Study of Entity Detection and Identification using Deep Learning Techniques a Survey

Abhishek Ratnaparkhi
Pimpari Chinchwad College of Engineering, Pune

Sushant Joshi
Pimpari Chinchwad College of Engineering, Pune

Rhushikesh Valiv
Pimpari Chinchwad College of Engineering, Pune

ABSTRACT
Real-time object detection is a recent trend in image processing that plays a very important role in detection of objects and identifying them. Also there are various tools for image processing to identify objects. There are also frameworks which uses end to end network and shows very good results in object detection. However, compared to more accurate but time-consuming frameworks, detection accuracy of existing real-time networks are still left far behind. In this survey paper we have studied different object identification techniques.

In this paper various frameworks like HyperNet, novel CAD YOLO Voxnet are studied. Various methods for object detection and identification like region generation, scale invariant detection, non maximum weighted, Sparse matrix distribution, Background modeling, Speed Up Robust Feature(SURF), Single Shot Detection(SSD) are also studied. R-CNN, Edge detection algorithms and Approach based studies are learned.

Keywords
CNN, deep learning, object detection, object tracking, object identification, edge detection.

1. INTRODUCTION
When we're shown an image, our brain instantly recognizes the objects contained in it. On the other hand, it takes a lot of time and training data for a machine to identify these objects. But with the recent advances in hardware and deep learning, this computer vision field has become a whole lot easier and more intuitive. We are constantly in search of methods to have a 'detection' or 'recognition' system as powerful as the human being. Some of the main tools and techniques studied are mentioned below.

1.1 YOLO
You Only Look Once[4] is good approach to object detection where they find object detection is a regression problem to separate the bounding boxes and relate the class assumption. a neural network can identify boxes and class probabilities directly from image. It can optimize directly from images YOLO is very fast in performance it can process images in real time at 45 frames per second compared to state of the art detection system YOLO makes more errors but it can perform better than DDM and R-CNN.

Idea of Yolo
It is more similar to FCNN (fully convolutional neural network and passes images once through FCNN and output is prediction. This architecture is spitting the image in grid and for each grid there are 5 bounding boxes and class probability for those bounding boxes. Object detection is a regression problem right from pixel to bounding box coordinates and class probability. Yolo trained on full image and can directly optimize the detection time. This model has many benefits over traditional method of object detection. Our base network runs at 45 frames per second with no batch processing on a Titan X GPU and a fast version runs at more than 150 fps. This means we can process streaming video in real-time with less than 25 milliseconds of latency YOLO sees the entire image during training and test time so it implicitly extract particular contexts information about classes. YOLO makes less background errors compared to Fast R-CNN.

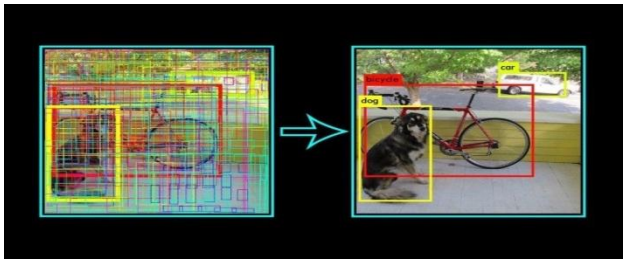

Fig 1: Image Reference (https://purplemaia.org/intern-update-kylyn-takes-on-machine-learning-for-fish-id/)

1.2 VoxNet
Its a 3D convolutional neural network for object detection. object detection is not easy for autonomous robot operating in the real world. VoxNet[7] is an architecture to tackle 3D information large amount of point cloud data by interacting a volumetric occupancy grid representation with supervised 3D-CNN. it evaluate approach of publically benchmarks using range sensors like LiDAR and RGBD CAD data. VoxNet archives accuracy beyond the state of the art while it label instantly

1.3 SqueezeDet
[6]It is a unified, small, low power fully convolutional neural networks for real time object detection for autonomous driving object detection is critical task for autonomous driving for achieving high class accuracy to ensure safety object detection for autonomous driving also require real time inference to guarantee vehicle control In this frameworks SqueezeDet propose a fully CNN for Object detection that goals with solving simultaneously above constraints. In this network they use convolutional layers not only extra features map but also an output layers to compute bounding, boxes and class probabilities.

The detection pipeline of this model only contain a single fast forward of neural network thus its extremely fast. this model is fully convolutional which leads to small model size but energy efficient. ultimately its shows that the model is very accurate achieving state of the art on KITTI dataset.

Multiple common object classes are contained in the given collection of images, which means this is totally unsupervised problem. Multiple objects or even no objects is contained in

some of the images. The project aims to incorporate state-of-the-art technique for object detection with the goal of achieving high accuracy with a real-time performance. A major challenge in many of the object detection systems is the dependency on other computer vision techniques for helping the deep learning based approach, which leads to slow and non-optimal performance. In this project, we use a completely deep learning based approach to solve the problem of object detection in an end to-end fashion. In this project, we will understand what object detection is and look at a few different approaches one can take to solve problems in this space. Then we will deep dive into building our own object detection system in Python.

2. LITERATURE SURVEY

Joseph Redmon [4] has used YOLO, a new approach to object detection they framed an object detection as a regression problem to separated bounding boxes and associated class probabilities. This architecture is extremely fast, it can be optimized end-to-end directly on detection performance. YOLO can processes images in real-time at 45 frames per second It can separate object detection into a single neural network. It also can predict bounding boxes in all classes for an image The YOLO design have high average perfection.

This system device input image into S*S grid cells if object centre is fall into the grid cell that particular grid cell is responsible for that object detection. Each grid cell has bounding boxes and confidence measurement of that boxes that confidence measurement reflect into how the model id confident about box containing an object and also about the accuracy. The confidence define as Pr(Object) * IOUtruth pred. If object find in that cell then score of that confidence should be zero. Otherwise confidence score to equal the intersection over union (IOU) between the predicted box and the ground truth.

Each box can consist of 5 predictions v, w, x, y and confidence The (v,w) coordinates represent the centre of the box relative to the bounds of the grid cell. Width and height can predict relative to images ultimately the confidence prediction represents the IOU between the predicted box and any ground truth box.

Each and every grid cell also predicts conditional class probabilities i.e(Class or Object). These probabilities are apply conditions on the grid cell containing an object. they only predict one set of class probabilities per grid cell, unconcern of the number of boxes B.

At test time they multiply the conditional class probabilities with the individual box confidence predictions.

Bichen Wu [6] propose SqueezeDet, a fully convolutional neural network for object detection that goal is to simultaneously satisfy high accuracy to make sure safety, object detection for autonomous driving also requires real time achieving speed to guarantee good vehicle control, as well as small model size and energy efficiency and effectiveness. In this work they use convolutional layers not only to get feature mapping, but also as the output layer to computing the boxes and class possibility. Detecting mechanism of this model contains single neural network so its very fast. this model is completely convolutional, which gives to small model size and good energy efficient. Ultimately, This model is very accurate experimentally and achieving state of the art accuracy on the KITTI[6] benchmark.

1. Speed. The detector should have real-time or faster interpretation speed to minimize the latency of control on vehicle.
2. Accuracy. object the detector should achieve fully recall with high definiteness on objects of interest.
3. Small model size. As proposed smaller model size leads to benefits of more efficient distributed management, Less energy consumption and more feasible embedded system deployment.
4. Energy efficiency. embedded processors which will targeting automotive market must suitable within a very smaller power and energy envelope so that the all above issues author proposed SqueezeDet, a fully convolutional neural network for object detection.

The detection pipeline of SqueezeDet is inspired by its features like small and effective as well as energy efficient. First, authors used stacked convolution filters to get a high dimensional, low resolution map for the input image. The author used ConvDet, as a convolutional layer to take the feature map as input and compute a large amount of object bounding boxes and identifying their categories.

Eventually they gets filter out these bounding boxes to obtain final detections. The backbone of this entire process i.e convolutional neural net (CNN) architecture of this network is SqueezeDet which achieves best level image accuracy with a model size of less than 5MB that can be compressed to 0.5MB. After modifying the SqueezeNet model with additional layers leading by ConvDet, the total model size is less than 8MB which is great achievement itself. The interpretation speed of this model can reach 57.2 FPS (frames per second) with input image resolution of 1242x375. Advantage is from the small model size and activation size, SqueezeDet has a much smaller memory and requires less DRAM accesses, so it consumes only 1.4J of energy per image on GPU, which is about 84 times lesser than a Faster R-CNN model that described in this paper . SqueezeDet is very accurate. One of the trained SqueezeDet models achieved the best average correctness in all three difficulty levels of detection in the KITTI object detection challenge.

Daniel Maturana [7] addresses the problem of predicting an object class label given a 3D point cloud segment, which may include background processing he proposed VoxNet, a framework to detect this problem by integrating a volumetric Occupancy Grid representation with a supervised 3D Convolutional Neural Network (3D CNN). authors evaluate approach on publicly available benchmarks using LiDAR, RGBD, and CAD data. VoxNet achieves perfect accuracy beyond the state of the art while labeling hundreds of frames per second.

Range sensors such as LiDAR and RGBD cameras are choice of sensor for now a days automotive vehicles, including cars, Aeroplanes and drones. this sensors are used for detecting barrier and understanding environment.

The main contribution of this paper is VoxNet, a basic 3D CNN architecture that can be applied to create fast and accurate object class detectors for 3D point cloud data

In the experiments, this architecture achieves state of-the-art accuracy in object identification tasks with three individual sources of 3D data such as LiDAR point clouds, RGBD point clouds and CAD models. The input to this algorithm point cloud segment which can originated from sliding box The segment gives intersection of a point cloud with a bounding box and it may include background processing .this system for this task has two main components: a volumetric grid

representing this estimate of occupancy and 3D CNN that predicts a class directly from the occupancy grid.

Volumetric Occupancy Grid
Occupancy grids represent the state of 3D lattice of random variables and maintain a probabilistic estimate of their occupancy as a function of incoming sensor data .

There are two reasons that author use occupancy grids.

It allows us to efficiently estimate free and occupied and unknown space from range measurements, even for measurements coming from different viewpoints and time instants. This representation is richer than those which only it considers occupied space against the free space such as gap between free and unidentified space can be valuable shape

They can be stored and manipulated with simple and efficient data structures. In this work, we use dense arrays to perform all our CNN processing, as we use small volumes and GPUs work best with dense data. To keep larger spatial extents in memory we use hierarchical data structures and copy specific segments to dense arrays as needed. on paper this allows to store a unbounded volume while using small occupancy grids for CNN processing.

Tao kong [1] proposed a framework for object detection and identification. It is based on Hyper Feature which integrates all the feature maps and then collects them in a uniform space so that object detection and identification gets faster than the other methods. Hyper net is hierarchical network method. It is very fast it operates at the speed of 5fps on GPU. due to its speed efficiency it used in real time object identification.

Hypernet first identifies the regions where objects are probably located by using R-CNN. It first extracts 2k region proposal with the help of selective search method. these are separated using CNN algorithm. Deeper CNN algorithm model(VGG16) provides 30% better improvement over the previous method on PASCAL VOC 2012 database.

Hypernet is advancement for problem of Fast R-CNN.it takes some amount of time to run for generating region proposals. Also Fast R-CNN may not work well with the small objects because of coarse pixel map size[1].

Hypernet gives solution for this problem it uses Fast R-CNN and combines region proposal and detection in unified network. on top of traditional ConvNet they add to convolutional layers to get output compute proposals and share features with Fast R-CNN. Author has referred PASSCAL VOC207 and PASSCAL VOC 2012 Datasets[20].

Huajun Zhou[5]- proposed unique scheme for object detection. It is a complete framework. It is much correct than the last networks. It additionally manages to keep up high speed results and determine the object in the sight. The CAD framework has introduced maxout[5] property to approximate the result between pixel and network prediction.

It uses R-CNN to spot the region of interest and so classifies these regions by multiple two class SVMS. Moreover It uses SPP-Net[5] , It helps to reshape region of interest instead of wrapping them to original size. It reduces computation size and increases the speed of detection of an object by implying spatial pyramid pooling(SPP).

Scale Invariant Detection is used in CAD end-to-end framework[5]. It indicates that detectors must produce the same prediction from different image patches multiple Detectors are trained to distinguish whether the maximum jaccard overlaps between d- different size patches and the ground true boxes are greater than the given threshold.

Chia-Hung[3]- proposed a system using hysteresis thresholding and motion compensation which construct spatial and temporal compensations respectively. This system employed to identify entity in video patrolling. In camera patrolling background modeling could be a core element that is employed to extract moving entity . It is nothing but object detection and moving object in this technique is foreground object.

They used CAVIAR dataset Wallflower dataset and CDNET, which have been widely used to assess experimental results in the area of video surveillance research.

System work in following manner:

1. Color and Texture Background Modeling.
2. Hystheresis Thresholding.
3. PCT combination.
4. Shape Mending.

1. Color and Texture Background Modeling
In many cases, moving objects in video usually come with Noise. In this situation, it will be a compensatory process leading to many false positive results. Therefore, we are targeting first Shadow reduction in Color and texture background modeling. Here two background modeling techniques used

i. GMM (Gaussian Mixture Model)
ii. Proposed texture background modeling

2. Hystheresis Thresholding
In most background modeling methods, the quality of The foreground object is highly dependent on a certain range. To determine whether the pixel falls within the range to determine whether the pixel falls within the range of the background distribution or not. Unfortunately, choosing the appropriate range for different frames are very impractical. to solve this problem, author constructed both color and texture background modeling methods.

3. Predominant Color and Texture Result s(PCT) Combination
As mentioned above, author have designed several foregrounds Various images with color and texture information. To capture all the benefits of each Background modeling method to reduce errors of In the preceding pictures, author propose the mixed approach described In the following paragraph. Because the first step is combining predominant color(PC) and predominant texture(PT) images. Predominant color and predominant texture contain very important information. In addition, Since shadows do not affect the predominant texture image, the combination predominant color and predominant texture can remove shadow on moving objects and Give a clear structure of skeleton. The combination begins with predominant texture. For the white pixel P in the predominant color, if there is any white pixel of the predominant color The neighboring p in the 3x3 range, is added to the pixel p The result of the combination, also known as the joint main color And texture result. This process also applies predominant color is processed after predominant texture. In other words, inside the PCT, Any white pixel in the predominant texture is at least one white pixel On predominant color, and vice versa. show predominant color and predominant texture Images, and result shows the convergence Result with morphology and connected

component.

4. Shape Mending

After the predominant color and Texture is obtained here shape mending process plays an important role in bounding box. because the cavities in the front object are obvious. The shape-mending approach uses consecutive supplementary color(SC) images Sorted according to their respective limitations in Descending order.The image was originally enhanced by SC1, this time a white pixel SC1 is added to the PCT if it is a neighbor of any white pixel In PCT in block 3x3. The result is called PSCT1. The The process is repeated and produces more PSCT results Until all SC images are used. Three-dimensional compensation method is proposed in this paper To detect objects in surveillance systems. First we Use a texture background modeling method that only detects The texture of the foreground object but can resist light Transformation and shadow interference. Second, we apply hysteresis Threshoulding over both texture and color background models Generate major and complementary images. The combination of key images shows the skeleton Moving objects, PCT. Finally The proposed motion history applies to spatial-temporal information To reduce cavity and fragment problems in the foreground

Objects. In this way the combined approach provides a three dimensional Compensation by texture, color, and leveraging

Spatial-temporal information.

Binoy B Nair [2] has proposed a system which is a combination of deep learning and stereovision for detecting an object and measures distance of next object. It gives labels to the entity. It uses framework convolutional neural network to find presence of entity and identify the entity which present in the range of fixed stereo camera. It uses 3D point cloud method to estimate the distance between camera and entity.

They developed a system which discover the presence of any hurdle within the camera vision region and classify the type of hurdle.

The previous method referred by the author uses Alexnet[2]. Where Alexnet can only tell what is the object, It doesn't tell where the object was and distance between object and camera. In traditional image processing algorithm having their cons. So they made an e a system which is confined system performs the task of detection of obstacle as well as distance estimation. This proposed system has been used on Nvidia® Jetson TX1 board using a Zed® stereo camera for detecting and estimating distance of obstacle which is in vision of camera.

Yunhang Shen [9] proposed a novel scheme to perform weakly supervised object localization, termed object-specific pixel gradient (OPG). The OPG is trained by using image-level annotations alone, which performs in an iterative manner to localize potential objects in a given image robustly and efficiently. In particular, we first extract an OPG map to reveal the contributions of individual pixels to a given object category, upon which an iterative mining scheme is further introduced to extract instances or components of this object. Moreover, a novel average and max pooling layer is introduced to improve the localization accuracy. In the task of weakly supervised object localization, the OPG achieves a state-of-the-art 44.5% top 5 error on ILSVRC 2013, which outperforms competing methods, including Oquab et al. and region-based convolutional neural, networks on the Pascal VOC 2012, with gains of 2.6% and 2.3%, respectively. In the task of object detection, OPG achieves a comparable performance of 27.0% mean average precision on Pascal VOC 2007. In all experiments, the OPG only adopts the off-the-shelf pretrained CNN model, without using any object proposals. Therefore, it also significantly improves the detection speed, i.e., achieving three times faster compared with the state of the-art method.

Guanbin Li.[8] address problem by developing a Cross-Modal Attentional Context (CMAC) learning framework, which enables the full exploitation of the context information from both RGB and depth data. Compared to existing RGB-D object detection frameworks, our approach has several appealing properties. First, it consists of an attention-based global context model for exploiting adaptive contextual information and incorporating this information into a region-based CNN (e.g., Fast RCNN) framework to achieve improved object detection performance. Second, our CMAC framework further contains a fine-grained object part attention module to harness multiple discriminative object parts inside each possible object region for superior local feature representation. While greatly improving the accuracy of RGB-D object detection, the effective cross-modal information fusion as well as attentional context modeling in our proposed model provide an interpretable visualization scheme.

3. RESULT COMPARISON

Method	**Input Resolution**	**Accuracy MAP (mean average precision)**	**Speed FPS(Frames per second)**
YOLO	448*448	71.8	45
SSD	512*512	76.8	19
R-CNN	1000*600	63.4	7

In the above table comparison of Yolo, R-CNN and Single Shot Detection (SSD) is given on the basis of mAP and FPS. All the techniques are tested on PASSCAL VOC 2007 and PASSCAL VOC 2012 datasets.

4. CONCLUSION

We have studied different image processing frameworks techniques, algorithms for detection and identification of a objects in image. Framework such as Voxnet, SqueezeDet, HyeperNet, Yolo, R-CNN, Hysteresis Thresholding algorithms are successfully studied. In each of the techniques mentioned above efficiency of system to the previous one in aspects of runtime and detection accuracy is increased. Yolo is the most efficient Technique in the current entity detection trend of image processing.

5. FUTURE SCOPE

The above defined object detection algorithm successfully tracks objects in consecutive frames. Our tests in sample applications show that using nearest neighbor matching scheme gives promising results and no complicated methods are necessary for whole-body tracking of objects. Also, in handling simple object occlusions, our histogram-based correspondence matching approach recognizes the identities of objects entered into an occlusion successfully after a split. However, due to the nature of the heuristic we use, our occlusion handling algorithm would fail in distinguishing

occluding objects if they are of the same size and color. Also, in crowded scenes handling occlusions becomes infeasible with such an approach, thus a pixel-based method, like optical flow is required to identify object segments accurately.

6. REFERENCES

[1] Tao Kong, Anbang Yao, Yurong Chen and Fuchung Sun. "HyperNet: Towards Accurate Region Proposal Generation and Joint Object Detection". in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016.

[2] Rahul and Binoy B Nair." Camera-based Object Detection, Identification andDistance Estimation" International Conference on Micro-Electronics and Telecommunication Engineering 2018.

[3] Chia-Hung Yeh, Chih-Yang Lin, Kahlil Muchtar, Hsiang-Erh Lai and Ming-Ting Sun." Three-Pronged Compensation and Hysteresis Thresholding for Moving Object.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick and Ali Farhadi." You Only Look Once: Unified, Real-Time Object Detection". in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016.

[5] Huajun Zhou, Zechao Li, Chengcheng Ning and Jinhui Tang." CAD: Scale Invariant Framework for Real-Time Object Detection" ICCV 2017

[6] Bichen Wu, Peter H. Jin, Kurt Keutzer and " SqueezeDet: Unified, Small, Low Power Fully Convolutional Neural Networks for Real-Time Object Detection for Autonomous Driving" in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[7] Daniel Maturana and Sebastian Scherer. " VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition" 2015.

[8] Guanbin Li, Yukang Gan, Hejun Wu, Nong Xiao and Liang Lin. "Cross-Modal Attentional Context Learning for RGB-D Object Detection"in proceedings of IEEE Transaction on Image Processing 2018.

[9] Yunhang Shen, Rongrong Ji, Changhu Wang, Xi Li, and Xuelong Li. "Weakly Supervised Object Detection via Object-Specific Pixel Gradient" in proceedings of IEEE Transaction On Neural Network And Learing system 2018.

[10] Z. Shi, T. M. Hospedales, and T. Xiang, "Bayesian joint modelling for object localisation in weakly labelled images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 10, pp. 1959– 1972, 2015.

[11] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Image co-localization by mimicking a good detectors confidence score distribution," in Proceedings of the European Conference on Computer Vision. Springer, 2016, pp. 19–34.

[12] D. Li, J.-B. Huang, Y. Li, S. Wang, and M.-H. Yang, "Weakly supervised object localization with progressive domain adaptation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3512–3520

[13] M. Cho, S. Kwak, C. Schmid, and J. Ponce, "Unsupervised object discovery and localization in the wild: Part-based matching with bottomup region proposals," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1201–1210.

[14] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, "Co-localization in realworld images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1464–1471

[15] R. Gokberk Cinbis, J. Verbeek, and C. Schmid, "Multi-fold mil training for weakly supervised object localization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2409–2416.

[16] Javeria Farooq."Object Detection and Identification using SURF and BoW Mode".978-1-5090-1252-7/16/$31.00 2016 IEEE

[17] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell.,37:1904–1916, 2014.

[18] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NIPS, 2015.

[19] S. Bell, C. L. Zitnick, K. Bala, and R. B. Girshick. Insideoutside net: Detecting objects in context with skip pooling and recurrent neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2874–2883, 2016.

[20] R. Girshick. Fast r-cnn. In *ICCV*, 2015.