Object Detection: A Comprehensive Review of the State-of-the-Art Methods

Akhil Kumar and Arvind Kalia
Himachal Pradesh University, Shimla, India
and
Akashdeep Sharma
Panjab University, Chandigarh, India

The process of localizing and classifying an object in a given sequence of images by computer vision systems is known as Object Detection. The work presented in the area of object detection is categorized into two broad categories. First category of work is based on traditional methods that deals with detection of an object in a single image having no or fewer deformations. The second category of work is based on evolutionary methods that deals with detection of multiple objects in a given image or a sequence of images having deformations. The Evolutionary methods of object detection addresses many core issues like fast detection, multi-view, multi-resolution, object part relation and deformations due to moving object and background. In this work, authors have presented a survey of the state-of-the-art methods of object detection. The object detection methods surveyed in this paper are Histogram of Oriented Gradients based Features, family of Region Proposal based Convolutional Neural Networks, Spatial Pyramid Pooling Network, family of You Only Look Once and Single Shot Detector. This work discusses the methods, training and evaluation aspects of evolutionary object detection methods based on Convolutional Neural Networks and Deep Learning. At the end, open research issues of object detection area are discussed.

Keywords: State-of-the-Art, Review, Object Detection, Research Issues.

1. INTRODUCTION

The key ability of computer vision systems is to perform object detection. Computer vision systems perform object detection by looking at the features of the object under consideration in the image sequence and the video. Several works in this area are presented in the past and new methods with high detection rate are proposed. The area of object detection can be classified into three categories. The first category of object detection methods is known as Traditional Methods. The traditional methods perform object detection by looking at the shape, color, texture and contours of the objects in the image sequence. Many works based on traditional methods are proposed in [Fischler and Elschlager 1973; Faloutsos et al. 1994; Vinod and Murase 1997; Grove et al. 1998; Jain et al. 1996]. The biggest drawback of the traditional methods is that they cannot perform classification and detection on multi-class datasets and where the object is moving or there are occlusions in the images. The traditional methods are not capable of doing detection in the images where the object is moving or the parts of the objects are deformed. In general, traditional methods are suitable for very small datasets and where very low computation cost is required.

The second category of object detection methods is based on feature learning and performing detection task on relatively big datasets and in the images where there are very few objects. The second category of object detection is also known as Intermediate Approach. An object detection method based on this category is proposed in groundbreaking work in [Dalal and Triggs 2005]. The work proposed is based on feature descriptors that learn characteristics of the objects by their shape. Local appearance of the objects is determined by edge directions and local intensity gradients. However, there are many drawbacks in this method. This method is only suitable for static imagery and not suitable for large datasets. The advantage of this method is that it is very fast, computation cost is low and predicts objects accurately with very low false positive rate. This method predicts with 99% accuracy on MIT Pedestrian dataset.[1]

The third category of object detection method is known as Evolutionary Approach. These methods are based on convolutional neural network. The convolutional neural networks are used to map the features and further employed to do the task of classification. These methods are very powerful, address many of the core issues of traditional and intermediate methods. The advancement in the area of evolutionary based methods has achieved the rate of 59 fps detection i.e. near to human eye visualization. In recent works [Girshick et al. 2014; He et al. 2014; Girshick 2015; Ren et al. 2015; He et al. 2017; Redmon et al. 2016; Liu et al. 2016], the Region Proposal based Convolutional Networks (R-CNN), Fast-RCNN, Faster-RCNN, Mask RCNN, Spatial Pooling Pyramid Networks, You Only Look Once and Single Shot Detector are few proposed methods based on convolutional neural networks. These methods have many advantages over traditional methods. These methods perform detection in multi-class datasets, in images with occlusions, where the background sequence is changing and in large datasets. The method such as Region Proposal based convolutional neural network has mean average precision of 62 on various object categories in PASCAL VOC 2011 database.[2] Successors of R-CNN, Fast-RCNN has mean average precision of 66 on PASCAL VOC 2012 database and Faster-RCNN has mean average precision of 75.9 on PASCAL VOC 2012 database. Single shot detector is the fastest among all the detectors. It detects at 59 fps and has mean average precision of 74 on PASCAL VOC 2012 and COCO database[3]. The drawback of these methods is that they are computationally expensive, prone to localization errors due to fast speed and specialized systems based on GPUs are required to train and evaluate these methods. However, there are many issues that are not addressed by any of the approaches. The issues are Active Vision, i.e., learning new classes of objects from the environment by self by the detectors, Multi-Modal detections, i.e., performing detections on images having objects at varying depths of the image, predicting by establishing relationship between object and its parts and localizing small objects in the images.

Figure. 1: Original image

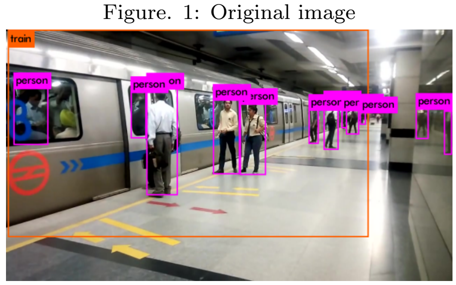


Figure. 2: Image after object detection

Figure 1 is the original image captured by a surveillance camera. Humans can identify that

[1]MIT CBCL: `http://cbcl.mit.edu/software-datasets/PedestrianData.html`
[2]PASCAL VOC: `http://host.robots.ox.ac.uk/pascal/VOC/`
[3]COCO Common Objects in Context: `http://cocodataset.org/#home`

the image is containing train and persons. Figure 2 is the image showing detection results for Figure 1 when passed through an object detection algorithm. The object detection algorithm localizes and classifies the objects present in the original image.

2. A BRIEF REVIEW OF OBJECT DETECTION RESEARCH

The early methods of object detection are based on features i.e. shape, color, contour and texture of the object under consideration. Several works based on these characteristics are proposed by Fischler and Elschlager [1973], Faloutsos et al. [1994], Vinod and Murase [1997], Grove et al. [1998], and Jain et al. [1996]. All these methods are able to perform detection task on single-class of object. These methods are not able to perform detection task on multi-class objects and if the object under consideration is prone to parts deformity due to moving background or high velocity movement of the object in a scene. Later, few works based on contextual information, deformation information and velocity information of object under consideration are proposed by Heikkila and Pietikainen [2006], Kass et al. [1988], Caselles and Coll [1996], and Wixson [2000].

With the advent of machine learning, growing information and specialized learning methods, new methods based on Sliding Window and Gradient Based Learning are introduced by Glumov et al. [1995] and LeCun et al. [1999]. In sliding window based method, a classifier is developed on the basis of an exhaustive search applied on a given image. The search is applied at different locations and scales of the image to recognize the features of the object. The learned feature by the classifier differentiates the object from the image. In alternate to this method, a method based on Bag-of-Words is proposed by Tsai [2012]. In this method, to verify the object in an image the image area is iteratively refined. This iteration process differentiates the object from the image. In gradient based learning method, the features of the object are represented and the represented features are learned by the neural network based classifiers. The classifier performs an exhaustive search for the learned features on a given sequence of images and perform detection tasks by matching learned features with the image features. If the learned feature representation is matched with the features of the image sequence then the detection is considered successful. The matching performed in this method is also based on sliding window. A summary of literature of object detection research is presented in Table I.

The literature in the area of object detection is primarily based on traditional techniques, i.e., performing object detection task on static imagery or in the images having very few deformations. The disadvantage of traditional techniques is their applicability on small datasets. With the advent of specialized systems with GPUs and growing size of image datasets new methods based on convolutional neural networks are proposed in recent years. In this work authors have covered all the evolutionary methods based state-of-the-art object detection methods.

The present day object detection methods are based on convolutional neural networks. The method of convolutional neural networks (CNNs) is proposed by Fukushima [1980] and LeCun et al. [1999]. The basic idea behind CNNs is neural networks. Like neural network, CNNs are made up of neurons with learnable weights and biases. Each neuron of the network receives several inputs, takes a sum over the weights, passes them through an activation function and finally responds with an output. The difference between CNNs and Neural Networks is that former function on volumes. The input in CNNs is a multi-channel image.

In CNNs as shown in Figure 3, an input image is represented as a matrix of pixels. The input image matrix is passed to the convolutional layers. The purpose of convolutional layers is to extract the features from the input image and pass the feature matrix to the pooling layer. The process applied by convolutional layers is known as Convolution. Next, Pooling operation introduced by [Ciresan et al. 2011] is applied to the extracted features provided by the convolutional layers. The purpose of pooling is to reduce the dimension of the feature matrix provided by the convolutional layer but to retain the most important information. Several pooling methods such as Max pool and Average pool are applied dependent on the type of information required. In general, Max pooling performs better. At last, Flattening is applied to the information matrix

Sr. No.	Paper Title	Author with Year	Method/ Technique Used
1	Vehicle Detection and Tracking Techniques: A Concise Review	Hadi et al. [2014]	Background subtraction, Feature based, Frame differencing and Motion based methods, Region, Contour, 3-D Model, Feature, Color and Pattern based tracking methods
2	Moving Object Detection: Review of Recent Research Trends	Kulchandani and Dangarwala [2015]	Background subtraction, Frame differencing, Optical flows and Temporal differencing based methods
3	Research of Object Recognition and Tracking Based on Feature Matching	Ahn and Rhee [2015]	SURF and SIFT
4	Object Detection: Current and Future Directions	Verschae and Ruiz-del Solar [2015]	Coarse to fine and boosted classifiers, Dictionary based, Deformable part based model, Deep learning and Trainable image processing architecture
5	A Survey on Object Detection in Optical Remote Sensing Images	Cheng and Han [2016]	Template matching, Knowledge based method, OBIA based such as image segmentation and Machine learning based methods such as HOG, Haar like features, SVM, Adaboost, CRF, SRC and Artificial neural networks
6	A Review of Object Detection Based on Convolutional Neural Network	Zhiqiang and Jun [2017]	Sliding window, HOG, SIFT, SVM, Adaboost, Non max suppression, Combine boxes, R-CNN, Fast-RCNN and Faster R-CNN
7	A Review and An Approach for Object Detection in Images	Sharma and Thakur [2017]	Sliding window, Contour based, Graph based, Fuzzy based and Context based methods
8	Soft Computing Based Object Detection and Tracking Approaches: State-of-The-Art Survey	Kaushal et al. [2018]	Neural networks, Fuzzy logic, Evolutionary techniques such as Fuzzy classifier and Fuzzy Kalman filter, Hybrid approaches such as Particle Swarm Optimization, Genetic algorithm and Hybrid Neural Networks and Expert system based approaches
9	Object Recognition based on Surface Detection - A Review	Boruah et al. [2018]	Knowledge representation
10	A Critical Review of Object Detection using Convolution Neural Network	Nisa and Imran [2019]	Convolutional Neural Networks, AlexNet and SVM

Table I: Object detection research summary

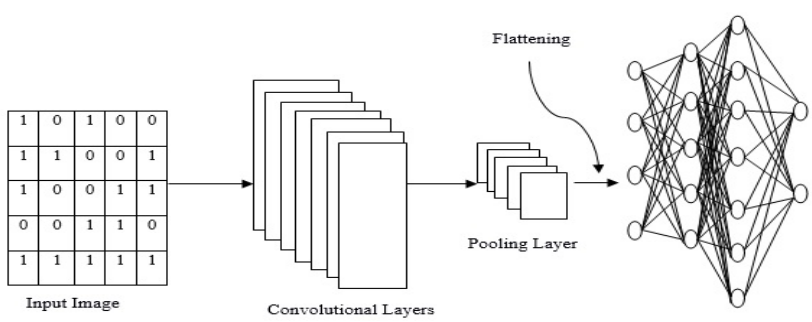


Figure. 3: A CNN for image classification model

provided by the pooling layer. The flattening layer converts the matrix provided by the pooling layer into a linear array. This linear array is fitted as input to nodes of the neural network. Furthermore, there are many other layers like Sigmoid function and Softmax function dependent on the type of classification. For classifying a binary class dataset sigmoid function is applied to make the full network. In case, if there are more than two classes for classification then Softmax function is applied to the network. The Softmax function is introduced in [Bridle 1990a] and [Bridle 1990b]. The scheme for detection adopted by object detection methods is depicted in Figure 4.

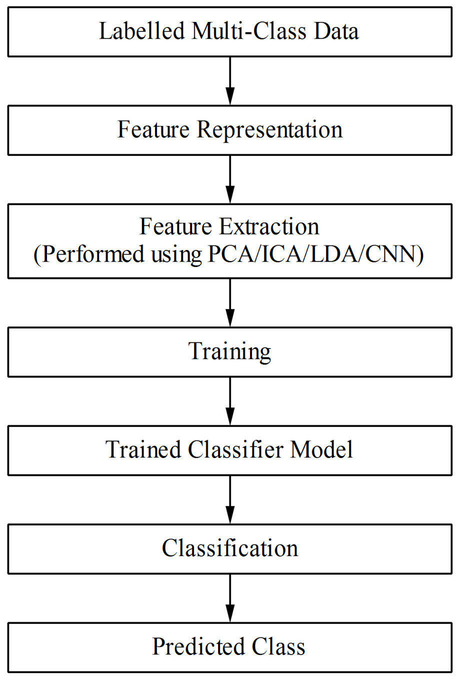


Figure. 4: General object detection strategy

3. THE STATE-OF-THE-ART OBJECT DETECTION METHODS

Object Detection methods can be broadly classified into three categories namely Traditional Approach, Intermediate Approach and Evolutionary Approach. The traditional approach of object detection is based on Feature, Template and Motion Information of the object under consideration. In this work, we have discussed the Intermediate and Evolutionary Approach of object detection. The intermediate approach of object detection is based on Support Vector Machine based classifiers. In groundbreaking work by Dalal and Triggs [2005], authors have proposed a Histogram of Oriented Gradient features based detection method for Pedestrian Detection. The evolutionary approaches of object detection are based on convolutional neural network based classifiers. As proposed in [Liu et al. 2016], the detector based on evolutionary approach can do detection at 59 fps. The evolutionary approach is classified in two categories- i). Region proposal based and ii). Classification based object detection methods. The solutions proposed to solve the problem of object detection are illustrated in Figure 5.

3.1 Histogram of Oriented Gradients

This method is proposed in groundbreaking work in the area of object detection in [Dalal and Triggs 2005]. This method is feature descriptor based, that characterize objects on the basis of shape. To identify objects local appearance edge directions and local intensity gradient is used.

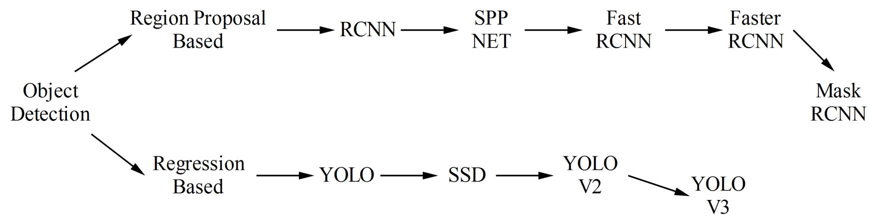


Figure. 5: Object detection problem and solution progression

3.1.1 *Method*

(1) In the first step, the image is divided into blocks. The block can be of size 16×16 pixels. The block is further divided into cells, i.e., a block of 16×16 pixels is divided into cells of 8×8 pixels. There can be several cells in a block. For these cells, at pixel level vertical and horizontal gradients are obtained. This is achieved by applying 1-D Sobel method proposed in [Gonzalez et al. 2004].

$$G_x\,(y,x) = Y(y,x+1) - Y(y,x-1) \tag{1}$$

$$G_y\,(y,x) = Y(y+1,x) - Y(y-1,x) \tag{2}$$

where $Y(y,x)$: Pixel intensity and coordinate x and y, $G_x\,(y,x)$: Horizontal gradients, and $G_y\,(y,x)$: Vertical gradients.

Next, magnitude and phase of the gradients are obtained using equation (i)

$$G\,(y,x) = \sqrt{G_x(y,x)^2 + G_y(y,x)^2}, \theta(y,x) = \arctan\left(\frac{Gy(y,x)}{Gx(y,x)}\right) \tag{3}$$

(2) In this step, for each cell histogram of gradients is computed. To get the histogram, for each angle Q bins are selected. The angle has unsigned orientation and due to this all angles below $0°$ are increased by $180°$.

(3) In this step contrast normalization is applied to the images as different images may have varying contrast level. In a single block obtained at step (1), normalization is applied on a histogram with vector v. The norm used is-

$$\text{L1-norm: } f = \frac{v}{(||v||_1 + e)} \tag{4}$$

$$\text{L2-norm: } f = \frac{v}{(||v||_2^2 + e^2)} \tag{5}$$

$$\text{L1-sqrt: } f = \sqrt{\frac{v}{(||v||_1 + e)}} \tag{6}$$

(4) In this step, to each detector window a descriptor is applied. For each detector window, the descriptor is constituted of all the histogram for all the cells of a block falling in that window. The descriptor obtained is used as feature information for recognition task and to perform training on the data.

(5) In this step, a linear Support Vector Machine based classifier is applied to classify the categories of the objects.

The method of HOG Features in presented in Figure 6. This method is originally tested on MIT Pedestrian detection dataset and perform detection task on static imagery. This method performs classification using Support Vector Machine based classifier but is not only tied to this method. The task of classification can be done with other machine learning algorithms once the gradients are computed and feature representation is obtained.

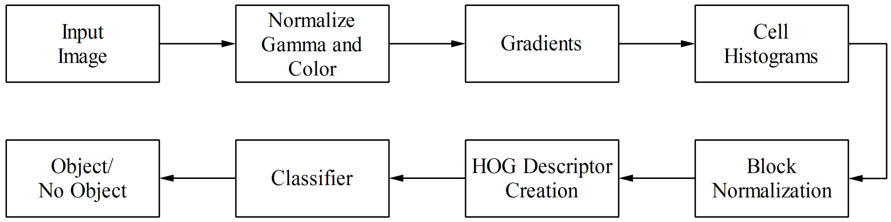


Figure. 6. Histogram of Oriented Gradient Features method

3.1.2 *Advantages*

—This method is computationally inexpensive.

—On MIT pedestrian dataset, the descriptors produced a detection miss rate of essentially zero at a 10-4 false positive rate. Hence, it is very accurate.

3.1.3 *Disadvantages*

—Not suitable for large dataset.

—Does detection for static imagery. Thus, not suitable for detection in videos.

3.2 Region Proposal Based CNN

This method is convolutional neural network based and functions on region proposals. An image can have large number of regions therefore, it is difficult and expensive to process each and every region. This method employs a different intuitive strategy. Instead of looking on large number of regions in an image this method looks for selective regions in the image to locate the object. This method uses selective search to extract region containing the object from other regions. This method is proposed by Girshick et al. [2014].

This method functions in two steps. In first step, Region proposals are generated using selective search and in second step, a convolutional neural network is trained to perform the task of object detection. The detailed pipeline of R-CNN is shown in Figure 7.

3.2.1 *Method.*

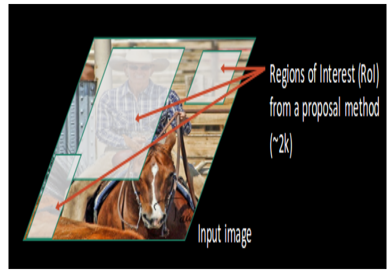
3.2.1.1 *Region Proposal Using Selective Search*

(1) Take the arbitrary size input image.
(2) Segmentation is applied to the input image so that multiple regions can be generated for the image.
(3) Based on color, texture, size similarity and shape compatibility several small regions are taken together to form a large region.
(4) Finally, from the large regions obtained in step (3), regions of interest are identified where the object detection is to be performed.

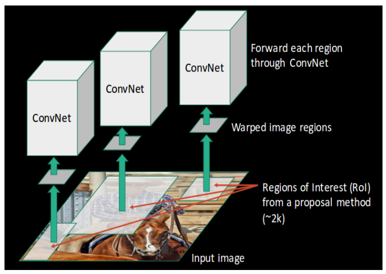
3.2.1.2 *Object Detection Using CNN*

(1) Take a pre-trained convolutional neural network model.
(2) Re-train the model. The last layer of the model is trained with number of classes that are to be detected.
(3) For each image, collect the region of interest. Reshape the region of interest to fit into the CNN model input.
(4) In this step, a Support Vector Machine based classifier is trained to classify the image into object and background. A binary Support Vector Machine is trained for each class.
(5) In this step, tight bounding box is applied across the images. This is performed by training a linear regression that classifies each category of object in the image.

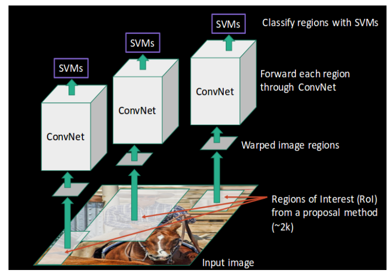
(a) Input image



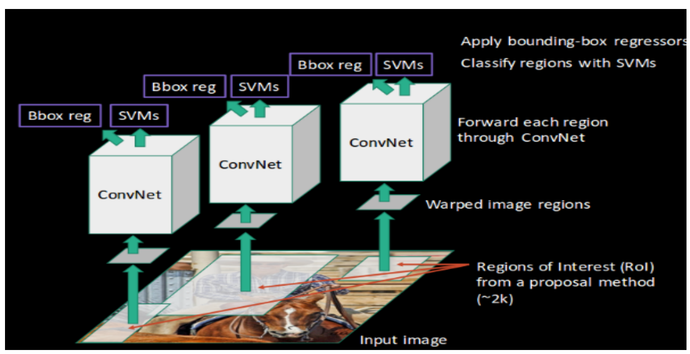
(b) Collect region proposal



(c) Fit Regions to CNN input



(d) Apply SVM to train the model



(e) Classification of each category of object after applying the tight bounding boxes

Figure. 7: R-CNN idea

3.2.2 *Advantages*

—This method allows detection of background objects in the image.

—It is less prone to localization errors, as only the region proposals are searched for presence of the object.

3.2.3 *Disadvantages*

—Based on selective search, a total of 2000 region proposal are extracted for each image.

—For each region, features are extracted using CNN Model. This is a computationally expensive task. For N images, $N * 2000$ CNN features will be calculated.

—The process of detection is a long process in this method. Firstly, the features are extracted for CNN and then a linear Support Vector Machine classifier is applied to identify the object. Next, for tightening the bounding box, a regression model is applied.

—RCNN takes approximately 40 seconds to detect an object in the image thus, it is very slow.

3.3 Spatial Pooling Pyramid Network

The Spatial Pyramid Pooling Network is a method that allows us to handle multi-scale images efficiently to perform the task of classification. This method is similar to Bag-of-Words method. This method scales up the performance of convolutional neural networks. This method is employed in visual recognition tasks. This method comprises of convolution layers and spatial pooling layers. The convolution layers perform task of extracting the feature maps and the spatial pooling layer standardize the output produced by convolution layer and further classify the output classes by passing the output through fully connected layers and Support Vector Machine or Softmax Layer based classifiers. Figures 8 and 9 show the detailed operation of Spatial Pyramid Pooling Network.

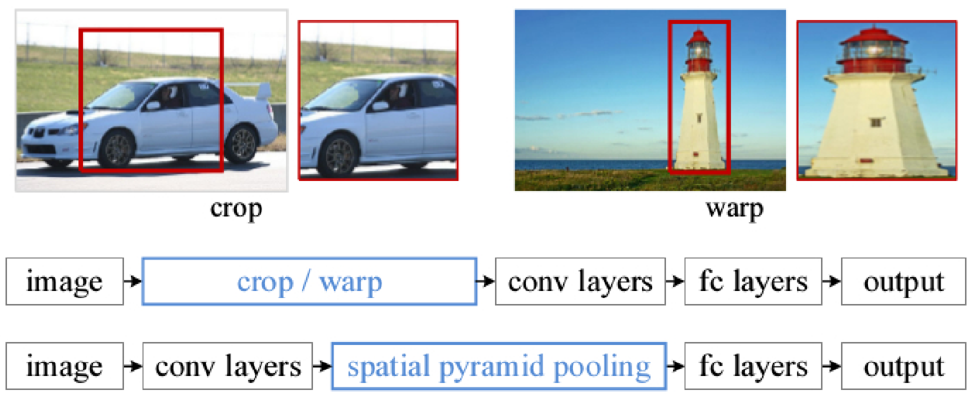


Figure. 8. Spatial Pyramid Pooling idea

3.3.1 *Method.*

3.3.1.1 *Convolution Layer*

(1) Take an arbitrary size input image.
(2) Pass the entire image to the convolution layer.
(3) Selective search is applied by the convolution layers to extract the feature maps from each region of the image.
(4) After passing through convolution layers, independent features for each region are computed by the pooling operation. This is done once the feature maps for each region is extracted by the selective search operation.

3.3.1.2 *Spatial Pooling Layer.* Since, arbitrary size images are taken by convolution layers but the output produced by them is of variable size. The standardization of variable size output is done by spatial pooling layer.

(1) The variable size output produced by convolution layers is passed to the spatial pooling layer.
(2) The spatial pooling layer applies an improved Bag-of-Words proposed by Tsai [2012] like method to standardize the variable size output provided by the convolution layer to fixed size vectors.
(3) The improved Bag-of Words approach maintains the spatial information by pooling in local bins.
(4) Next, the network is trained and classification of output is performed using regression layer by SVM based classifiers.

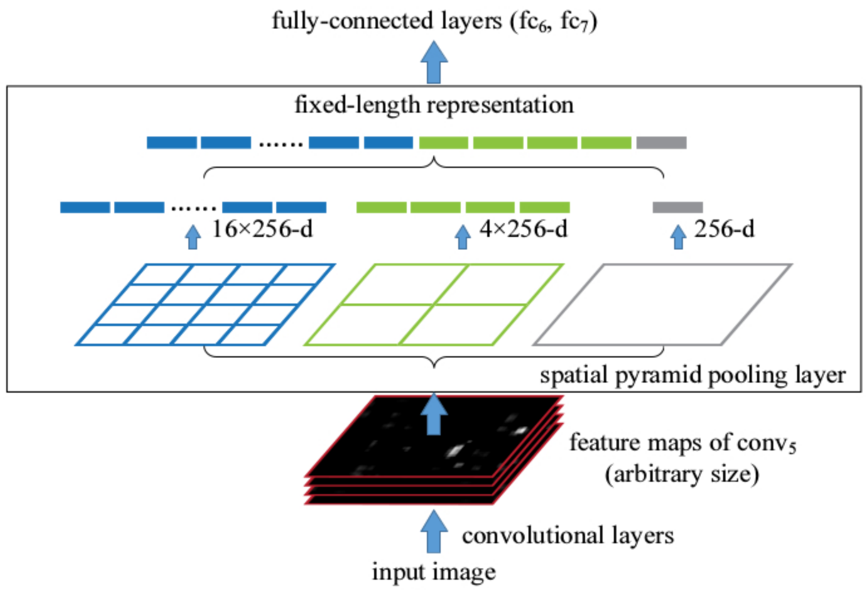


Figure. 9. Pooling Layer idea

3.3.2 *Advantages*

—Computationally less expensive than R-CNN. It performs the task of feature collection by passing the region proposals to convolution layers.

—Perform detection not only on arbitrary size input images but multi-scale images also.

3.3.3 *Disadvantages*

—The training process in SPP Net is not end-to-end. Feature collection is done by convolution layers, the spatial pooling layer maintains the spatial information bins and the classification is done by regression layer. Thus, it is a lengthy process.

—Not suitable for real-time detection.

3.4 Fast R-CNN

This method is proposed by Girshick [2015] and is extension to Region proposal CNN. In R-CNN, a CNN is run 2000 times to extract proposals per image. This makes R-CNN computationally expensive. To reduce this computational expensiveness, authors proposed the method of extracting only proposals by running CNN only once.

To make R-CNN fast, in this method the CNN runs only once to extract the proposals from one image and then share the computation across the 2000 regions. In this method, the input image is fed to the CNN and in turn CNN generates the convolutional feature maps. From these feature maps, the region proposals are extracted. Next, using Region of Interest (ROI) pooling layer, all proposed regions are reshaped to fixed size and fed to the fully connected network. The detailed operations of F-RCNN is presented in Figure 10.

3.4.1 *Method*

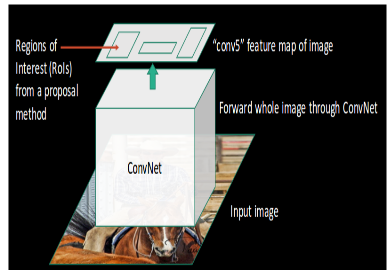
(1) Take input image of arbitrary size.

(2) The image is fed to a convolutional neural network to generate regions of interest.

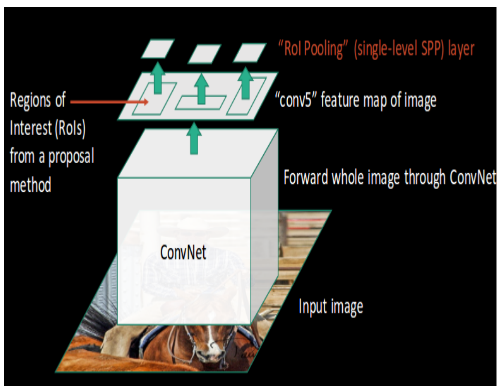
(3) To all the regions of interests, a Region of Interest pooling layer is applied to reshape. Next, the reshaped regions are passed through a fully connected network.

(4) On top of the fully connected network, a Softmax Layer is applied to classify the object categories. In parallel to Softmax Layer, the output bounding box coordinates a linear regression layer is applied to predict the output classes.

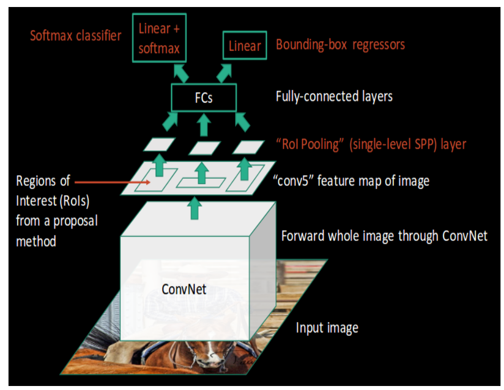
(a) Input Image



(b) Generate region of interest



(c) ROI layer applied to reshape



(d) Image classification

Figure. 10: Fast RCNN idea

3.4.2 *Advantages*

—In R-CNN, 2000 proposals are fed to the CNN. This makes R-CNN computationally expensive. In Fast-RCNN, only one proposal is fed to the CNN to generate the feature map.

—Only one model is employed to extract feature map, classification and generating bounding boxes for the output classes.

3.4.3 *Disadvantages*

—This method also employs the selective search method to extract the regions of interest which is a time-consuming process.

—F-RCNN takes approximately 2 seconds to detect an object in the image. Thus, it is not suitable for real-time detection.

3.5 Faster RCNN

This method is introduced by Ren et al. [2015] and basically it is an extension to the Fast-RCNN method. This method overcomes the issue of slow detection by replacing the selective search method of extracting the region proposals in Fast-RCNN by a Region Proposal Network. The Region Proposal Network is used for extracting the image feature maps and to generate the object proposals. Each object proposal is assigned an objectness score as output. Faster R-CNN idea is presented in Figure 11.

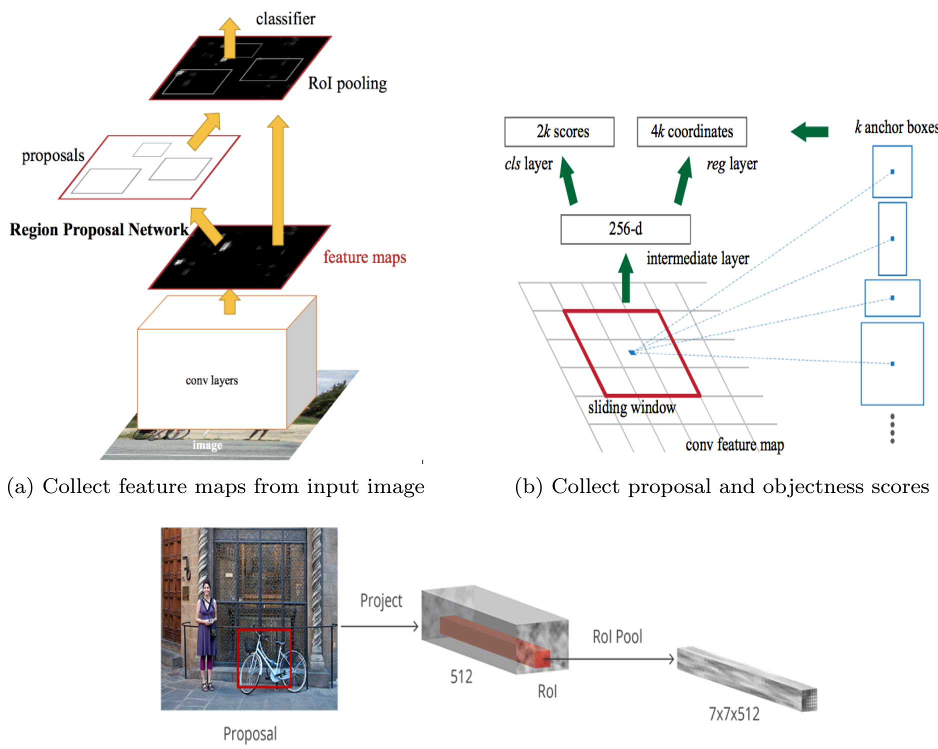
3.5.1 *Method*

(1) An input image is passed to the convolutional neural network to obtain the feature map.

(2) On the feature map Region Proposal Network is applied. In return, Region Proposal Network provide region proposals with their objectness score.

(3) To bring all region proposals to same size, Regions of Interest Pooling layer is applied to the region proposals.

(4) In last, to the proposals passed through the covolutional network, a Softmax Layer is applied and at top of it a Regression Layer is applied to classify the objects along with their bounding boxes.



(a) Collect feature maps from input image

(b) Collect proposal and objectness scores

(c) Apply ROI Pooling layer to bring all proposals on same level

Figure. 11: Faster RCNN idea

3.5.2 *Advantages*

—Much faster than R-CNN. It replaces selective search by Region Proposal Networks. It makes this method computationally less expensive than its predecessors R-CNN and F-RCNN.

—Faster-RCNN take approximately 0.2 seconds to perform detection on an image thus, it is very fast.

3.5.3 *Disadvantage*

—In this method multiple layers are functioning one after another to generate the feature maps, region proposals, bounding boxes and to perform classification. Due to this, object proposal generation takes time and the performance of next layers is dependent on the previous layers.

3.6 Mask RCNN

This method is an extension to Faster-RCNN and is proposed by [He et al. 2017]. The method proposed is a simple and flexible framework for object instance segmentation. This method extends features of Faster-RCNN and in parallel to Faster-RCNN functions; this method performs prediction for object mask. The image mask obtained is used to do the prediction of a class at pixel level. Mask RCNN perform detection at 5 frames-per-second. This method allows estimating human poses in images. The operation of Mask R-CNN is presented in Figure 12.

3.6.1 *Method.* Mask RCNN is a combination of Faster RCNN and Fully Convolutional Network. The steps involved in this method are-

3.6.1.1 *Faster-RCNN.* Faster-RCNN is employed on the image to obtain the class and bounding boxes. This step does the task of object detection. The steps of this method are discussed in section 3.5.

3.6.1.2 *Fully Convolutional Network.* In this step, Fully Convolutional Network is applied on the class and bounding boxes obtained in step 1 and the pixel wise boundary of the object classes is obtained. It is applied to perform semantic segmentation.

(1) Select an arbitrary size image.
(2) Using Convolution layers and Maxpool layers decompose the original image to its 1/32th size.
(3) In this step, on 1/32th size granule image class prediction is done.
(4) Lastly, using up sampling and deconvolution layers the granule 1/32th size image is reformed to the original size image.

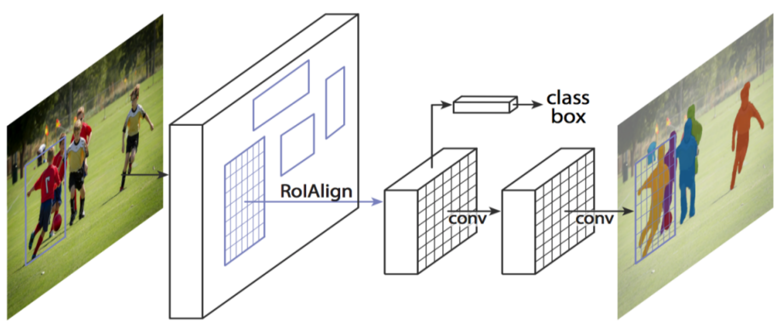


Figure. 12. Mask RCNN idea

In Image, ROI layer + first Convolution layer is used to extract Regions of interest, bounding boxes and the two Convolution layers does pixel wise boundary.

3.6.2 *Advantage*

—It performs detection at pixel level boundary of object classes. More accurate bounding boxes are generated due to accurate boundary of the objects in the image.

3.6.3 *Disadvantages*

—The detection task is computationally expensive due to two parallel methods, i.e., Faster-RCNN and Fully Convolutional Network running simultaneously.

—Detection rate is slow, i.e., 5 frame per second. Thus, it is not suitable for real-time detection.

3.7 You Only Look Once (YOLO)

This method is proposed by Redmon et al. [2016]. This method utilizes the complete top most feature map to predict bounding box scores and confidences for multiple categories. The idea behind YOLO is illustrated in image 13. YOLO is stated as a real-time object detector which performs detection at 45 fps and on PASCAL VOC dataset has a mean average precision of 63.4. This method is trained on COCO database object categories. The YOLO operation is illustrated in Figures 13 and 14.

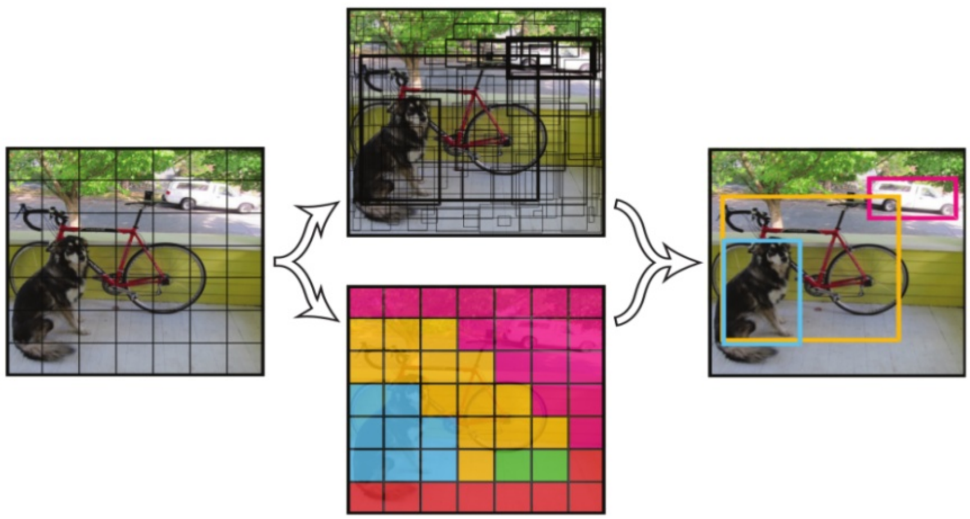


Figure. 13. YOLO idea

3.7.1 *Method*

(1) The input image is divided in $S \times S$ grid. Each grid is a cell responsible to predict the object centered in that grid cell.

(2) Each grid cell predicts B bounding box and their confidence score. Confidence scores are defined as $Pr(Object) * IOU_{pred}^{truth}$, confidence score indicates the likeliness of presence of object ($Pr(Object) \geq 0$) and shows confidence of its prediction, (IOU_{pred}^{truth}).

(3) In this step, in parallel to step (2), regardless of number of boxes, for each grid cell Conditional Class probability C as $\Pr(Class_i|Object)$ is also predicted. Contribution is calculated only for the grid cell containing the object.

(4) Next, individual box confidence prediction is multiplied with conditional class probabilities to determine the class-specific confidence scores for each box. This step is performed at test time as-

$$\Pr(Object) * IOU_{pred}^{truth} * \Pr\left(Class_i|Object\right) = \Pr\left(Class_i\right) * IOU_{pred}^{truth} \tag{7}$$

The existing class specific objects in the box probabilities, the fitness between the predicted box and the object are taken into consideration.

Following loss function is optimized at training time-

$$
\begin{aligned}
&\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] + \\
&\lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left[(\sqrt{w_i} - \sqrt{\widehat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{ij}^{obj} \left(C_i - \hat{C}_i \right)^2 + \\
&\lambda_{nobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{I}_{i}^{nobj} \left(C_i - \hat{C}_i \right)^2 + \sum_{i=0}^{S^2} \mathbb{I}_{i}^{obj} \sum_{c \,\in\, classes} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}
\tag{8}
$$

For a certain cell $i(x_i, y_i)$ denote the center of the box relative to the bounds of the grid cell. (w_i, h_i) is normalized width and height relative to the image size. C_i, represents confidence scores. I_i^{obj}, indicates the existence of objects. I_{ij}^{obj}, denotes that the prediction is conducted by the j^{th} bounding box predictor.

The loss function penalizes the classification errors only when there is presence of an object in that grid cell. Similarly, the bounding box coordinate errors are penalized when the predictor has achieved highest Intersection of Union (IOU) for the ground truth box.

The YOLO model is based on DarkNet model that has 24 convolution layers and 2 fully connected layers. Few convolution layers are 1×1 reduction layers and 3×3 convolution layers that construct ensemble of the inception module.

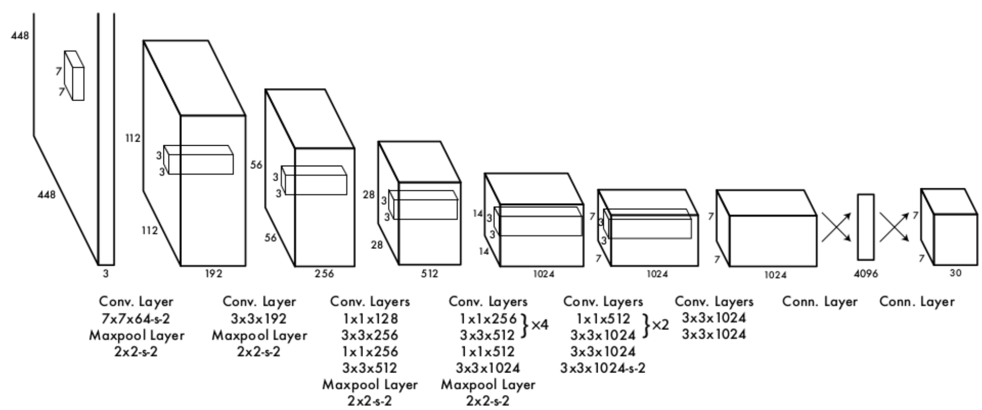


Figure. 14. YOLO architecture

3.7.2 *Advantages*

—Very fast. Does detection at 45 fps.

—Generalized object representation is understood by the network. This method predicts fairly well on artwork images.

—Faster version of this model is based on small architecture and perform detection at 155 fps.

3.7.3 *Disadvantages*

—Prone to localization errors.

—Struggle in detection of small objects.

3.8 YOLO V2

YOLO V2 is an extension of original YOLO discussed in section 3.7. This method is proposed by [Redmon et al. 2016]. In this method YOLO detects 9000 categories of objects using hierarchical classification with a 9418 node WordTree. In this method samples are combined from COCO database and 9000 object categories of ImageNet database[4]. For every COCO data, YOLO sample four ImageNet data. Detection data in COCO database is used for learning the objects and the classification is performed with ImageNet samples. YOLO 9000 evaluates its results from 200 categories of ImageNet object detection dataset. ImageNet share 44 categories with COCO. There are total 156 categories that are uncommon in COCO and ImageNet and YOLO learn feature map for those categories and perform detection task. On the 156 categories of objects, the YOLO V2 achieves mean average precision of 16.0. The overall mAP of YOLO V2 is 78.6 on VOC 2007 database.

3.8.1 *Method.* The working method of YOLO V2 is similar to the original YOLO discussed in section 3.7. However, few additions are proposed to the original method. The additions are discussed below.

(1) **Batch Normalization-** In YOLO V2, Batch Normalization is added to all convolution layers. This reduces the fitting and helps in regularizing the model
(2) **High resolution classifier-** Original YOLO takes input image of size 224×224. The YOLO 9000 takes input image of size 448×448, i.e., the doubled image resolution for training on ImageNet dataset.
(3) **Anchor Boxes-** In this method Anchor Boxes are introduced as they were in Region Proposal Networks and Faster-RCNN. This improves the Recall but reduce the accuracy. This leads to prediction of more bounding boxes per image. To calculate Anchor Boxes, this method uses k-means clustering.
(4) **Fined-Grain Features-** YOLO V2 predicts on feature map of size 13×13 which is smaller than original YOLO. This leads to detection of small objects accurately as well as large objects.
(5) **Multi Scale Training-** YOLO V2 can learn from varying scale images ranging between 320×320-608×608.
(6) **Feature Extractor-** This method employs DarkNet 19 as its backbone architecture for classification. This backbone architecture has 19 convolutional layers and 5 max pooling layers. For classification, a Softmax Layer is applied at top of the last convolutional layer.

3.8.2 *Advantages*

—Can learn new classes by doing generalization.
—Can detect small objects accurately.

3.8.3 *Disadvantage*

—Prone to localization errors but lesser than original YOLO.

3.9 YOLO V3

YOLO V3 is proposed as incremental improvement to its predecessors YOLO and YOLO V2 by [Redmon and Farhadi 2018]. The improvements proposed in this method scale up the mean average precision on COCO dataset to 57.9. YOLO V3 employs a single neural network to the full image and at test time predictions are made on global context of information present in the image. This method divides the image into regions and for each region predicts bounding boxes and class probabilities. The bounding boxes are weighted by the class probabilities.

[4]ImageNet: `http://image-net.org/`

3.9.1 *Method.* The method of YOLO V3 is similar to YOLO and YOLO V2 with few modifications proposed to improve training and increase performance. In this method a better backbone classifier is proposed along with predictions on multi-scale images. The modifications are discussed below.

Bounding Box Predictions- YOLO V3 like YOLO V2 employs Anchor Boxes to determine the image clusters. As YOLO V3 is a single network and from the same network the loss of objectiveness and classification is calculated separately. The objectiveness score by YOLOV3 is predicted by logistic regression where the complete overlap of bounding box over the the ground truth object is represented by 1. For one ground truth object only 1 bounding box is predicted. Both classification loss and detection loss will infer an error, if there is a change in the value of logistic regression. For other values than this best 1, only the error will incur in detection loss.

(1) **Class Predictions-** This method instead of using a Softmax Layer uses independent logistic classifiers for each class. This enables the method to do multi-class classification.
(2) **Predictions across scales-** Three different scales are employed to do detection at varying scales and accordingly YOLO V3 predicts the boxes. Like Pyramid Pooling as discussed in section 3.3 of this work, features are extracted from different scales.
(3) **Feature Extractor-** This method employs DarkNet 53, a backbone architecture for extracting the features. DarkNet 53 has 53 convolutional layers, residual and shortcut connections. YOLO V2 used DarkNet 19 as its backbone architecture for feature extraction. DarkNet 53 used in YOLO V3 is deeper than DarkNet 19 thus more features are extracted by YOLO V3.

3.9.2 *Advantages*

—Improved precision for small object detection.
—Less localization errors.
—Due to addition of feature pyramid method, the predictions for same objects increases significantly at varying scales.

3.9.3 *Disadvantage*

—Precision can be improved for medium and large objects.

3.10 Single Shot Detector

This method is proposed by [Liu et al. 2016] and it is the fastest among all the methods discussed above. This method works on the concept of bounding box and has replaced the concept of region proposals. In this method, pre-defined boxes look for the presence of objects. This method is based on a feed-forward convolutional neural network that integrates several systems into one. This method uses convolution layer to learn convolutional feature maps from the previous layer and run small convolution filters over the feature maps to predict the class scores and bounding boxes. The base network for this method is VGG-16. This method has 74% mean average precision on Pascal VOC 2012 and COCO object detection dataset. Due to its high speed of detection this method is suitable for embedded devices. The SSD operation is shown in Figure 15 and Figure 16.

3.10.1 *Method*

(1) First, a Convolution Neural Network is trained with bounding box and classification object. Bounding box is the regression function and classification objective is the loss function.
In this step, a Fully Connected (FC) layer or a Convolution layer that act as Fully Connected layer is applied to gather activation from layers to infer classification and location. The convolution layer produces the final object classes by passing the input image from a fixed-size collection of bounding boxes and the object class presence in bounding box is measured using bounding box scores.

(2) In order to classify the object and filter the multiple bounding boxes around the same object Non-Max suppression is applied. The final output classes are produced after applying non-max suppression on the bounding box scores. The non-max suppression hides the bounding boxes with low scores and highlights only the classes in the bounding boxes with maximum bounding box score.

(3) At the training time, to relate the predictions during the training and the ground truth Intersection of Union (IoU) is applied.

The loss function of single shot detector is very complex. The loss function manages many objectives (Regression, classification, to check if there is object or no object is managed by the loss function).

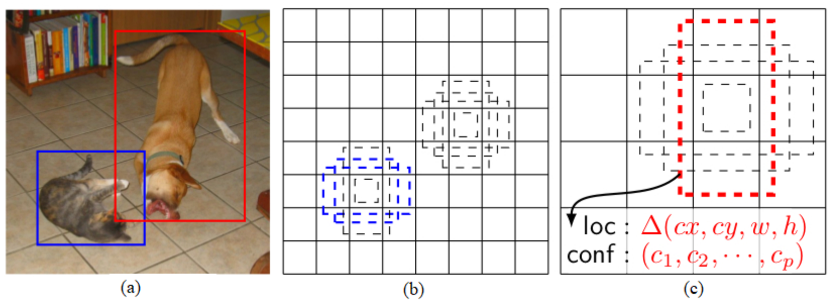


Figure. 15. SSD framework (a) Image with ground truth boxes. (b) In convolutional fashion, default boxes at aspect ratio in 4×4 and 8×8 scale is collected for different resolution for the feature maps. (c) For all object categories in the default boxes shape offsets and confidence scores are calculated.

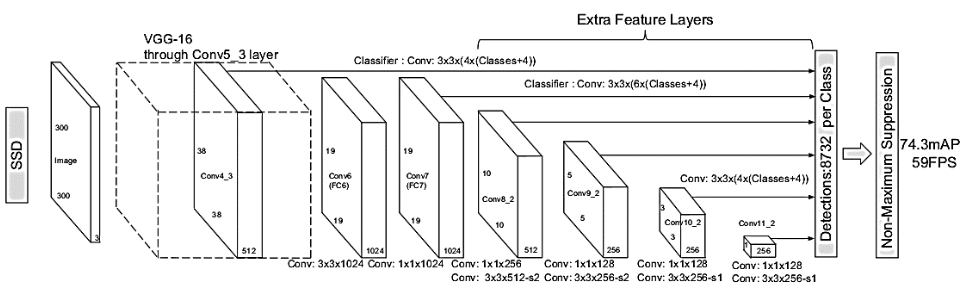


Figure. 16. SSD architecture

3.10.2 *Advantages*

—High speed and accurate detection due to a greater number of bounding boxes. Detector runs at 59 fps on 300×300 size input image. Multi box is applied at a greater number of layers. This leads to better detection as the detector run on multiple features at different image resolutions.

—This method does detection in multi-resolution images.

3.10.3 *Disadvantages*

—This method is based on base VGG-16 network and 80% of the time is spent in training the model. The performance of the method can be improved by reducing the training time.

—This method confuses objects belonging to the same class. This is due to location sharing of multiple classes.

—The features of small objects are not spread across all the feature maps. Thus, this method finds difficulty in detection of small objects.

A qualitative comparison of the state-of-the-art methods based on dataset, accuracy of detection, features, issues and applications is presented in Table II.

Method	**Dataset**	**Accuracy**	**Features**	**Issues**	**Applications**
Histogram of Gradients (HOG)	MIT Pedestrian Test	99%	Computationally inexpensive, very fast detection, very low false positive and miss rate	Not suitable for non-static images, Not suitable for large dataset, Old approach	Pedestrian detection, Face detection
Region Proposal based Convolutional Neural Network (RCNN)	PASCAL VOC 2011	62 mAP	Less localization errors	Background detection problem, computationally expensive as segmentation and region proposal process is performed	Object detection. Object categories include bike, car, bottle, cat, chair etc.
Spatial Pyramid Pooling Network	ImageNet -ILSVRC 2014	35.1 mAP	Less computation cost than R-CNN, Does detection on multi-scale images	Lengthy model, Not suitable for real-time detection	Visual recognition
Fast - RCNN	PASCAL VOC 2012	66 mAP	Less computational cost than RCNN, Less number of steps for region proposal extraction, Less localization errors.	Based on selective search, take much time to extract regions of interests, Not suitable for real-time detection	Object detection. Object categories include bike, car, bottle, cat, chair etc.
Faster-RCNN	PASCAL VOC 2012	75.9 mAP	Fast detection at 0.2 seconds per image, Less expensive than RCNN, FRCNN	Generate region proposals slow, Background detection problem	Object detection
Mask-RCNN	COCO Test set	Mask Average Precision of 35.7%	Object detection at pixel boundary level, Accurate detection of object class	Detection at 5 fps, i.e., very low, Computationally expensive	Human pose retrieval
You Only Look Once (YOLO)	PASCAL VOC 2007	63.4 mAP	Very fast near to human eye visualization, Detect background accurately	Prone to localization errors	Object detection, Artwork detection
YOLO 9000	PASCAL VOC 2007	78.6 mAP	Allows generalization for learning new classes, Detect small objects, train from varying image scales	Less localization errors as compared to original YOLO	Object detection
YOLO V3	COCO	57.9 mAP	Less localization errors for small objects, more features are extracted at varying scales	Precision can be improved for medium and large objects	Object detection
Single Shot Detector (SSD)	PASCAL VOC 2012 and COCO	74 mAP	Multiple systems, Multi box detection leads to better detection, Detector run at multiple resolutions that helps in gathering more features	Does poor detection on small objects, much time is wasted on VGG-16 training which effects the overall performance	Object detection

Table II: A qualitative comparison of the state-of-the-art methods Method.

A comparison of state-of-the-art object detection methods based on computational factors is illustrated in Table III.

Method	**Approach**	**Multi-Scale Input**	**Learning Factor**	**Loss Function**	**Softmax Layer**	**End-to-End Train**	**Platform**
Region Proposal based Convolutional Neural Network (RCNN)	Selective Search	No	Stochastic Gradient Descent, Belief Propagation	Classification Loss, Bounding Box Regression	Yes	No	Caffe/ Matlab
Spatial Pyramid Pooling Network	Edge Boxes	Yes	Stochastic Gradient Descent	Classification Loss, Bounding Box Regression	Yes	No	Caffe/ Matlab
Fast - RCNN	Selective Search	Yes	Stochastic Gradient Descent	Class Log Loss and Bounding Box Regression	Yes	No	Caffe/ Python
Faster-RCNN	Region Proposal Network	Yes	Stochastic Gradient Descent	Class Log Loss and Bounding Box Regression	Yes	Yes	Caffe/ Python
Mask-RCNN	Region Proposal Network	Yes	Stochastic Gradient Descent	Class Log Loss and Bounding Box Regression and Semantic Sigmoid Loss	Yes	Yes	Tensorflow/ Keras/ Python
You Only Look Once (YOLO)	Anchor Boxes with Non-Max Suppression	No	Stochastic Gradient Descent	Class Sum-Squared Error Loss, Bounding Box Regression, Object Confidence and Background Confidence	Yes	Yes	Darknet/ C Language
YOLO 9000	Anchor Boxes with Non-Max Suppression	No	Stochastic Gradient Descent	Class Sum-Squared Error Loss, Bounding Box Regression, Object Confidence and Background Confidence	Yes	Yes	Darknet/ C Language
YOLO V3	Anchor Boxes with Non-Max Suppression	No	Stochastic Gradient Descent	Class Sum-Squared Error Loss, Bounding Box Regression, Object Confidence and Background Confidence	Yes	Yes	Darknet/ C Language
Single Shot Detector (SSD)	No Proposal Based Approach	No	Stochastic Gradient Descent	Class Softmax Loss and Bounding Box Regression	No	Yes	Caffe/ C++ Language

Table III: Characterization of object detection methods on the basis of architecture

4. RESEARCH ISSUES

In this section various research issues of object detection area are presented. The methods discussed in this work address few of the issue but many issues are still the open area of research.

4.1 Active Vision

The methods discussed in this work learn object categories by training the model by feature representation of the object classes. The object detection methods proposed so far does not learn classes of new objects by transfer learning, i.e., learning new categories of objects from the environment without supervised learning. If the methods start learning object features and classes of self then this can save cost of manual training by many folds. This area can contribute a lot in the area of robotics and autonomous systems.

4.2 Background Problem and Image Inconsistency

The methods discussed in this work neglects presence of objects in the background. The method of Mask-RCNN addresses this issue by applying pixel level segmentation but it boundaries only the foreground objects neglecting the background. The process of pixel level segmentation increases the computation cost and slow down the detector. The other issue of image inconsistency is addressed by single shot detector method which feeds feature maps of the input image by zooming the input image to the Fully Connected layer, but this method suffers with problem of localizing and detecting the small objects.

4.3 Localizing Small Objects

The method of Single Shot Detector and You Only Look Once suffer problem of localization of small objects in the image. Many methods based on Anchor Box, Bounding Box, Non-Max Suppression and Intersection of Union are proposed in the work discussed in this paper. This area is of biggest concern as the detectors with fast detection rate are suffering from this issue. A solution focusing Data Fusion, i.e., geometric parameters other than x-coordinate and y-coordinate for anchor box and bounding box should be proposed in future.

4.4 Multi-Modal Detection

At varying depth of images captured through satellite cameras and thermal cameras it is difficult to detect presence of an object in the image. In future most of the surveillance will be done from images captured through satellite cameras thus for task of pedestrian detection, place detection and vehicle detection this area is to be addressed for accurate detection.

4.5 Object Part Relation

No method discussed in this work address the issue of what to detect first. Object or its part? As this creates a dilemma. In aerial images, both drone and bird present the same features when projected from one side. Human eyes can differentiate between a drone and a bird but for an object detection method and computer vision system it is a very difficult task. Methods should be developed establishing connection between objects and their parts.

4.6 Optimize Deep Learning Models

All convolutional neural network based methods discussed in this work are trained and evaluated on large datasets. The deep learning based methods should be optimized to learn features from small datasets and perform detection on same.

5. CONCLUSION

This work discusses the various state-of-the art object detection methods and presents a comparison of the same. The working method of all the object methods is discussed. The study signifies that the current day object detectors based on convolutional neural network are very fast and are able to do detection at real-time. The YOLO V2, YOLO V3 and Single Shot Detector method are very fast and can detect small objects with very low localization error rate, however, detection of small objects is prone to errors. It should be well addressed to decrease the errors. The Single Shot Detector is fastest among all and does object detection at 59 frames per second. A significant work should be done to improve the efficiency of Mask RCNN method as it is based on semantic segmentation. Improving Mask RCNN can lead the area of object detection to do predictions at pixel level information of the image. Solution should be provided to decrease the computation cost of the Mask RCNN. Not enough methods are present to detect the background objects, this area is also to be addressed. Furthermore, all the object detection methods are based on general object classes. The methods are required to be trained with more classes and in future more detectors with active vision should be developed so that the object detectors can learn new classes of the objects from the environment without manual training process. As the detectors can now detect very fast, i.e., near to human visualization, such solutions should be proposed that the object detectors are integrated with surveillance systems to do real-time detection.

REFERENCES

Ahn, H. and Rhee, S.-B. 2015. Research of object recognition and tracking based on feature matching. In *Computer Science and its Applications*. Springer, 1071–1076.

Boruah, A., Kakoty, N. M., and Ali, T. 2018. Object recognition based on surface detection-a review. *Procedia computer science 133*, 63–74.

Bridle, J. S. 1990a. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*. Springer, 227–236.

Bridle, J. S. 1990b. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In *Advances in neural information processing systems*. 211–217.

Caselles, V. and Coll, B. 1996. Snakes in movement. *SIAM Journal on Numerical Analysis 33,* 6, 2445–2456.

Cheng, G. and Han, J. 2016. A survey on object detection in optical remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing 117*, 11–28.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., and Schmidhuber, J. 2011. Flexible, high performance convolutional neural networks for image classification. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

Dalal, N. and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. IEEE, 886–893.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., and Equitz, W. 1994. Efficient and effective querying by image content. *Journal of intelligent information systems 3,* 3-4, 231–262.

Fischler, M. A. and Elschlager, R. A. 1973. The representation and matching of pictorial structures. *IEEE Transactions on computers 100,* 1, 67–92.

Fukushima, K. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics 36,* 4, 193–202.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

Girshick, R., Donahue, J., Darrell, T., and 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*. 580–587.

Glumov, N., Kolomiyetz, E., and Sergeyev, V. 1995. Detection of objects on the image using a sliding window mode. *Optics & Laser Technology 27,* 4, 241–249.

Gonzalez, R. C., Woods, R. E., and Eddins, S. L. 2004. *Digital image processing using MATLAB*. Pearson Education India.

Grove, T., Baker, K. D., and Tan, T. 1998. Colour based object tracking. In *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*. Vol. 2. IEEE, 1442–1444.

Hadi, R. A., Sulong, G., and George, L. E. 2014. Vehicle detection and tracking techniques: a concise review. *arXiv preprint arXiv:1410.5894*.

He, K., Gkioxari, G., Dollr, P., and Girshick, R. 2017. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

He, K., Zhang, X., Ren, S., and Sun, J. 2014. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, Cham, 346–361.

Heikkila, M. and Pietikainen, M. 2006. A texture-based method for modeling the background and detecting moving objects. *IEEE transactions on pattern analysis and machine intelligence 28,* 4, 657–662.

Jain, A. K., Zhong, Y., and Lakshmanan, S. 1996. Object matching using deformable templates. *IEEE Transactions on pattern analysis and machine intelligence 18,* 3, 267–278.

KASS, M., WITKIN, A., AND TERZOPOULOS, D. 1988. Snakes: Active contour models. *International journal of computer vision 1,* 4, 321–331.

KAUSHAL, M., KHEHRA, B. S., AND SHARMA, A. 2018. Soft computing based object detection and tracking approaches: State-of-the-art survey. *Applied Soft Computing 70*, 423–464.

KULCHANDANI, J. S. AND DANGARWALA, K. J. 2015. Moving object detection: Review of recent research trends. In *2015 International Conference on Pervasive Computing (ICPC)*. IEEE, 1–5.

LECUN, Y., HAFFNER, P., BOTTOU, L., AND BENGIO, Y. 1999. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*. Springer, 319–345.

LIU, W., ANGUELOV, D., ERHAN, D., SZEGEDY, C., REED, S., FU, C.-Y., AND BERG, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*. Springer, 21–37.

NISA, S. U. AND IMRAN, M. 2019. A critical review of object detection using convolution neural network. In *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*. IEEE, 154–159.

REDMON, J., DIVVALA, S., GIRSHICK, R., AND FARHADI, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

REDMON, J. AND FARHADI, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

REN, S., HE, K., GIRSHICK, R., AND SUN, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

SHARMA, K. U. AND THAKUR, N. V. 2017. A review and an approach for object detection in images. *International Journal of Computational Vision and Robotics 7,* 1/2, 196–237.

TSAI, C.-F. 2012. Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence 2012*.

VERSCHAE, R. AND RUIZ-DEL SOLAR, J. 2015. Object detection: current and future directions. *Frontiers in Robotics and AI 2*, 29.

VINOD, V. V. AND MURASE, H. 1997. Video shot analysis using efficient multiple object tracking. In *Proceedings of IEEE International Conference on Multimedia Computing and Systems*. IEEE, 501–508.

WIXSON, L. 2000. Detecting salient motion by accumulating directionally-consistent flow. *IEEE transactions on pattern analysis and machine intelligence 22,* 8, 774–780.

ZHIQIANG, W. AND JUN, L. 2017. A review of object detection based on convolutional neural network. In *2017 36th Chinese Control Conference (CCC)*. IEEE, 11104–11109.

Mr. Akhil Kumar is a Ph.D. student at Himachal Pradesh University, India. He has obtained B.Tech and M.Tech from USIT, GGS Indraprastha University, India. His research interests include Object Detection and Deep Learning.

Dr. Arvind Kalia is a Professor at Department of Computer Science, Himachal Pradesh University, India. He has obtained Ph.D. from Punjabi University, Patiala. He has published more than 100 research papers in international journals and conferences. His research interests include Software Engineering, Data Mining and Computer Networks.

Dr. Akashdeep Sharma is an Assistant Professor at UIET, Panjab University, India. He has obtained his Ph.D. from GND University, Amritsar. His research interests include Deep Learning, Video Analytics and Sensor Networks.