IoU = 0.5)	COCO 2016 (IoU = 0.5)	Real-time speed
YOLO	63.7%	Not given	57.9%	Not given	Not given	Yes
SSD	83.2%	Not given	82.2%	48.5%	Not given	No
YOLO V2	78.6%	Not given	Not given	44.0%	Not given	Yes
Faster-RCNN	78.8%	Not given	75.9%	Not given	Not given	No
R-CNN	Not given	62.4%	Not given	Not given	Not given	No
Mask R-CNN	Not given	Not given	Not given	Not given	62.3%	No
NASNet	Not given	Not given	Not given	43.1%	Not given	No
R-FCN	82.0%	Not given	Not given	53.2%	Not given	No



(a) IOU Curve of SSD512

(b) IOU Curve of SSD300

(c) IOU Curve of Yolo V3

Fig. 8 a IOU Curve of SSD512. **b** IOU Curve of SSD300. **c** IOU Curve of Yolo V3. Our proposed approach gives us the highest accuracy of 96.7%. **a**–**c** Shows the different curves implemented on SSD312, SSD300, and YOLO

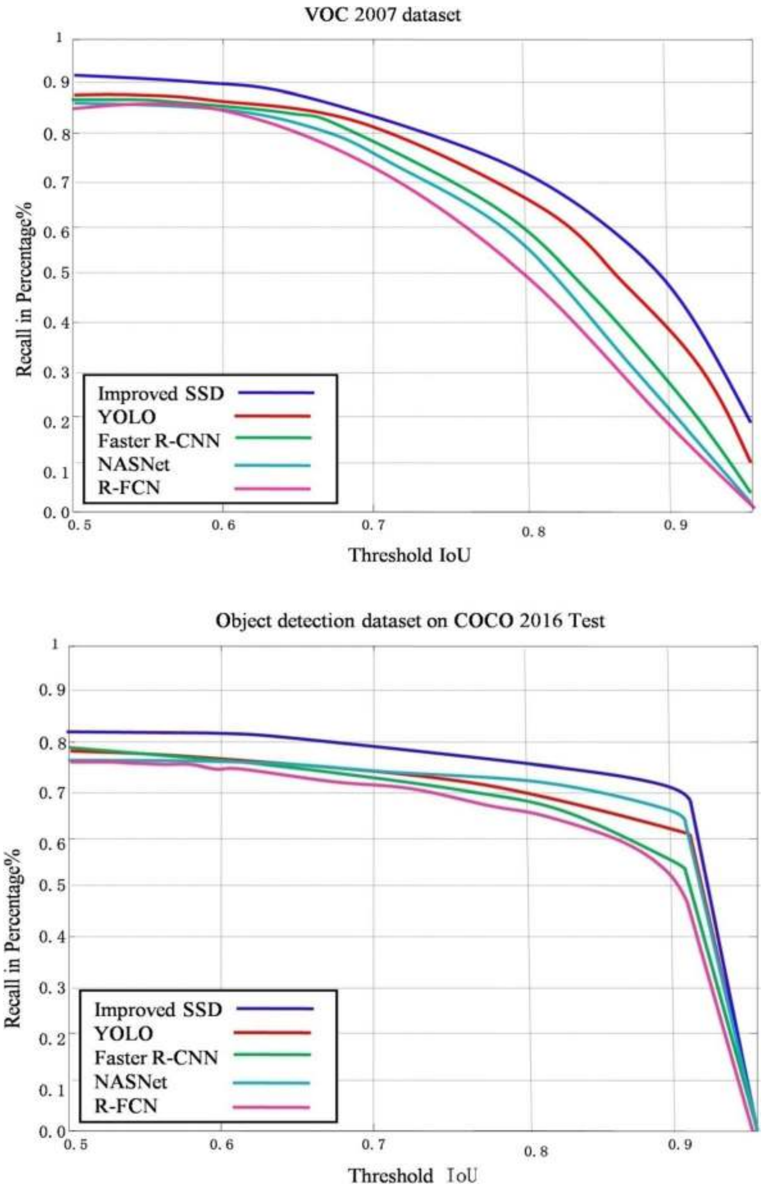


Fig. 9 Represent the different recall curves by using object detection algorithm on a particular threshold value of IoU. The proposed improved SSD approach has higher recall value, i.e., 0.9 compare with YOLO, faster R-CNN, NASNet, R-FCN as shown in **a**. **b** Demonstrates the graph of recall % vs threshold IoU compared with other object detection technique such YOLO, faster R-CNN, NASNet, and R-FCN on object detection (OD) dataset from our Centre for Image processing Lab. The recall value of improved SSD algorithm is 79.8% when the different value of IoU is applied

aspect ratio, and FPS with other previous models, which indicates that the proposed algorithm achieves a higher mAP, uses more frames to gain good speed, and obtains acceptable accuracy for detecting objects from color images. This paper points out that the algorithm uses truth box to extract feature maps. Future research can extend our proposed algorithm by training the datasets for micro-objects.

Abbreviations

CNN: Convolution neural network; F-CNN: Faster convolutional neural networks; F-RCNN: Faster region convolutional neural networks; SSD: Single shot multi-box detector; mAP: Mean average precision; FPS: Frames per second;

IoT: Internet of Things; VR: Virtual reality; YOLO: You Only Look Once; 5G: Fifth generation; OD: Object detection; VOC: Visual object classes; COCO: Common objects in context; GPU: Graphics processing unit; FPN: Feature Pyramid Network; API: Application program interface; ILSVRC: ImageNet Large Scale Visual Recognition Challenge; RPN: Region Proposal Network; DSSD: Deconvolutional Single Shot Detector; OpenCV: Open source computer vision; F(m): Feature maps; IoU: Intersection over Union; VGG: Visual Geometry Group

Authors' contributions

AK conceived of the study and carried out the experiment. HL and ZZ conducted the literature view and helped motivate the research. AK drafted the initial manuscript and all authors participated in further improving the manuscript. All authors read and approved the final manuscript.

Funding

The paper declares no funding support.

Availability of data and materials

The research uses the Pascal VOC and COCO object detection (OD) datasets from the center for image processing lab at Vardhaman College of Engineering. Datasets are available upon request.

Competing interest

The authors declare that they have no competing interests.

Author details

[1]Department of Computer Science & Engineering, Vardhaman College of Engineering, Hyderabad, India. [2]Coggin College of Business, University of North Florida, Jacksonville, FL 32224, USA. [3]Logistics and E-Commerce School, Zhejiang Wanli University, Ningbo, Zhejiang 315100, China.

Received: 23 December 2019 Accepted: 1 October 2020
Published online: 17 October 2020

References

1. Y. Zhong, Y. Yang, X. Zhu, E. Dutkiewicz, Z. Zhou, T. Jiang, Device-free sensing for personnel detection in a foliage environment. IEEE Geoscience and Remote Sensing Letters **14**(6), 921–925 (2017). https://doi.org/10.1109/LGRS.2017.2687938
2. S.Z. Su, S.Z. Li, S.Y. Chen, G.R. Cai, Y.D. Wu, A survey on pedestrian detection. DianziXuebao **40**(4), 814–820 (2012). https://doi.org/10.3969/j.issn.0372-2112.2012.04.031
3. M. Zeng, J. Li, Z. Peng, The design of top-hat morphological filter and application to infrared target detection. Infrared Physics & Technology **48**(1), 67–76 (2006). https://doi.org/10.1016/j.infrared.2005.04.006
4. L. Deng, D. Yu, Deep learning: methods and applications. Foundations and Trends® in. Signal Processing **7**(3–4), 197–387 (2014). https://doi.org/10.1561/2000000039
5. J. Redmon, S.Divvala, R.Girshick, A. Farhadi, 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). https://doi.org/10.1109/cvpr.2016.91.
6. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105)(2012). https://doi.org/10.1145/3065386.
7. H. Jiang, E. Learned-Miller, Face detection with the faster R-CNN. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017) (pp. 650-657)(2017, May). IEEE.. https://doi.org/10.1109/fg.2017.82
8. R. Girshick, J. Donahue, T. Darrell, J. Malik, 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). https://doi.org/10.1109/cvpr.2014.81.
9. X. Peng, C. Schmid, Multi-region two-stream R-CNN for action detection. In European conference on computer vision (pp. 744-759). Springer, Cham. (2016, October).https://doi.org/10.1007/978-3-319-46493-0_45.
10. J. Redmon, A.Angelova, Real-time grasp detection using convolutional neural networks. In 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1316-1322). IEEE(2015, May). https://doi.org/10.1109/ICRA.2015.7139361.
11. R. Girshick, Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448)(2015). https://doi.org/10.1109/iccv.2015.169.
12. S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99) (2015). https://doi.org/10.1109/tpami.2016.2577031.
13. Y.J. Cao, G.M. Xu, G.C. Shi, Low altitude armored target detection based on rotation invariant faster R-CNN[J]. Laser & Optoelectronics Progress, 55(10): 101501(2018). https://doi.org/10.3788/LOP55.101501.
14. J. Dai, Y. Li, K. He, J. Sun, R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems (pp. 379-387)(2016).
15. T.Y. Lin, P.Dollár, R.Girshick, K. He, B. Hariharan, S.Belongie, Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125)(2017). https://doi.org/10.1109/cvpr.2017.106.
16. J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7263-7271)(2017). https://doi.org/10.1109/cvpr.2017.690.
17. S. Ioffe, C.Szegedy,Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.(2015).
18. C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-First AAAI Conference on Artificial Intelligence. (2017, February)
19. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, A.C. Berg, Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37)(2016, October). Springer. Cham.. https://doi.org/10.1007/978-3-319-46448-0_2

20. Y.M. Wei, J.C. Quan, Y.Q.Y. Hou, Aerial image location of unmanned aerial vehicle based on YOLO V2[J]. Laser & Optoelectronics Progress, 54(11): 111002(2017). DOI:https://doi.org/10.3788/LOP54.111002.
21. J. Redmon, A. Farhadi, Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.(2018).
22. C.Y. Fu, W. Liu, A.Ranga, A. Tyagi, AC Berg, Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659.(2017).
23. Z. Li, F. Zhou, FSSD: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960.(2017).
24. J. Jeong, H. Park, N. Kwak, Enhancement of SSD by concatenating feature maps for object detection. arXiv preprint arXiv:1705.09587.(2017).
25. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778)(2016). https://doi.org/10.1109/cvpr.2016.90.
26. W. Liu, ARabinovich, AC Berg, Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579.(2015).
27. J.Q. Wang,J.S. Li,X.W. Zhou,X. Zhang, Improved SSD algorithm and its performance analysis of small target detection in remote sensing images[J]. Acta Optica Sinica, 39(6): 0628005(2019).https://doi.org/10.3788/AOS201939.0628005.
28. T.Y. Lin, P. Goyal, R.Girshick, K. He, P. Dollár, Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988)(2017). https://doi.org/10.1109/iccv.2017.324.
29. Y. LeCun, Y.Bengio, G. Hinton,Deep learning. nature 521(2015). https://doi.org/10.1038/nature14539.
30. A. Kumar, S.P.Ghrera, V. Tyagi, An ID-based secure and flexible buyer-seller watermarking protocol for copyright protection. Pertanika Journal of Science & Technology, 25(1)(2017).
31. A. Kumar, Design of secure image fusion technique using cloud for privacy-preserving and copyright protection. International Journal of Cloud Applications and Computing (IJCAC), 9(3), 22-36(2019). https://doi.org/10.4018/IJCAC.2019070102.
32. A. Kumar, S. S. S. S. Reddy and V. Kulkarni, "An object detection technique for blind people in real-time using deep neural network," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, pp. 292-297.
33. Szegedy, S. Reed, D. Erhan, D.Anguelov, S.Ioffe, Scalable, high-quality object detection. arXiv preprint arXiv:1412.1441.(2014).
34. V. Thakar, W. Ahmed, M.M.Soltani, J.Y. Yu, Ensemble-based adaptive single-shot multi-box detector. In 2018 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6) (2018, June). IEEE. https://doi.org/10.1109/ISNCC.2018.8530893.
35. P. Saimadhu, How The Logistic Regression Model Works, https://dataaspirant.com/2017/03/02/how-logistic-regression-model-works/, Accessed 2 March, 2017.
36. Jiang, R. Luo, J. Mao, T. Xiao, Y. Jiang, Acquisition of localization confidence for accurate object detection. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 784-799)(2018). https://doi.org/10.1007/978-3-030-01264-9_48.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.