Label Space Context Network For Small Object Detection

A Sheetal Reddy, [1] N Dinesh Reddy[2] and K. Madhava Krishna[1]

***Abstract*— Numerous methods address the problem of scene understanding for robot navigation. Object detection has been at the pivot for robot navigation and scene interpretation compared to other scene understanding methods. Although current object detection algorithms show state-of-the-art accuracy on multiple datasets, they fail to accurately detect small objects this can be attributed to the features trained on large regions of the image. Small object detection plays an important role in robot navigation, scene understanding etc as most of the objects observed are generally small until observed from close. We use a context network based method to improve the accuracy of small object detection as the surrounding objects play an important role in detection of objects. This is close to mimicking human behaviour where we search for small objects in the vicinity of larger related objects. We formulate this as a label space learning method compared to an end-to-end learning algorithm. We use the state-of-the art YOLO network and train a context network as an auxiliary task for better localization of the small objects in an image. We show improvements on the object detection accuracy for small objects not detected byt the YOLO method.**

I. INTRODUCTION

Object detection has been a widely studied field of computer vision. Its one of the basic problem of perception and has wide applications in scene understanding, robotics and recognition tasks. Although there has been substantial progress in object detection for the past few decades, detecting small objects has not been a widely addressed problem. One of the reasons for this can be attributed to the fact that object detection so far was not enjoying high accuracy on the available datasets. With the dawn of the deep learning era object detection has become a well studied problem with high detection accuracy. This has propelled us to study and solve detecting of small objects as its a important requirement for robot navigation.

Detecting small objects is a challenging task due low visibility of the object in the image. The size of the object makes the detection a challenging task and direct feature learning based methods do not show great accuracy on these tasks. We propose to solve the small object detection as a scene understanding problem. The underlying principle behind our formulation is that objects co occur in an image and exploiting this information can boost the detection accuracy i.e. using contextual information for improving the detection accuracy. There have been multiple methods which have previously exploited contextual information in scene understanding, object detection and other problems but none of the methods specifically formulate the task for small object detection and specifically as a label space learning problem. The intuition behind the label space learning compared to an end-to-end learning is that the method can be incorporated with any object detection or semantic segmentation pipelines for small object detection. Further small objects generally have a relation with larger objects in the vicinity and learning this in label space gives higher generality compared to learning it end-to-end. The label space compatibility is learned using a neural network based filter learning.

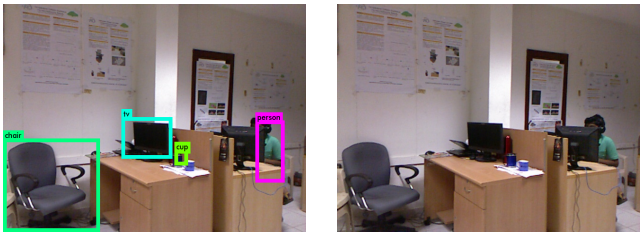

Fig. 1. The figure depicts the output of object detection pipelines(LEFT) compared to our small detection pipeline(RIGHT)

We build a framework which can be incorporated on top of any trained object detection algorithm and can fine-tune the network to detect small objects. The underlying principle is that object detection scores for small objects are inherently calculated by the network but due to less confidence in the detection most of the detection are not used. we come up with a context based label similarity network to boost the confidence of these small detection. That is for example we have multiple hypothesis of a keyboard in the image but the general object detection algorithms reject the proposals due to their less likelihood. Our context network boosts the label probabilities for keyboard which are around desktops and mouse by learning the relation between these co occurring labels. This is inherently true for all small objects, they are always associated with another larger object which gets good detection accuracy. We build a context based network in the probabilistic label space of the YOLO detection output and show improvements in the accuracy of object detection.

The main contributions of the paper are:

- A novel pipeline for small object detection for robot navigation
- Label space learning for context network boosting in re usability of the weights for other object detection and semantic segmentation pipelines.
- Ground truth annotations for small objects in cluttered environments for evaluation with other methods

[2] with affiliation to Robotics Institute, Carnegie Mellon University, Pittsburg, USA

[1] with affiliation to RRC, International Institute of Information Technology, Hyderabad, INDIA

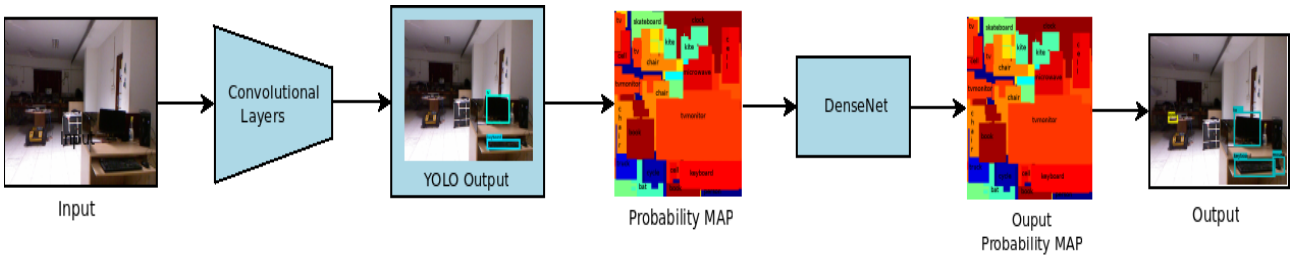

Fig. 2. We demonstrate the pipeline for the detection of small objects. The input is passed through a convolution network which was pretrained on COCO dataset. These detections are converted into class probabilty maps and passed through a contextual network to learn the relations between objects.

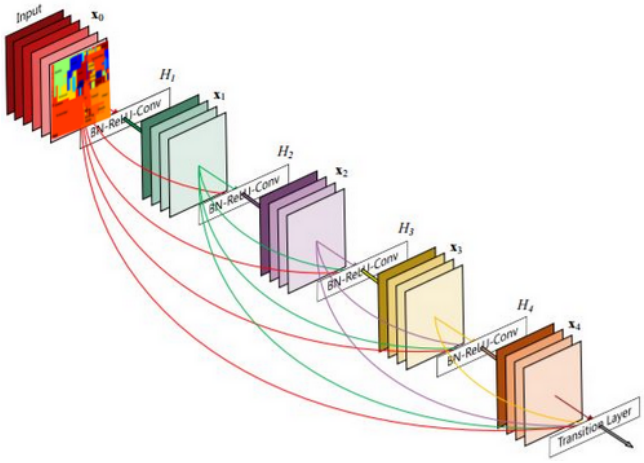

Fig. 3. Dense net Description

II. RELATED WORK

Vast amount of object detection pipelines have been proposed to detect objects from images directly by running a classifier on regions of images [1]. object detection has been a widely addressed problem and multiple datasets [2] [3] [4] have been available to address the problem. Plethora of methods [5] [6] [7] [8] [9] [10] have been proposed for addressing the problem of object detection and show significant accuracy on the current challenging datasets. Most of the above methods run a classifier on a region of a image to detect and object. These methods show a progress in the object detection algorithms and show significant improvement in accuracy. The methods proposed don not incorporate any contextual information for object detection.

There has been a significant amount of work incorporating context as a key for object detection. Some of the methods [11] [12] [13] [14] [15] [16] [17] [18] [19] use context information for improving the accuracy of object detection and recognition. Most of the above methods exploit the object relations in the real world to classify the objects in the image space. Our method is an extension of these methods but has a novel learning paradigm and can detect small objects in images. Compared to the above methods we propose a label space learning of the contextual information instead of learning it from the weights.

All of the above methods have been addressing the problem of object detection as a bounding box classification which uses a sliding window based method to classify. Recent end to end training methods [20] for object detection using neural networks have shown state-of-the-art results on multiple datasets. [20] propose a method to predict the objectness score and classification score and combine them to give the object detection score. This method is fast and trained end-to-end and gives accurate results on many object detection tasks. But it fails to detect small objects because of its grid based prediction pipeline. We propose a method to extend this method to small object detection by exploiting the contextual information. We convert the output of the current learned pipeline to a label probabilistic space and learn the label relations.

Small object detection has not been a widely addressed problem in the vision community. This can be attributed to the fact that object detection has not shown great accuracy on the current until the recent advances in neural networks. Earlier work on small object detection is mostly about detecting vehicles using hand-crafted features and shallow classifiers [21]. A considerable amount of work has been done in searching for small objects particularly in large rooms. [22] gives a solution for search and localization of objects using a monocular camera with zooming features to overcome the problem of low resolution images of small objects from far away distance.

[23] proposes a VOP framework to identify discriminative viewpoints to recognize small objects having distinctive features only in specific views. [24] addresses small object detection problem by augmenting R-CNN []algorithm by proposing their own regional proposal generator. R-CNN and its variants use region proposals instead of sliding windows to find objects in images. The region proposal generator captures the objectness of small objects and gives small set of proposals compared to RCNN and its variants but their approach doesn't claim to give a real time performance. RCNN takes more than 40 seconds for testing a single image.

System Overview: We describe the process of context Probability Map III. We describe the object detection of sematic segmentation pieline converion to class probabilty map in sec III-A. We train the context network using deep learning architectures as described in section III-C . The results and evaluations comparing our method are shown in section IV.

III. CONTEXT NETWORK FOR SMALL OBJECT DETECTION

Small object detection has been a challenging problem. There has been plethora of work attempting to solve this problem. We address the small object detection as a context exploitation problem. Using contextual information of an image improves the likelihood of small objects is an interesting direction. The network we propose learns the relationship between object classes using neural networks and enhances the probability of small object detection. The probability maps generated from the object detection network is passed into a deep network for learning contextual information. We describe the pipeline and different modelues of our method.

A. Object detection Network

Most of the detection systems re-purpose classifiers or localizers to perform detection. These models apply the model at different image locations and scales. Some recent methods apply a single neural network to the full image. These network divide the image into regions and predicts bounding boxes and probabilities for each region. These bounding boxes are then weighted by the class probabilities. Both the pipelines have a inherent computation that each pixel of the image is classified into a class probability. These class probabilities are converted to bounding box predictions using an objectness and class probability scores. we follow the architecture of YOLO [20] that learns to predict the objectness score and class probabilities. in this approach the image is divided into multiple fine grids and then each grid computes a 3 bounding boxes which have the grid within them. The algorithm further computes the class probability of each grid cell. These class probabilities and bounding box scores are concatenated to give the final bounding box for each image. To create the context network, we take these class probabilities from the grid space and compute a per pixel class probabilities. The computation of the class probabilities is explained in the following section.

B. Class Probability Map

This section is going to brief on conversion of the yolo output to class probability maps. The architecture of the object detection divides the input image isn't SxS grid, and if the center of the object falls into a grid cell, then that grid cell is responsible for detecting the object. So each grid cell predicts B bounding boxes and confidence scores for the boxes. Simultaneous we have each grid cell predicting C conditional class probabilities. These probabilities are conditioned on the grid cell containing an object. Out pipeline uses the class probabilities and the bounding box predicted for each grid and creates a per pixel class probability map. The final max value of per pixel class probabilities using the per grid class probabilities and bounding boxes is depicted in the image 4. For each pixel in the image we sum the class probabilities over each bounding box the pixel lies in:

$$P(c, x_i) = \sum_{b \in B_{x_i}} (P_b(c, x_i)) \quad (1)$$

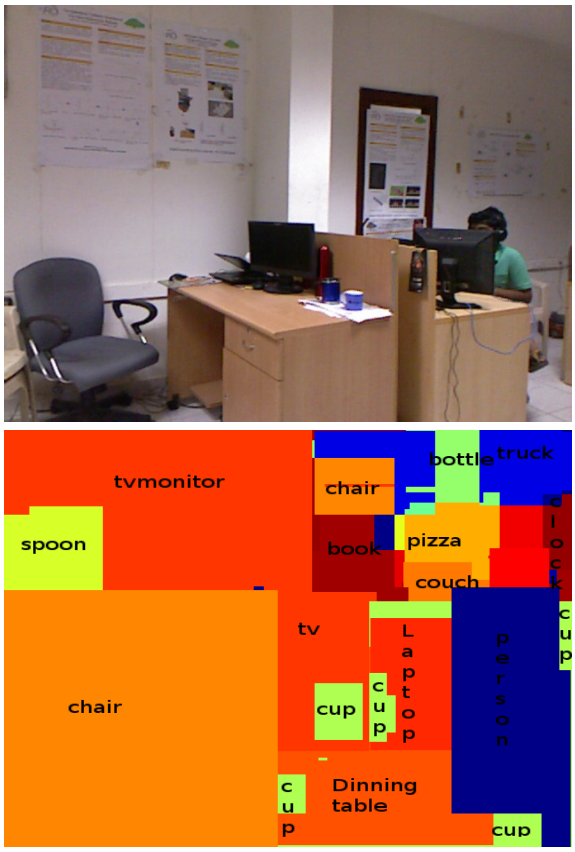

Fig. 4. We show the image of conversion of the output of object detection pipeline to a class probability map. The top image is a input image and the bottom image is a max of the class probability maps. We label the image with respect to the class of the object for better understanding of the image. (best viewed in color)

Where $P(c, x_i)$ is the probability of class c for x_i pixel of the image. B_{x_i} are all the bounding boxes which have pixel x_i. $P_b(c, x_i)$ is the probability of class c for pixel x_i. This produces a vector of $vxmxc$ where n and m are the height and width of the image and c is the number of classes. We pass this as input to a context network to learn the per pixel relationship between object classes. This per pixel context network can be used with any object detection pipeline for small object detection.

C. Context Network:

The context network learns the relationship between the objects in an image and tries to improve the detection of related objects. We propose a densenet based learning in label space. There are two components of this network we would like to point our. The network should be able to learn the relation between multiple classes in the image and it should simultaneous need to retain the labels learned from the earlier object detection pipelines. We have experimented with multiple architectures to get this element of the network working and found that the current architecture best exploits these constraints.

We assume that each pixel in an image has a d-dimensional feature vector f_i. Learning the context information using

a Densnets. A standard CNN [25], applies a non-linear transformation H_l to the input layer as $x_l = H(x_{l-1})$. Similarly the extension of this was proposed by Resnet [26] by introducing a residual block so the resulting output was given as $x_l = H(x_{l-1}) + x_{l-1}$. The recent DenseNets [27] have further extended the architecture by concatenating multiple block features and can be given as:

$$x_l = H_h(x_{l-1}, x_{l-2}, ..., x_0) \quad (2)$$

Where x_l is the output of densenet in the l^{th} layer. We believe having these feature concatenate layers is specifically useful for label space learning as the simple cnns might lose the previous information because of less number of relations learned compared to number of parameters of the network. The output of the network is a similar d-dimensional feature vector with class probabilities. We apply a softmax layer to the current output and train the network with smaller object detection groundtruth. We use ADAM to train the current method.

D. Small object detection

We convert the learned probability map from object class probability to object detection scores. The probability map has a learned representation for small object detection. we use the objectness score from the object detection pipeline and combine the object class probability from the learned context network to give the score of small objects. We follow the framework of [20] for combining the objectness score and the probability map. Although the framework uses grid based probability it can be scaled to pixel based probability. The main difference between the proposed method and our method is the thresholding. We give the detection based on the size of the object. To classify the object as detected we scale the threshold of detection based on the size of the number of pixels the class probabilty occupies in the bounding box.

IV. EVALUATION AND RESULTS

A. Dataset:

We compose our small object detection dataset with a set of images from Microsoft COCO [2] dataset. Microsoft COCO is a dataset containing images in the wild with annotations for groundtruth object detection and semantic segmentation. It further has information on the instance segmentation as well for better detection. We further have hand labelled our own small object detection dataset for in an indoor environment with multiple support structures and small objects, we will here forth refer to this dataset as LAB dataset. These scenes are challenging as they contain many small objects with cluttered backgrounds unlike PASCAL VOC dataset or COCO dataset. These images are taken from the same indoor background at different time instances with different locations for small objects to generalize our detection. We select 6 object categories from all the classes which we consider as small objects. The object categories that are selected are Monitor, Keyboard, Mouse, cup, bottle, Laptop, book. Manual annotation of each scene from the LAB dataset with the 6 classes is performed. We have classified the dataset into 600 training examples and 100 testing examples to evaluate our algorithm. We will be making the dataset available to boost the research in the direction of small object detection as its a very important problem for mobile robot navigation.

Method	Small Objects	Large Objects
YOLO	10.4	60.3
Our Method	40.2	60.3

TABLE I

WE DO A QUANTITATIVE COMPARISON OF OUR METHOD OVER THE YOLO METHOD FOR OBJECT DETECTION. WE SHOW THE ACCURACY OF OUR METHOD

B. Qualitative evaluation

We observe that detecting small objects is not a trivial task for vision algorithm. To show the results of small object detection, we first finetuned the YOLO network on the small object detection dataset and have observed that finetuning the network for multiple small objects does not scale to detecting small objects. In figure 5 we show the results of the YOLO network augmented with small object detection. We furthur observed that by finetuning the YOLO we were able to get small object probabilities in our class probability map but the confidence of the class probability was low due to multiple instances of the object in different background. This motivated us to learn the label space based probability map to bosst these low probability maps of small objects surrounded by co occurring objects. In figure 5 sequence 1 class probability map shows the detection of two mugs on the table but the yolo output only detected one mug because of low confidence score. Our method was able to boost the confidence of the mun because of the surrounding labels being table and resulted in accurate prediction of mug on the table. We observe multiple such instance in the sequence which resulted in such boosts in detection accuracy.

C. Quantitative evaluation

to compare our method quantitatively with other detector algorithms. We calculate the mAP of our detection comparison with YOLO. mean Average precision is defined as $mAP = \sum_{i=1}^{N} \sum_{k=1}^{n} P(k)/min(m,n)/N$ where P(k) is the precision values of the bounding box predictions for the range of k=0 to k=1. This is a standard metric used by multiple methods for object detection accuracy compilation. We follow the method on the test sequence of the dataset and find that our method performs better than YOLO in detecting small objects. This can be attributed to the fact that contextual network has successfully learned the relation between objects boosting in accuracy for smaller objects.

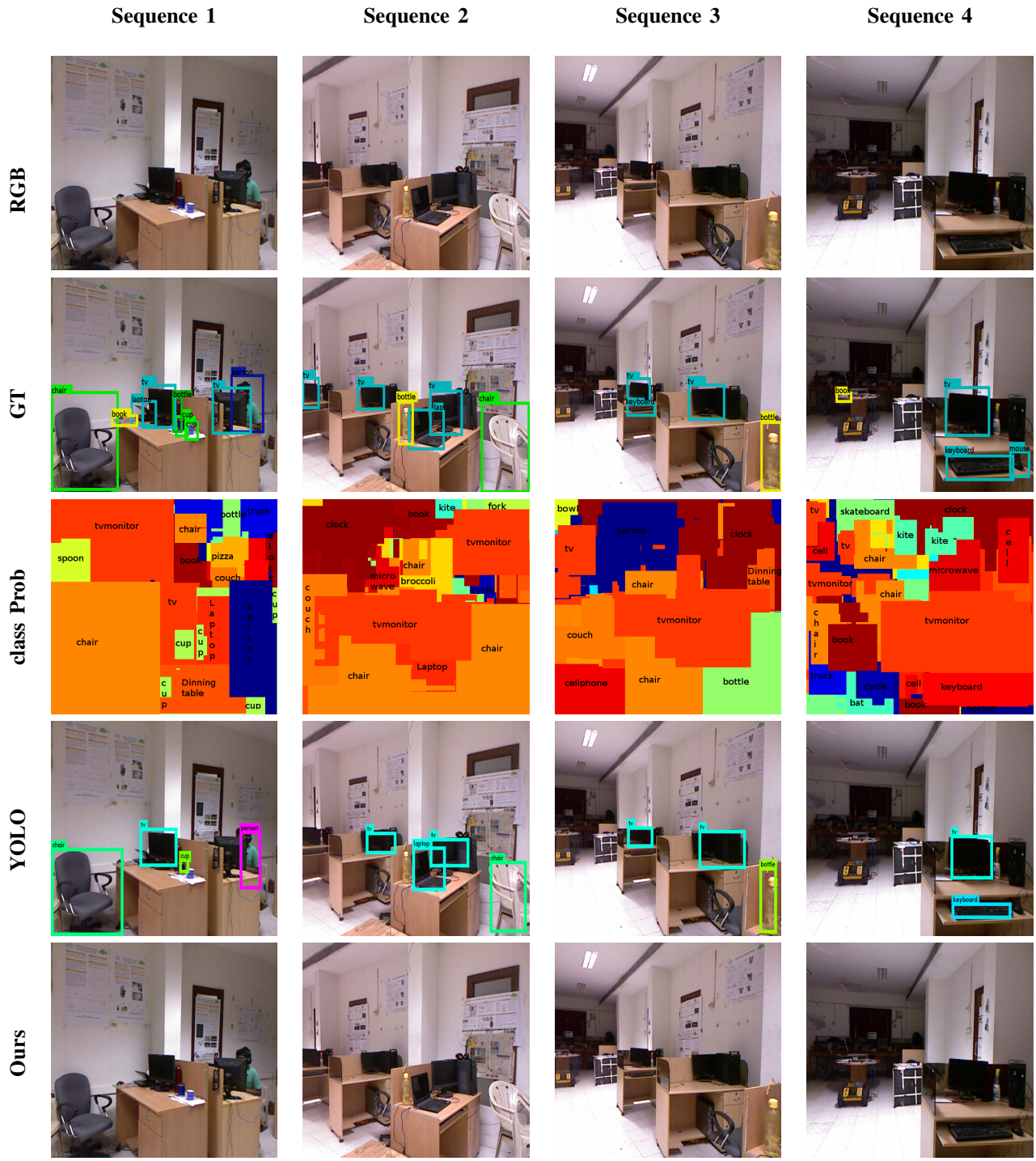

Fig. 5. Qualitative evaluation of joint labels with Ground Truth annotations on LAB dataset. Top to Bottom: (1) Input image from LAB sequence (2) Ground Truth for Object detection with augmented small object detection (3) The results of the conversion from context YOLO to class probability network (4) (5) (Best viewed in color)

For larger objects the mAP remains the same due to high likelihood of the objects probability in the maps. These are the two main insights for using a label space probability map instead of trying to learn small object detection from scratch.

we further have tested different architectures to learn the context map. The CNN based context maps could not successfully retain the information from the previous probability maps. The current dense architecture is the most suitable architecture for label space learning problems because of its multi layer feature maps concatenation method helping in retain the information learned by YOLO architecture. This has been analytically proven because the CNN based contextual learning shows poor accuracy on the large object detection task.

V. CONCLUSION

We proposed a context network specifically trained for small object detection and show improvement on the state-of-the-art object detection algorithms for small object detection. Our method can be used in conjecture with any other detection pipeline as a precursor. This network is lightweight and simple and can be ported to any other object detection algorithm as it learns the context in label space and is agnostic to the object detection pipeline.

We hope that this direction of research will open the frontiers for small object detection in navigation environments and scene understanding. This can further improve the scope of research in the direction of obstacle detection and avoidance as most of the current methods use depth based methods to solve the problem. We address the problem using image based architecture.

REFERENCES

[1] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, pp. I–I, IEEE, 2002. 2

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, 2014. 2, 4

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. 2

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010. 2

[5] C. H. Lampert, M. B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008. 2

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893, IEEE, 2005. 2

[7] K. Mikolajczyk, B. Leibe, and B. Schiele, "Multiple object class detection with a generative model," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1, pp. 26–36, IEEE, 2006.

[8] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 26, no. 11, pp. 1475–1490, 2004. 2

[9] V. L. Mustafa zuysal, Pascal Fua, "Fast keypoint recognition in ten lines of code," 2007. 2

[10] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 90–97, IEEE, 2005. 2

[11] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," *Computer Vision–ECCV 2008*, pp. 30–43, 2008. 2

[12] A. Torralba, K. P. Murphy, W. T. Freeman, M. A. Rubin, *et al.*, "Context-based vision system for place and object recognition.," in *ICCV*, vol. 3, pp. 273–280, 2003. 2

[13] D. Hoiem, A. A. Efros, and M. Hebert, "Geometric context from a single image," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, pp. 654–661, IEEE, 2005. 2

[14] A. T. Aude Oliva, "Modeling the shape of the scene: A holistic representation of the spatial envelope," 2001. 2

[15] V. Hedau, D. Hoiem, and D. Forsyth, "Thinking inside the box: Using appearance models and context based on room geometry," *Computer Vision–ECCV 2010*, pp. 224–237, 2010. 2

[16] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1271–1278, IEEE, 2009. 2

[17] N. R. Thomas Kollar, "Utilizing object-object and object-scene context when planning to find things," 2009. 2

[18] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer vision and image understanding*, vol. 114, no. 6, pp. 712–722, 2010. 2

[19] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 520–527, 2007. 2

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," pp. 779–788, 2016. 2, 3, 4

[21] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, pp. 1250 – 1265, 2011/06// 2011. 2

[22] K. Sjö, D. G. López, C. Paul, P. Jensfelt, and D. Kragic, "Object search and localization for an indoor mobile robot," *CIT. Journal of Computing and Information Technology*, vol. 17, no. 1, pp. 67–80, 2009. 2

[23] M. S. Karthik, S. Mittal, G. Malik, and K. M. Krishna, "Decision theoretic search for small objects through integrating far and near cues," in *Mobile Robots (ECMR), 2015 European Conference on*, pp. 1–6, IEEE, 2015. 2

[24] C. Chen, M.-Y. Liu, O. Tuzel, and J. Xiao, "R-cnn for small object detection," in *Asian Conference on Computer Vision*, pp. 214–230, Springer, 2016. 2

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012. 4

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. 4

[27] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. 4