© 2019 IJSRST | Volume 6 | Issue 6 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology
DOI : https://doi.org/10.32628/IJSRST

# Visualization of 2D and 3D Object Detection System

**T. Sridevi[1], S. Sushanth[2]**
[1]Associate Professor, Department of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, India
[2]Student, B.E, Department of CSE, Chaitanya Bharathi Institute of Technology, Hyderabad, India

## ABSTRACT

Existing 3D representations for RGB-D images capture the local shape and appearance of object categories, but have limited power to represent objects with different visual styles. The detection of small objects is also challenging because the search space is very large in 3D scenes. However, we observe that much of the shape variation within 3D object categories can be explained by the location of a latent support surface, and smaller objects are often supported by larger objects. Based on this simple sensor modality for practical applications, deep learning- based monocular 3D object detection methods that overcome significant research challenges are categorized and summarized. This paper gives a visualization of 2d and 3d object detection system.

Keywords : 3D object detection, centralization, objects

## I. INTRODUCTION

Recent approaches to 3D object detection exploit different data sources. Camera based approaches utilize either monocular or stereo images. However, accurate 3D estimation from 2D images is difficult, particularly in long ranges. With the popularity of inexpensive RGB-D sen- sors such as Microsoft Kinect, Intel RealSense and Apple PrimeSense, several approaches that utilize depth information and fuse them with RGB images have been developed. They have been shown to achieve significant performance gains over monocular methods. Deep learning networks have increasingly been extending the generality of object detectors. In contrast to traditional methods in which each stage is individually hand-crafted and optimized by classical pipelines, deep learning networks achieve superior performance by automatically deriving each stage for feature representation and detection. In addition, new approaches for data-driven representation and end-to-end learning with a substantial number of images have led to significant performance improvements in 3D object detection. With the evolution of deep representation, object detection is being widely used in robotic manipulation, autonomous driving vehicles, augmented reality, and many other applications, such as CCTV systems.

Beyond the significant progress in image-based 2D object detection, 3D understanding of real-world objects is an open challenge that has not been explored extensively thus far. In addition to the most closely related studies, we focus on investigating deep learning-based monocular 3D object detection methods. For location-sensitive applications, conventional 2D detection systems have a critical limitation in that they do not provide physically correct metric information on objects in 3D space. Hence, 3D object detection is an interesting topic in both academia and industry, as it can provide relevant solutions that significantly improve existing 2D-based applications.

Camera sensors that capture color and texture information have emerged as an essential imaging

modality in many computer vision applications. The passive camera sensors do not interfere with other active optical systems, and always work well with them when needed. For image-based deep representations that encode depth cues, monocular images are also highly cost-effective. Owing to considerable accumulations of annotations for RGB databases, the data-driven representations using deep neural networks make monocular 3D object detectors even more advantageous without expensive depth-aware sensors or cameras at additional viewpoints.

To understand the major breakthroughs and current progress in practical 3D object detection, we contribute to the literature by reviewing recent developments in deep learning-based state-of-the-art 3D object detection with monocular RGB databases.

The task of target detection is to find all Region of Interests (ROI) in the image and determine their positions and categories. Due to the different appearance, shape and attitude of various objects, as well as the interference of lighting, shielding and other factors during imaging, object detection has continuously been a challenging problem in the field of computer vision.

However, existing 3D detection algorithms suffer some common problems. Given diverse objects in the same category, modeling different visual styles is often very challenging, and ground truth annotations of 3D cuboids can vary among different human annotator. Moreover, objects with smaller physical size are hard to detect because the search space in the whole scene is very big, and bottom-up proposals typically contain many false positives. State-of-the-art 3D object features, such as COG and TSDF, are calculated for a grid of voxels within each hypothesized 3D cuboid. A major cause of feature in-consistency across different object instances is variation in the location of the supporting surface contained by many indoor objects. We treat the height of the support surface as a latent variable, and use it to distinguish different visual styles of the same object category.

Modeling support surface can also help detect smaller objects like monitors, lamps, TVs, and pillows. Since small objects are typically placed on the supporting surfaces of large objects, we first detect large objects on the ground and predict their support surface location, and then search for small objects on top of support surface areas. The reduced search space for small objects naturally reduces false positives and improves performance.

## II. LITERATURE REVIEW

Given an image with a pixel grid representation, object detection is the task of localizing instances of objects with bounding boxes of a certain class. An important contribution in solving the 2D object detection problem is the use of region-based convolutional neural networks (R-CNNs), which involves two main stages: region proposal and detection. The region of interest (ROI) of an image is proposed on the basis of certain assumptions, such as color, texture, and size. The ROI is cropped to feed a CNN that performs the detection. By combining prior knowledge and labeled datasets, the two-stage detection framework has emerged as a classical model in both 2D and 3D object detection.

Another important algorithm for object detection is the YOLO algorithm. It does not have a separate region proposal stage; instead, it divides an input image roughly into an $N \times N$ grid. Based on each grid cell, localization and classification tasks are performed together in a unified regression network, followed by further post-processing. Early end-to-end approaches performed poorly in the detection of small or occluded objects. As new datasets are being developed,

there have been significant innovations in end-to-end networks. As fewer proposal steps with hand-crafted features are involved in single-stage methods, they are computationally less complex than multi-state approaches that usually prioritize detection accuracy. In practice, there was active competition between multi-stage and single-stage methods for object detection tasks. 3D object detection is similar to this overall flow.

The goal of 3D object detection systems is to provide 3D-oriented bounding boxes for 3D objects in the 3D real world. The 3D cuboids can be parameterized by 8-corners, 3D centers with offsets, 4-corner-2-height representations, or other encoding methods. In monocular 3D object detection methods, we seek the oriented bounding boxes of 3D objects from single RGB images. Similarly to 2D-image-based object detection systems, monocular 3D object detection methods can be also categorized into two main type. From a taxonomic point of view, we have extended them to six sub-categories, according to the main distinguishing features of each sub-category. We have summarized the main features of ten high-quality datasets, such as descriptions with quick links, input data types, contextual information for different applications, the availability of synthetic RGB images, the number of 3D object instances/categories, the number of training/testing images, and lastly, other related references, which can be used for future research. We have briefly explained key features of the most representative works for each category and the related databases, computational time, and so on. All of those methods use powerful algorithms that can only run on a high-performance system using GPUs, and we did not pay attention to lightweight deep learning models for lower-power embedded/mobile systems.

Based on a general understanding of object detection, we review 11 datasets for monocular 3D object detection and more than 29 recent algorithms. The unique properties of 3D object detection systems, such as different data representations and the availability of both 2D and 3D annotations, make the 3D detection frameworks more complicated and interesting.

2D Object detection We highlight some of the most related work in the rich literature on object detection.

[2] introduced the histogram of oriented gradient (HOG) descriptor to model 2D object appearance. Building on HOG, [4] use a discriminately-trained part-based model to represent objects. This method is effective because it explicitly models object parts as latent variables, which implicitly encode object style variations. More recently, many papers have used convolutional neural networks (CNNs) to extract rich features from images. These methods achieve state-of-the-art performance and efficient detection speed, but in cluttered indoor scenes, accurate 2D object localization remains a challenging task.

3D Object Detection Increasingly, real-world computer vision systems often incorporate depth data as additional input to increase accuracy and robustness. There have recently been significant advances in methods for 3D object classification, point cloud segmentation, room layout prediction, 3D object context and 3D shape reconstruction. Here, we focus on the related problem of 3D object detection.

In outdoor scenes, object localization with 3D cuboids has become a new standard in the popular KITTI autonomous driving benchmark. 3D detection systems model car shape and occlusion patterns using lidar or stereo inputs, and may also incorporate additional bird's eye view data. However, methods for outdoor 3D detection are usually focused on the identification of vehicles and pedestrians in open

scenes, and do not generalize to more challenging detection tasks in cluttered scenes.


Figure 1: Visualization of 3D object detection system

In indoor scenes, a larger number of object categories is common, and categories have greater shape and style variations. Because indoor objects are often heavily occluded by their cluttered environments, localizing objects with 3D cuboids instead of 2D bounding boxes can be more useful. Some work aligns 3D CAD models to objects in RGB-D inputs, as evaluated on the small- scale NYU Depth dataset, but the computational cost is usually expensive. A simple 3D convolutional neural network was designed to detect simple objects in real time. Other work utilizes pretrained 2D detectors or region proposals as priors, and localizes 3D bounding boxes via a separate CNN Those methods can achieve good performance with great computational speed, but are very sensitive to the accuracy of 2D object proposals. [6] introduce the clouds of oriented gradient (COG) to represent 3D cuboids and perform holistic scene under- standing with a cascaded prediction framework. Al- though this work achieves state-of-the-art performance on the SUN RGB-D dataset for 10 large object categories, it cannot be directly used to detect smaller objects because it requires exhaustive search in 3D space. In addition, the COG feature does not capture object style variations.

## III. 3D SHAPE INFORMATION

While Mono3D, an optimization-based pioneering method, does not show satis- factory accuracy and speed, its successor, Mono3D++, achieves improved performance with better template matching, as does the Ceres toolbox. Mono3D++ uses coarse and fine 3D hypotheses to infer the object shape and pose from one RGB image. Specifically, a fine representation for vehicles is generated by morphable wireframe models with differ- ent shapes and poses. For lower sensitivity to 2D landmark features, a coarse representation aims to model 3D bounding boxes to improve stability and robustness. For joint energy minimization with a projection error, three priors are considered, namely, vehicle shape, a ground plane constraint, and unsupervised monocular depth.

3D shape information-based methods tend to become slow when the number of shape templates or object poses increases, because hand-crafted steps for comparing them are required for optimization. To tackle this problem with some physical quantities, [7] proposed a new image feature based on orientation histograms of random projection images from CAD models. Similarly, coarse initialization was adopted for 3D poses of texture-less objects. In [8], temporally consistent, local color histograms were used for pose estimation and segmentation of rigid 3D objects. For handheld objects, the statistical descriptors can be learnt online within a few seconds.

Instead of optimizing separate quantities, [4] proposed a multi-tasking network structure for 2D and 3D vehicle analysis from a single image. For simultaneous part localization, visibility characterization, vehicle detection, and 3D dimension estimation, the many-tasks network (MANTA) first detects 2D bounding boxes of vehicles in multiple refinement stages. For each detection, it also gives the 3D shape template, part visibility, and part coordinates of the detected vehicle even if some parts are not visible. Then, these features are considered to estimate the vehicle localization and orientation using 2D–3D correspondence matching. To access the 3D information of the test objects, the vehicle models are searched for template matching. The real-time pose

and orientation estimation uses the outputs of the network in the inference stage. At the time of publication, this approach was the state-of-the-art approach using the KITTI 3D benchmark in terms of vehicle detection, 3D localization, and orientation estimation tasks.

An input image is passed forward to the deep MANTA net- work where convolution layers with the same weights have the same color. The existing architecture is split into three blocks. With these networks, the object proposals are refined iteratively until the final detection that is associated with the part's coordinate, the part's visibility, and the template similarity. Moreover, non-maximum suppression (NMS) re- moves some redundant detections. Based on the outputs, the best 3D shape is chosen in the inference stage. 2D and 3D pose computation is then performed with the associated shape. In the ROI-10D algorithm, a monocular deep network directly optimizes a novel 3D loss formulation and then lifts a 2D bounding box to 3D shape recovery and pose estimation. Using CAD templates and synthetic data augmentation, deep feature maps are generated and combined to obtain the shape dimensions. Then, shape regression is performed to obtain the object information. In particular, the pose distributions are well analyzed in the KITTI 3D benchmark. In metrically accurate pose estimation, learning synthetic data is useful for increasing the pose recall; however, some hand-crafted modules such as 2D and 3D NMS have a strong influence on the final results.

## IV. CONCLUSION

Modeling the height of the support surface as a latent variable leads to improved detection performance for large objects, and contains the search space for small object detectors. Via a cascaded prediction framework our detector achieves state-of-the-art performance on the SUN RGB-D dataset, demonstrating the effectiveness of modeling support surfaces in 3D object detection. They are being employed in various practical applications such as autonomous vehicles and robotics. We believe that the current gap between mature 2D-based methods and nascent 3D-based methods can be rapidly bridged on the basis of the intensive review presented herein. This paper given a visualization of 2d and 3d object detection system.

## V. REFERENCES

[1] Almasi, Omid N, Rouhani, Modjtaba, "A new fuzzy membership assignment and model selection approach based on dynamic class centers for fuzzy SVM family using the firefly algorithm", Turkish Journal of Electrical Engineering & Computer Sciences, 4: 1–19, 2016.

[2] Arunkumar Sangiah, Arunkumar Thangavelu, Venkatesan MeenakshiSundram, "Cognitive Computing for Big Data Systems over IoT", Gewerbestrasse, 11, p.6330, 2018.

[3] Bavya N, Arunkumar T, Adalarasu K, "A Comprehensive Survey on IoT Technologies in Health Care System", Research Journal of Pharmacy and Technology, Vol. 11, Issue 7, pp. 3157-3162, 2018.

[4] Chen, X., Zheng, Z., Yu, Q., Lyu, M.R., "Web service recommendation via exploiting location and QoS information", IEEE Trans. Parallel Distrib. Syst., 25(7), pp. 1913 – 1924, 2014.

[5] F. Ding, A. Song, E. Tong, J. Li, "A smart gateway architecture for improving efficiency of home network applications", Journal of Sensors, 2016.

[6] C. Tao, X. Ling, S. Guofeng, Y. Hongyong, H. Quanyi, "Architecture for monitoring urban infrastructure and analysis method for a smart-safe city", In Sixth International Conference on

Measuring Technology and Mechatronics Automation, pp. 151–154, 2014.

[7] Umadevi K S, ArunkumarThangavelu, “An optimal medium access slot allocation for wimedia medium access control protocol using firefly algorithm”, International conference on Microelectronic Devices, Circuits and Systems (ICMDCS), pg 1-3, DOI: 10.1109/ICMDCS.2017.8211711, August 2017.

[8] C. Xiaojun, L. Xianpeng, X. Peng, “IoT-based air pollution monitoring and forecasting system”, In 2015 International Conference on Computer and Computational Sciences (ICCCS), pp. 257–260, 2015.

[9] Yang, X. S., “Nature-Inspired Metaheuristic Algorithms”, Luniver Press, pp.242-246, 2008. ISBN 978- 1-905986-10-1.

[10] Yin, Jianwei, Wei Lo, Shuiguang Deng, Ying Li, Zhaohui Wu, and Naixue Xiong, “Colbar: A collaborative location-based regularization framework for QoS prediction”, Information Sciences, 265, 68-84, 2014.

## Cite This Article :

T. Sridevi, S. Sushanth, "Visualization of 2D and 3D Object Detection System", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 6 Issue 6, pp. 382-387, November-December 2019.

Journal URL : https://ijsrst.com/IJSRST207454