# YOLO Based Object Detection System: A Survey

**Prof.Dipti Chaudhary[1], Purva Dhote [2], Aarti Alte [3], Aditi Sherkhane[4], Tanvi Dhewade[5]**

*[1]Professor(Computer Engineering), Dr. D. Y. Patil Institute of Technology, Pimpri, Pune,Maharashtra, India.*
*[2,3,4,5]BE(Computer Engineering), Dr. D. Y. Patil Institute of Technology, Pimpri, Pune,Maharashtra, India.*

---

***Abstract***– The Objective is to detect of objects using You Only Look Once (YOLO) approach. This method has several advantages as compared to other object detection algorithms. In other algorithms like Convolutional Neural Network, FastConvolutional Neural Network the algorithm will not look at the image completely but in YOLO the algorithm looks the image completely by predicting the bounding boxes using convolutional network and the class probabilities for these boxes and detects the image faster as compared to other algorithms.

*Key words* – Convolutional Neural Network, Fast-Convolutional Neural Network, BoundingBoxes, YOLO.

## 1. INTRODUCTION

This project will be developed to make the life of blind people easy. This is a camera based system to scan the product behind the image and read the description of the product with the help of Id stored in the product. This is very beneficial in case of finding out the description of packaged goods to the blind people and thus helping them in deciding to purchase a product or not especially which are packaged. This is because it becomes very difficult for the blind people to distinguish between the packaged goods. In order to use this system, all the user needs to do is capture the image on the product in the device which then resolves the product which means it scans the image to find out the Id stored. Thus this application really benefits blind and visually impaired people and thus making their work of identifying products easy.

This is very easy to use and affordable as it requires a scanner to scan the product and a camera phone to take the picture of the image containing the product. This is now easy to implement as most of the devices today have the required resolution in order to scan the product to identify the Id stored in it and read out the product description. This project can be implemented in any shopping mall, supermarket, Book stores, Medical stores etc.

## 2. BACKGROUND

The aim of object detection is to detect all instances of objects from a known class, such as people,cars or faces in an image. Generally, only a small number of instances of the object are present in the image, but there is a very large number of possible locations and scales at which they can occur and that need to somehow be explored. Each detection of the image is reported with some form of pose information. This is as simple as the location of the object, a location and scale, or the extent of the object defined in terms of a bounding box. In some other situations, the pose information is more detailed and contains the parameters of a linear or non- linear transformation.

For example for face detection in a face detector may compute the locations of the eyes, nose and mouth, in addition to the bounding box of the face. An example of a bicycle detection in an image that specifies the locations of certain parts is shown in Fig 1.

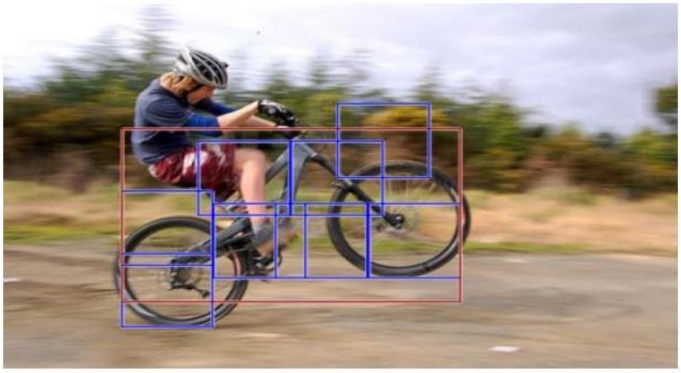

**Fig-1:** Input Image

The pose can also be defined by a three-dimensional transformation specifying the location of the object relative to the camera. Object detection systems always construct a model for an object class from a set of training examples. In the case of a fixed rigid object in an image, only one example may be needed, but more generally multiple training examples are necessary to capture certain aspects of class System show by way of

experiments on challenging UPC-A barcode images from five different databases that this approach outperforms competing algorithms. Implemented on a Nokia N95 phone, this algorithm can localize and decode a barcode on a VGA image (640 480, JPEG compressed) in an average time of 400-500 ms variability.

Convolutional implementation of the sliding windows Before we discuss the implementation of the sliding window using convents, let us analyze how we can convert the fully connected layers of the network into convolutional layers. Fig. 2 shows a simple convolutional network with two fully connected layers each of shape .

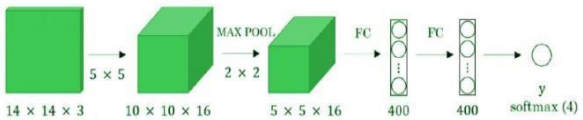


**Fig-2:** Simple Convolution Network

A fully connected layer can be converted to a convolutional layer with the help of a 1D convolutional layer. The width and height of this layer is equal to one and the number of filters are equal to the shape of the fully connected layer. An example of this is shown in Fig 3.

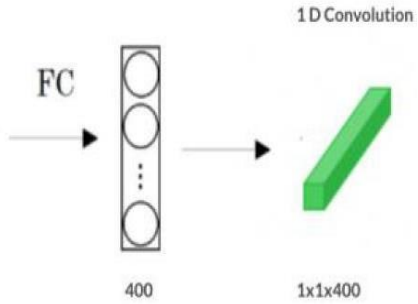


**Fig-3**

We can apply the concept of conversion of a fully connected layer into a convolutional layer to the model by replacing the fully connected layer with a 1-D convolutional layer. The number of filters of the 1D convolutional layer is equal to the shape of the fully connected layer. This representation is shown in Fig 4 Also, the output softmax layer is also a convolutional layer of shape (1, 1, 4), where 4 is the number of classes to predict.

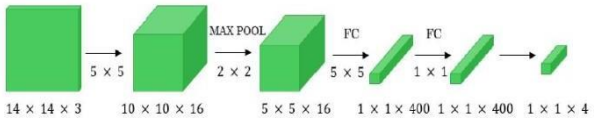


**Fig-4**

Now, let's extend the above approach to implement a convolutional version of the sliding window. First,

let us consider the ConvNet that we have trained to be in the following representation (no fully connected layers).

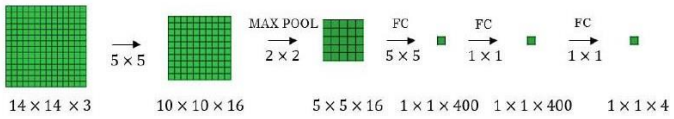


**Fig-5**

Let's assume the size of the input image to be 16 × 16 ×

**3.** If we are using the sliding window approach, then we would have passed this image to the above ConvNet four times, where each time the sliding window cropsthe part of the input image matrix of size 14 × 14 × 3 and pass it through the ConvNet. But instead of this, we feed the full image (with shape 16 × 16 × 3) directly into the trained ConvNet (see Fig. 6). This results will give an output matrix of shape 2 × 2 × 4. Each cell in the output matrix represents the result of the possible crop and the classified value of the cropped image. For example, the left cell of the output matrix(the green one) in Fig. 6 represents the result of the first sliding window. The other cells in the matrix represent the results of the remaining sliding window operations

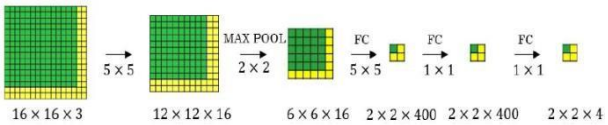


**Fig-6**

The stride of the sliding window is decided by the number of filters used in the Max Pool layer. In the example above, the Max Pool layer has two filters, and for the result, the sliding window moves with a stride oftwo resulting in four possible outputs to the given input. The main advantage of using this technique is that the sliding window runs and computes all values simultaneously. Consequently, this technique is really fast. The weakness of this technique is that the position of the bounding boxes is not very accurate. A better algorithm that tackles the issue of predicting accurate bounding boxes while using the convolutional sliding window technique is the YOLO algorithm. YOLO stands for you only look once which was developed in 2015 by Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. It is popular because it achieves high accuracy while running in real-time. This algorithm requires only one forward propagation pass through the network to make the predictions. This algorithm

divides the image into grids and then runs the image classification and localization algorithm (discussed under object localization) on each of the grid cells. For example, we can give an input image of size 256 × 256. We place a 3 × 3 grid on the image.

## 3. MATERIALS AND METHODS

### 3.1 Data Collection

To collect the necessary data for evaluation, we captured many images through pre-installed cameras in the location. The captured images are matched with the stored images either deterministically or probabilistically

### 3.2 Data Preprocessing

For preprocessing YOLO Algorithm is used, YOLO model divides an image into S×S grid. Each grid cell predicts B bounding boxes, and boxes' confidence scores for the prediction and detect if a class falls in the boxes. The confidence is defined as P r(object) × IOUtruth pred , which represents the confidence of a class in the box and accuracy of the box coordinates. Thus, each box has five parameters to predict: x, y, w, h and confidence. Each grid cell also predicts P r(Classi |Object). Thus the confidence for each box is P r(Classi |Object) × P r(object) × IOUtruth pred = P r(Classi) × IOUtruth pred . The overall variables to be predicted can be represented as a S × S × (B × 5 + C) tensor.

## 4. METHODS AND TERMS

### 4.1 YOLO Algorithm

This is done using the Yolo algorithm which accepts the images and converts it into gray scale and gives the results in the form of binary value.0 for black and 1 for white. It converts the large image into 50-50 cube and is then processed using Yolo algorithm. It divides the cube by 25 units each. It takes first image as input and then by pre-processing the final output is been drawn after respective expected output. The pre-processing has total 4*5=20 cubes for result analysis in which the first is input cube and the last is output cube.

Step 1: YOLO first takes an input image:

Step 2: The framework then divides the input image intogrids (say a 3 X 3 grid):

Step 3: Image classification and localization are applied on each grid. YOLO then predicts the bounding boxes and their corresponding class probabilities for objects (ifany are found, of course).


**Fig-8:** Input Image into Grid

### 4.2 RCNN

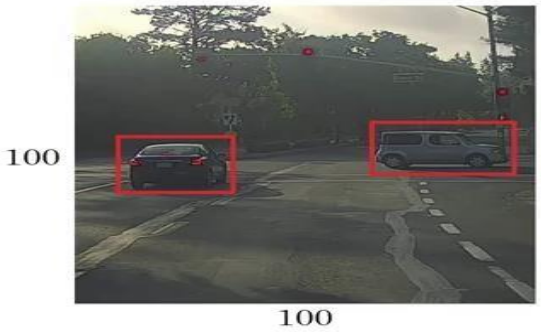


**Fig-7:** Input Image

Region based convolutional neural networks (RCNN) algorithm uses a group of boxes for the picture and then analyses in each box if either of the boxes holds a target. It employs the method of selective search to pick those sections from the picture. In an object, the four regions are used.

### 4.3 Histogram of Oriented Gradient

HOG is a feature descriptor that is extensively applied in various domains to distinguish objects by identifying their shapes and structures. Local object structure, pattern, aspect,and representation can usually be characterized by the arrangement of gradients of local intensity or the ways of edges.

### 4.4 Spatial Pyramid Pooling

Spatial Pyramid Pooling (SPP) is a pooling layer that removes the fixed-size constraint of the network, i.e. a CNN does not require a fixed-size input image.

Specifically, we add an SPP layer on top of the last convolutional layer. The SPP layer pools the features and generates fixed-length outputs, which are then fed into the fully-connected layers (or other classifiers). In other words, we perform some information aggregation at a deeper stage of the network hierarchy (between convolutional layers and fully-connected layers) to avoid the need for cropping or warping at the beginning.

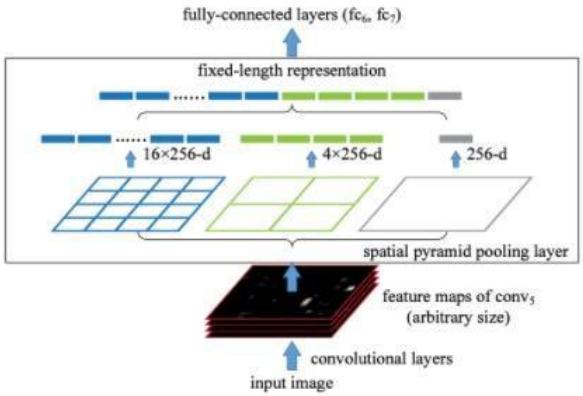


**Fig-9:** A Network Structure with a spatial pyramid pooling layer

## 5. CONCLUSION

The difficulty of blind peoples to identify the objects helps us to think upon this and try to ease their life by making device to identify and read out the object and make the decision according it. To identify the objects and read it out, this system uses YOLO algorithm to build an efficient and smart object detection model and then converts the name of the object which is text into the voice so that blind peoples can identify. This algorithm can be implemented in various fields to solve some real-life problems like security, monitoring traffic lanes or even assisting visually impaired people with help of audio feedback.

## 6. FUTURE WORK

With regards to future implementations, The system is expected to become a compact using glasses and bluetooth. It can be used for personal use to identify objects. It can be used in any shopping mall, supermarket, Book stores, Medical stores etc. It can be used while walking on the road. Visually Impaired person can identify the product. Once product is identified there will be voice of that product.

## REFERENCES

[1] Joseph Redmon and Anelia Angelova, Real-TimeGrasp Detection Using Convolutional Neural Networks (ICRA), 2015.

[2] A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

[3] Saurabh Gupta, Ross Girshick, Pablo Arbelaez and Jitendra Malik, Learning Rich Features from RGBD Images for Object Detection and Segmentation (ECCV), 2014.

[4] Tadas Naltrusaitis, Peter Robison, and Louis-Phileppe Morency, 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking (CVPR), 2012.

[5] Andrej Karpathy and Fei-Fei Li, Deep VisualSemantic Alignments for Generating Image Descriptions (CVPR), 2015.

[6] David Brown, Tom Macpherson, and Jamie Ward, Seeing with sound? exploring different characteristics of a visual-to-auditory sensory substitution device. Perception, 40(9):1120–1135, 2011.

[7] Liam Betsworth, Nitendra Rajput, Saurabh Srivastava, and Matt Jones. Audvert: Using spatial audio to gain a sense of place. In Human-Computer Interaction– INTERACT 2013, pages 455–462. Springer,2013.