Literature Survey on Object Detection using YOLO

Rekha B. S.[1], Athiya Marium[2], Dr. G. N. Srinivasan[3], Supreetha A. Shetty[4]

[1,3]Professor, Dept. of Information Science and Engineering, R V College, Karnataka, INDIA
[2,4]Dept. of Information Science and Engineering, R V College, Karnataka, INDIA

---***---

Abstract - *Object detection is important for computer vision. The problems such as noise, blurring and rotating jitter, etc. with images in real-world have an important impact on object detection. The objects can be detected in real time using YOLO (You only look once), an algorithm based on convolutional neural networks. This paper addresses the various modifications done to YOLO network which improves the efficiency of object detection.*

Key Words: *Object detection, YOLO, Convolution neural networks, light field camera, pedestrian detection, obstacle detection*

1. INTRODUCTION

Object detection plays an important role in computer vision, automatic vehicles, industrial automation etc. Detecting objects in real time is a challenging task. Deep learning in object detection is better than traditional target detection. Deep learning methods include Region proposal object detection algorithms wherein it generates region proposal networks and then classify them. These include SPP-net, Region-based Convolutional Neural Networks, Fast-RCNN, Faster-RCNN etc. Regression object detection algorithms like SSD and YOLO generate region proposal networks and classify them at the same time. This paper summarizes the various real time object detection approaches based on YOLO (You Only Look Once).

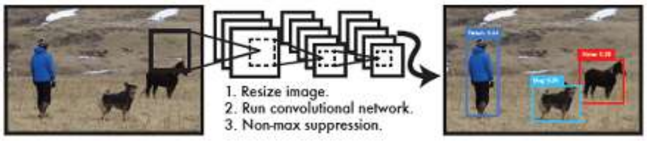

Fig. 1. The YOLO Detection System . Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence. *[1]*

The paper is organized as follows: Section 2 briefs about the base architecture of YOLO. The next section outlines the applications that uses architecture, datasets used in the experiment, experimental results, pros & cons of YOLO architecture, conclusion and future work.

2. ARCHITECTURE

● Unified Detection

The components involved in object detection are separated and unified into a single neural network by the YOLO architecture. All bounding boxes are simultaneously projected from the entire image, which means that the network can handle the entire image with all its objects.

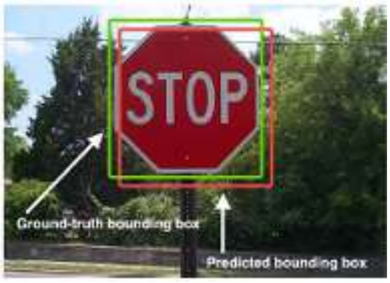

Fig. 2. Understanding Ground truth and predicted bounding boxes in object detection.

● The Methodology

The basic steps that are followed in object detection using YOLO according to the paper presented by Joseph Redmon et. al [1], are as follows:

- An S x S grid is obtained by dividing the image. B bounding boxes are predicted by each grid cell.
- Intersection of Union (IoU): It is a metric for evaluation in object detection. Consider Figure 1 where the areas of Predicted bounding boxes and Ground Truth are shown. We calculate IoU as:

$$IoU = \frac{Area\ of\ overlap}{Area\ of\ union} \qquad (1)$$

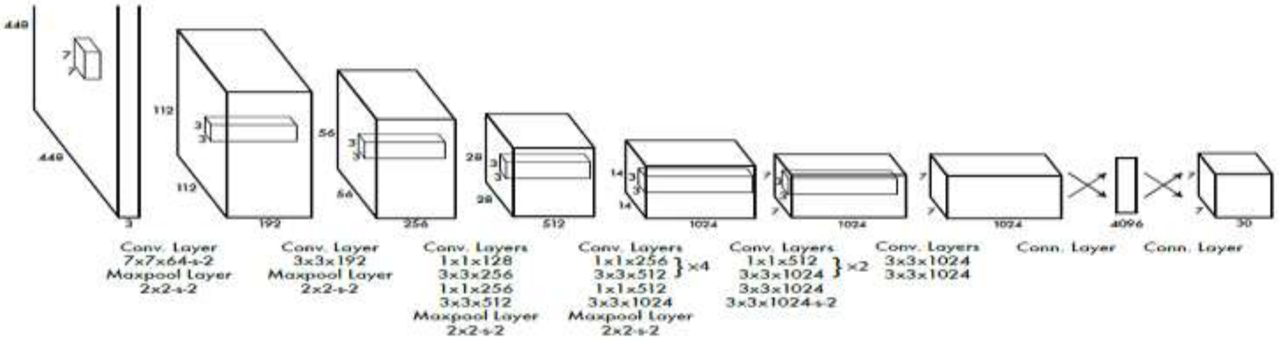

Fig. 3. The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1 × 1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224 × 224 input image) and then double the resolution for detection [1].

Confidence score for each bounding box is given as Pr(Object) * IoU_{pred}^{truth} . The Confidence score for cells with no object should be zero. A total of 5 predictions are made for every bounding box: x, y, w, h and confidence along with C.

The center of the box relative to the grid square is (x,y).

w is the width and h is height that is relative to the entire image.

C is the conditional class probabilities.

At test time, the following formula gives confidence scores of every class for each of the boxes:

$$\Pr(Class_i|\text{Object})*\Pr(\text{Object})*IoU_{pred}^{truth} = \Pr(Class_i)*IoU_{pred}^{truth} \quad (2)$$

This gives us two values, with what probability a class will be a part of the box and with what confidence it fits that box.

The predictions are encoded as S × S × (B * 5 + C). When evaluating YOLO on PASCAL VOC [6], which has 20 labelled classes i.e. C = 20, take S = 7, B = 2. The final prediction is given as a 7 * 7 * 30 tensor.

Design

The model is implemented as a CNN and evaluated on PASCAL VOC dataset for detection [6]. The first few convolutional layers of the network are used to get the features from images and the fully connected layers are to predict the probabilities and the respective coordinates. There are 24 convolutional layers in the model that are followed by 2 fully connected layers. The complete network is drawn in Figure 3.

3 APPLICATIONS AND THEIR ARCHITECTURES

3.1 Object Detection in degraded images

Real world images are sometimes noisy, blurred, rotated or jittered. Detecting these images is an important part of object detection. Chengji Liu et. al [2] proposes an image degradation model that uses images that are degraded for the test set. The degraded images were run on a standard model, which was trained on regular images. The source network was then modified by training the model with the degraded images, which were obtained by performing some degradation processes on them.

The accuracy of the test set is calculated on both the models and are compared. Then the training set is modified again by performing further complex degradation processes and a more generalized model for detection is obtained from this. This was also compared with the standard test performance. The final object detection model obtained thus is optimized and the generalization ability had been enhanced, while the accuracy improved. For the ease of description of Image degradation following assumptions were made:

- Consider the lower left corner of the image to be the point of origin (0,0), the width of the image would be the x-axis coordinate and the image height be y-axis coordinate. Which means the image was considered to be in the first quadrant.
- Considering that blurring happens because of overlapping images which is caused by a uniform linear motion of the object. Let the images, without any blur and noise be represented by g(x, y), and the blurred images be represented as h(x, y). The formula is:

$$h(x,y) = \int_0^t g(x + c_x t, y + c_t)dt$$
(3)

where,

- t - exposure time,
- c_x- speed on the x-axis and
- c_y- speed on the y-axis

- Rotation of the image of width x and height y to new width x' and height y' with angle of rotation β is given as

$$(x'y') = (x\ y) * \begin{pmatrix} \cos\beta & \sin\beta \\ -\sin\beta & \cos\beta \end{pmatrix} \quad (4)$$

- Noise: To add noise to make the image degraded, Gaussian and salt & pepper noise is considered. Gaussian noise follows the Gaussian curve for disturbance. The formula is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (5)$$

where μ and σ represents mathematical expectation and standard deviation respectively. For salt & pepper the noise equation is:

$$f(x) = \begin{cases} P_a & x = a \\ P_b & x = b \\ 1 - P_a - P_{b,} & otherwise \end{cases} \quad (6)$$

In this experiment, 0.08 was considered as the value of salt & pepper noise, which means that 0 or 255 was randomly assigned to 8% pixels in the images.

- Cropping: The images were cropped randomly. The width and height cropped was between range 0 to 0.15.

Dataset: Traffic signs in 1,652 images from ImageNet was the dataset for the experiment. Out of these the standard training set was formed by 1,318 images and the test set was formed by 334 images. Furthermore, in order to analyze the performance of the model, 334 images in the real world were considered to form the test set. Degeneration processing was applied on all these images.

Experimental Result & Conclusion:

- The experiment was carried out on the single standard model by training the model without degrading the images. Then, on the test sets a single image degradation was performed and fed to the standard model. The image degradation resulted in lowering of average precision. This showed that for detection of degraded images, the model trained with standard sets has low robustness and was not generalized.
- Then the model was trained by performing an image degradation on the training set. This results in different degenerative models. It was observed that the average precision levels improved for these models.
- Next the general degenerative model was evaluated against the standard model. The average precision for degraded images was better in general degenerative model compared to the standard model.

Thus, the model trained using degraded training image sets was found to have higher generalization and better robustness.

3.2 Pedestrian Detection

Wenbo et. al [3] proposes a YOLO-R, as an improvement on YOLO, a new network structure that increased accuracy of the network when it predicted shallow pedestrian features. Two additional layers were incorporated to the early YOLO architecture called Passthrough layers.

The two Passthrough layers are the Route layer and the Reorg layer.

- Route layer - This layer passes characteristic pedestrian information from the identified layer to the present layer and then it uses the Reorg layer.
- Reorg layer - this layer first reorganizes the feature maps to map features of the new Route layer to the next layer's feature map.

Dataset [3]: For pedestrian detection the dataset that is used is the INRIA dataset. The dataset has images and the annotations corresponding to each image. A total of 614 positive samples are present in the dataset. 2416 pedestrians are considered as the positive samples. 288 of the positive samples with 1126 pedestrian pictures in them were considered as the test set. Most of the pictures are high definition and have people standing up with more than 100 pixels as height.

Experimental Result & Conclusion:

TABLE I: COMPARISON OF THE RESULTS OF THE TWO ALGORITHMS ON THE INRIA TEST SET [3]

	YOLO-v2	YOLO-R
Precision	97.37%	98.56%
Recall	89.33%	91.21%
IoU	74.46%	76.18%

$$P = \frac{TP}{(TP+FN)} \quad \textbf{(7)}$$

$$R = \frac{TP}{(TP+FP)} \quad \textbf{(8)}$$

where, P and R are Precision and Recall respectively.

TP - is True Positive values given as the number of outputs that correctly predicted person on street as pedestrian.
FP - is False Positive given as the number of outputs that predicted as pedestrians, other objects which are not pedestrians.

FN - is False Negative given as the number of samples that identified person on the street as something else.

The Loss curves in the training process of the two algorithms are compared-
YOLO-R detects pedestrian that is missed by YOLO-2.
Obstructed pedestrian is detected by YOLO-R.
False pedestrian detection made by YOLO 2 is rectified.

LAMR (Log-Average miss RATE) is the estimation index for evaluating the performance of the new passthrough layers for the algorithm. LAMR talks about the association between the false negative per image (FPPI) and missed rate. The performance is better for lower FPPI values.

3.3 Obstacle Detection using Light Field camera

An unconventional obstacle detection algorithm for an indoor environment is proposed in the paper [4]. The algorithm makes use of YOLO architecture for detecting objects and a light field camera for capturing images. It classifies identified objects into classes and marks them in the image. The input to algorithm is provided by the RGB images from the light field camera. The model is trained on more than 100 classes of common everyday objects. Some of the earlier sensing techniques are:

An energy function is minimized and the image is separated from its background using an IR line.

An algorithm that uses the technique of fusing laser data with vision for obstacle detection.

A detection system that is commonly used in pedestrian detection and on road vehicle detection uses deep learning for predictions. This algorithm needs lots of data to train the model and hence the cost of system is usually high. For an indoor environment system this is not useful.

For self-driving cars, there are algorithms written to predict and analyze brake lights using deep learning and so on.

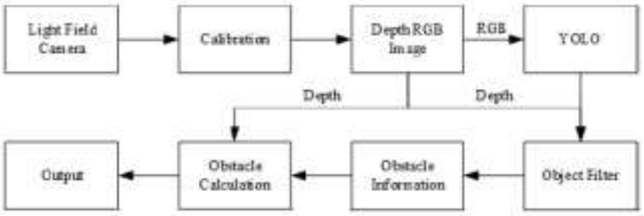

Fig 4. Framework of the proposed method to detect obstacles [4].

Dataset [4]: The model was built for detecting obstacles indoors. Training data was found by downloading common obstacle images and labelling them before they are sent as training data to the model.

Experimental Result & Conclusion: The images of the obstacles found in common were labeled and used for training YOLO. Then the filter is applied to remove the unconcern obstacles. To prove the effectiveness of this obstacle detection algorithm, different types of scene including chairs, books, pedestrian are demonstrated.

3.4 Identification and Detection of Automotive Door Panel Solder Joints

The paper [5] gives an algorithm using deep learning techniques for providing the category and position of solder joints of automotive door panels identification in real time. The solder joints location in automotive door panels alters many times because of the variety and complexity of work environments. As such, if the location of the solder joints is not identified precisely by automation, instructing and programming has to be repeated frequently by manual interference. This affects the efficiency and quality of the welding and results in lower intelligence and automation of the manufacturing line. YOLOv1 and YOLOv2 are not suitable since the size of solder joint of a car door is smaller. Hence, YOLOv3 [8] algorithm is used for detecting the small solder joints more precisely which differs from YOLOv2 only in the final result. Multiple levels of predictions are adopted in which prediction is done on different size feature maps and the predictions from these are combined to get the final output. Each grid cell in 52*52 feature maps contain only one object which can precisely detect the solder joints.

Dataset [5]: As the dataset, 61 front door and 57 rear door solder joints of car were considered. There are three types of solder joints, Rectangular, semi-circular and circular. Out of 553 door panel photos that were taken, 447 door panels were used for training and validations. The remaining 106 were used are test set.

YEAR	AUTHOR	TITLE & ARCHITECTURE	DOMAIN & PROBLEM STATEMENT	CONCLUSION & RESULT
2016	Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi	You Only Look Once: Unified, Real-Time Object Detection	A fast and simple approach to detecting real time images was introduced in this paper as You Only Look Once. The model was built to detect images accurately, fast and to differentiate between art and real images.	In comparison with Object detection techniques that came before YOLO, like R-CNN, YOLO introduced a single unified architecture for regression go image into bounding boxes and finding class probabilities for each box. This meant that YOLO performed much faster and also provided more accuracy. It could also predict artwork correctly.
2018	Chengji Liu, Yufan Tao, Jiawei Liang, Kai Li1, Yihang Chen	Object Detection Based on YOLO Network	A generalized object detection network was developed by applying complex degradation processes on training sets like noise, blurring, rotating and cropping of images. The model was trained with the degraded training sets which resulted in better generalizing ability and higher robustness.	The experiment showed that the model trained with the standard sets does not have good generalization ability for the degraded images and has poor robustness. Then the model was trained using degraded images which resulted in improved average precision. It was proved that the average precision for degraded images was better in general degenerative model compared to the standard model.
2018	Wenbo Lan, Jianwu Dang, Yang-ping Wang, Song Wang	Pedestrian Detection Based on YOLO Network Model	The network structure of YOLO algorithm is improved and a new network structure YOLO-R was proposed to increase the ability of the network to extract the information of the shallow pedestrian features by adding passthrough layers to the original YOLO network.	The YOLO v2 and YOLO-R network models were tested on the test set of the INRIA data set. The experimental results show that the YOLO-R network model is superior to the original YOLO v2 network model. The number of detection frames reached 25 frames/s, basically meeting the requirement of real-time performance.
2018	Rumin Zhang, Yifeng Yang	An Algorithm for Obstacle Detection based on YOLO and Light Filed Camera	An obstacle detection algorithm in the indoor environment is proposed which combines the YOLO object detection algorithm and the light field camera and will classify objects into categories and mark them in the image.	The images of the common obstacles were labeled and used for training YOLO. The object filter is applied to remove the unconcern obstacle. Different types of scene, including pedestrian, chairs, books and so on, are demonstrated to prove the effectiveness of this obstacle detection algorithm.
2019	Zhimin Mo1, Liding Chen1, Wen-jing You	Identification and Detection of Automotive Door Panel Solder Joints based on YOLO	A method for identifying the solder joints of automotive door panels based on YOLO algorithm that provides the type and location of solder joints in real time. For detecting the small solder joints more precisely, this paper adopts YOLO algorithm which adopts multi-level predictions, predicting on different size feature maps and combining the prediction results to obtain the final result.	The YOLO algorithm, proposed identifies the position of the solder joints accurately in real time. This is helpful to increase the efficiency of the production line and it has a great significance for the flexibility and real-time of the welding of automobile door panels.

Experimental Result & Conclusion: The proposed YOLO algorithm identifies the position of the solder joints accurately in real time. It increased the flexibility and efficiency of the automobile production line along with the welding of automobile door panels in real time.

4. PROS AND CONS

Benefits of using YOLO [1]:

- It is extremely fast compared to other real time detectors which came before it as it uses a Unified Model where the detection is seen as a single regression problem and there is no complex pipeline, just a neural network run on the image.
- It makes less errors than Fast R-CNN as it can see the bigger context because YOLO, unlike Fast R-CNN, can globally reason the image when making predictions. YOLO sees the entire image and encodes some of the contextual information about all classes and their appearance.
- YOLO has learnt generalized representations of objects. YOLO successfully differentiates natural images against art work.

The limitations of YOLO are:

- Small object detection, such as a flock of birds, is a problem as there is a spatial restriction on bounding boxes with each cell being able to predict only two boxes and have one class.
- Problems when generalizing objects of abnormal aspect ratios and configurations.
- Loss function will treat the errors of small or large bounding boxes as same

5. SOFTWARE ARCHITECTURE STYLES FOLLOWED

This paper was conducted as an architecture survey for YOLO models. Here we try to look at the architecture followed by YOLO and also how the basic architecture was modified and used in other areas. A unified architecture style is introduced in YOLO [1], which includes both regression for finding bounding boxes in images as well as finding confidence scores of object classes for each bounding box. The convolutional layers in the base architecture can be said to be of a layered architecture style, with each layer providing functionality only to the layer that follows it. Each of the application-oriented papers that were researched, used the basic architecture along with some modifications, like adding extra layers to the neural network, compressing the number of layers in the network, adding modules for providing additional functionality etc.

6. CONCLUSIONS

In this paper, we have surveyed the YOLO architecture, YOLO network model for object detection, pedestrian detection, obstacle detection and solder joint detection.

- A unified model for object detection which is easy to build and is trained straight on full images. YOLO also generalized well to new domains used in applications that rely on fast, robust object detection.
- A degenerative model built for detecting degraded images like blurred and noisy images has the model being trained with these degraded images. This model performed better in terms of detecting degraded images and coped better with complex scenes.
- For detection shallow pedestrian features, a YOLO v2 network was modified by adding three Passthrough layer to them. The number of detection frames can reach 25 frames/s, which meets the demands of real-time performance.
- To recognize indoor obstacles a new method of using deep learning along with a light field camera was used. The method identifies the obstacles and perceives its information.
- YOLO applied to automobile door panel welding panel lines can identify and detect solder joint accurately. The algorithm can also detect the position of the solder joints and more.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Joseph Redmon,Santosh Divvala,Ross Girshick,Ali Farhadi. ”You Only Look Once: Unified, Real-Time Object Detection” [J].2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016:779788.

[2] Chengji Liu ,Yufan Tao ,Jiawei Liang ,Kai Li1 ,Yihang Chen, ”Object Detection Based on YOLO Network” 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC 2018)

[3] Wenbo Lan, Jianwu Dang, Yangping Wang and Song Wang, ”Pedestrian Detection Based on YOLO Network Model” 978-1-5386-60751/18/$31.00 ©2018 IEEE

[4] Rumin Zhang, Yifeng Yang, ”An Algorithm for Obstacle Detection based on YOLO and Light Filed Camera”, 2018 Twelfth International Conference on Sensing Technology (ICST)

[5] Zhimin Mo1, Liding Chen1, Wenjing You1 ”Identification and Detection of Automotive Door Panel Solder Joints based on YOLO” 978-1-72810106-4/19$31.00 ©2019 IEEE

[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1):98–136, Jan. 2015.

[7] Software Architecture in Practice, 3rd Addison-Wesley Professional ©2012, ISBN:0321815734 9780321815736

[8] Joseph Redmond, Ali Farhadi, ”YOLOv3: An Incremental Improvement”, University of Washington