3D目标检测技术的研究进展

王永森 刘宏哲

(北京联合大学北京市信息服务工程重点实验室 北京 100101)

摘 要 近年来,随着人工智能的发展,尤其是自动驾驶汽车的发展,使目标检测技术成为研究的热点。现如今,得益于深度学习技术的发展及应用,2D目标检测技术已经非常成熟,相比于传统的目标检测方法,其检测准确度有了很大的提高。但是,2D目标检测还无法满足自动驾驶汽车等产业的需要,还需获取更准确的物体目标信息,所以3D目标检测技术成为了人们进一步的研究热点。基于此,文中对目前的3D目标检测方法进行了介绍,并给出了各类方法的基本思想和一般处理流程,然后对典型的检测方法进行了分析,最后给出了可能的研究方向。

关键词 目标检测,3D目标检测,激光雷达点云,单目视觉,双目视觉,自动驾驶汽车,人工智能

Study Progress of Advances in 3D Object Detection Technology

WANG Yong-sen LIU Hong-zhe

(Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China)

Abstract In recent years, with the development of artificial intelligence, especially the development of autonomous vehicles, target detection technology has become a research hotspot. Nowadays, thanks to the development and application of deep learning technology, 2D object detection technology is very mature, and the detection accuracy is greatly improved compared with the traditional target detection method. However, 2D object detection cannot meet the needs of industries such as autonomous vehicles, and more accurate object target information is needed. Therefore, 3D object detection technology has become a hotspot for further research. Based on this, the paper introduced the current 3D objectdetection method, and gave the basic ideas and general processing flow of various methods, then analyzed the typical detection methods, and finally gave the possible research directions.

Keywords Object detection, 3D object detection, Lidar point cloud, Monocular vision, Binocular vision, Autonomous vehicle, Artificial intelligence

1 引言

目前,在智能驾驶领域,目标检测变得非常重要,随着深度学习技术的发展,目标检测技术已经非常成熟,被大量应用于工业界。常规的目标检测方法都是对图像进行2D目标检测,但是在无人驾驶、机器人、增强现实的应用场景下,普通2D检测并不能提供感知环境所需要的全部信息,2D检测仅能提

本文受北京市属高校高水平教师队伍建设支持计划项目(IDHT20170511)资助。

王永森(1994—),男,硕士生,主要研究方向为数字图像处理、计算机视觉、深度学习,E-mail:861935973@qq.com;刘宏哲(1971—),女,博士,教授,主要研究方向为语义计算、数字图像处理、人工智能等。

供目标物体在二维图片中的位置和对应类别的置信度，但是在真实的三维世界中，物体都是有三维形状的，大部分应用都需要目标物体的长、宽、高以及偏转角等信息。所以，3D 目标检测应运而生，成为了目前的研究热点。目前，出现了很多具有挑战性的数据集，比如 KITTI 数据集[1]、Pascal3D+数据集[2]等，可见，3D 目标检测技术已经变得越来越重要。

3D 目标检测是视觉感知、运动预测和自动驾驶规划的重要基础，尤其在自动驾驶汽车领域，获取到目标障碍物的三维信息，可以提高对目标的分析精度，对于后续自动驾驶场景中的路径规划和控制具有至关重要的作用。比如，在进行前方车辆目标测距时，如果对 2D 目标检测结果进行分析，就会出现很大的测距误差，但是，如果使用 3D 目标检测，得到目标的三维信息，就可以获取到目标的精准下边沿，再进行测距时就可以很大程度上降低测距的误差。因此，3D 目标检测技术在智能驾驶领域中应用极为广泛，有着广阔的应用前景。文献[3]提出了一种新方法，称为 Deep MANTA(Deep Many-Tasks)，用于从给定图像中进行多任务车辆分析。他们引入了一个强大的卷积网络，用于同时进行车辆检测、零件定位、可视性表征和 3D 尺寸估算。文献[4]提出了在自动驾驶领域中使用单目视觉进行 3D 物体检测。其方法首先生成一组候选类特定对象推荐区域，然后通过卷积神经网络管道运行以获得高质量的对象检测。文献[5]提出了一种在自动驾驶的环境中生成高质量的 3D 检测推荐区域的方法。该方法利用立体图像以 3D 边界框的形式存放推荐区域。其将问题表述为最小化编码物体尺寸先验、地平面以及几个深度特征的能量函数，通过这些特征推断出三维空间、点云密度和目标到地面的距离。文献[6]提出了一种在自动驾驶场景中的高精度三维物体检测方法。该方法为多视图 3D 网络，是一种感知融合框架，它将点云数据和图像数据作为输入并预测定向的 3D 边界框。该网络由两个子网组成：一个用于 3D 对象建议生成，另一个用于多视图特征融合，候选区域生成网络从点云的俯视图表示中有效地生成 3D 候选框。文献[7]提出了一种在点云上进行 3D 语义分割的方法，该方法没有其他 3D 数据格式中出现的问题。文献[8]提出了一种新颖的 3D 物体探测器，它可以利用激光雷达和相机来得到非常精确的定位。他们设计了一种端到端的学习架构，利用连续卷积将图像和雷达特征图融合在不同的分辨率水平。文献[9]研究了室内和室外环境中深度传感器捕获的 RGB-D 数据的三维物体检测问题。他们利用了降维算法和成熟的二维物体检测框架，开发了 Frustum PointNet 框架。在 KITTI 检测的基准测试中，该方法有着明显优势。

本文主要对目前 3D 目标检测技术的研究进展进行分析，并探索进一步的研究方向。

2 3D 目标检测技术的分析

目前 3D 目标检测技术正处于高速发展时期，主要是综合利用单目相机、双目相机、多线激光雷达来进行 3D 目标检测。从成本上讲，激光雷达成本最高，其次是双目相机，最后是单目相机；从准确率上讲，激光雷达精度最高，其次是双目相机，最后是单目相机。不过，随着技术的不断发展，出现了越来越多的 3D 目标检测方法。

2.1 基于激光雷达的 3D 目标检测方法

文献[10]提出了一种使用激光雷达获取点云数据进行 3D 目标检测的方法。为了将高度稀疏的激光雷达点云与区域建议网络(Region Proposal Network，RPN)进行结合，目前大多数的工作都集中在手工制作的特征表示上，例如，鸟瞰投影。如图 1 所示，在这项工作中，他们提出了 VoxelNet 网络框架，该方法不需要人工制作 3D 点云特征，它是一个通用的 3D 检测网络，将特征提取和包围框预测统一到一个阶段，是一个端到端可训练的深层网络。具体而言，VoxelNet 网络将点云划分为等间距的三维体素，并通过新引入的体素特征编码(Voxel Feature Encoding，VFE)层将每个体素中的一组点变换为统一的特征表示。以这种方式，点云被编码为描述性的体积表示，然后将其连接到 RPN 层以产生检测候选框。KITTI 车辆检测的基准实验表明，VoxelNet 方法大幅度优于现有的基于激光雷达的 3D 检测方法。

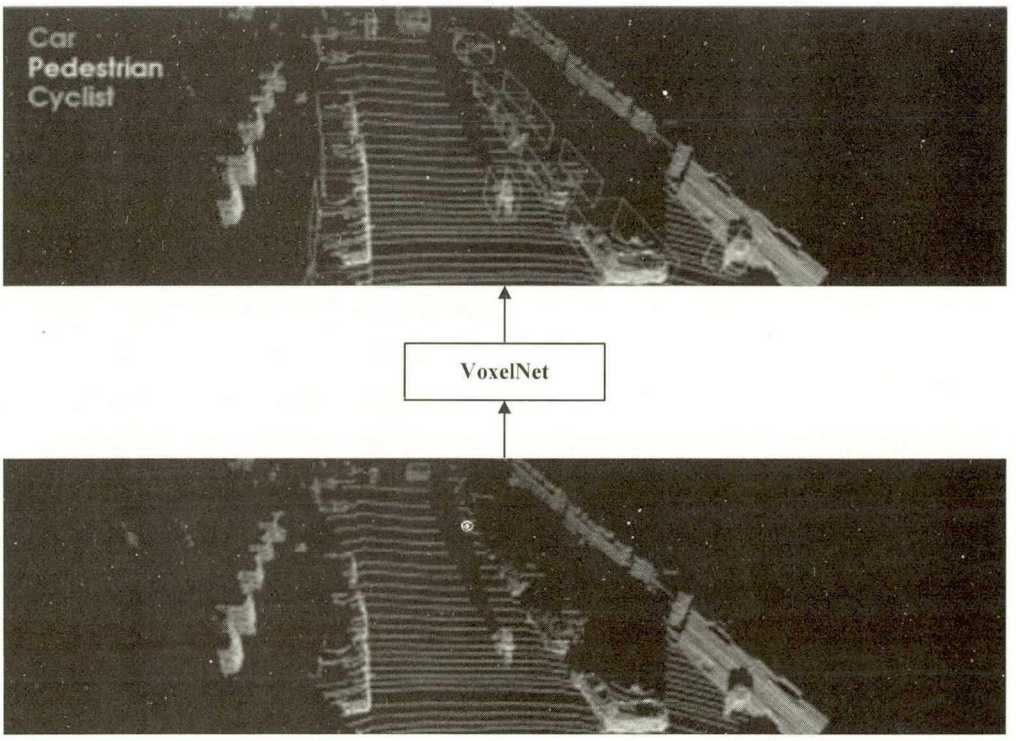
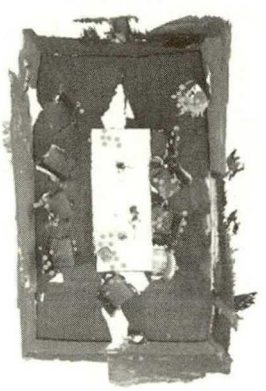
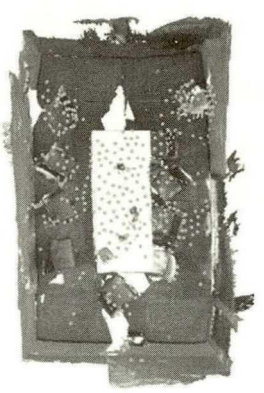


图 1　VoxelNet 网络

文献[11]重点分析了 3D 检测中的一个难题，即一个 3D 物体的质心可能远离任何表面点，因此很难用一个步骤准确地回归。其提出了一种 VoteNet 网络用于解决这个问题。这是一个基于深度点集网络和霍夫投票的端到端 3D 目标检测网络。该网络直接处理原始数据，不依赖 2D 检测器。在图像中，目标中心附近通常存在一个像素，但在点云中却并非如此。由于深度传感器仅捕获物体的表面，因此 3D 物体的中心很可能在远离任何点的空白空间中。因此，基于点的网络很难在目标中心附近聚集场景上下文。简单地增加感知域并不能解决这个问题，因为当网络捕获更多的上下文时，它也会包含更多的附近的对象和杂物。为此，我们提出为点云深度网络予一种经典霍夫投票机制。如图 2 所示，通过投票，基本上生成了靠近对象中心的新的点，这些点可以进行分组和聚合，以生成 Box Proposals，投票有助于增加检测的上下文，从而增加了准确检测的可能性。与传统的多独立模块、难以联合优化的霍夫投票相比，VoteNet 是端到端优化的。具体来说，在通过主干点云网络传递输入点云之后，我们对一组种子点进行采样，并根据它们的特征生成投票。投票的目标是使其到达目标中心。因此，投票集群出现在目标中心附近，然后可以通过一个学习模块进行聚合，生成 Box Proposals。如图 3 所示，当目标点远离目标中心的情况下，投票更有帮助。该模型是一个强大的 3D 物体检测器，它是纯几何的，可以直接应用于点云。



(a)BoxNet(no voting)



(b)VoteNet

图 2　投票模型

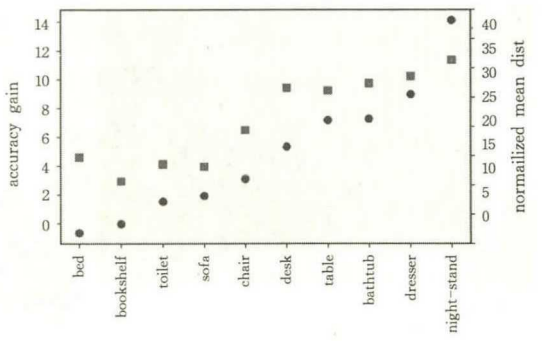


图 3 投票模型结果

我们在两个具有挑战性的 3D 目标检测数据集上评估了本文方法:SUN RGB-D 数据集[12]和 ScanNet 数据集[13]。在这两个数据集上,仅使用几何信息的 VoteNet 明显优于使用 RGB 和几何甚至多视图 RGB 图像的现有技术。研究表明,投票方案支持更有效的上下文聚合,并验证了当目标中心远离目标表面时,VoteNet 网络能够提供最好的效果。

2.2 基于单目相机的 3D 目标检测方法

文献[14]提出一种方法,只使用单目相机得到的 RGB 图像进行 3D 目标检测。他们对单个图像进行 3D 对象检测和姿势估计。与使用回归对象的 3D 边界框的技术相比,他们的方法首先使用深度卷积神经网络回归相对稳定的 3D 对象属性,然后将这些估计与 2D 对象边界框提供的几何约束组合,以产生完整的 3D 边界框。第一个网络输出使用新颖的混合离散连续损耗估计 3D 物体方向,其明显优于 L2 损失。第二个输出回归 3D 对象尺寸,与已有方案相比,其具有相对较小的变化,并且通常可以针对许多对象类型进行预测。这些估计与 2D 边界框施加的平移几何约束相结合,能够恢复稳定且精确的 3D 对象姿势。如图 4 所示,他们将卷积层得到的共享特征图分为 3 个分支进行处理,一个用于估计感兴趣区域对象的尺寸,一个用于计算偏移角度,一个用于计算目标置信度。

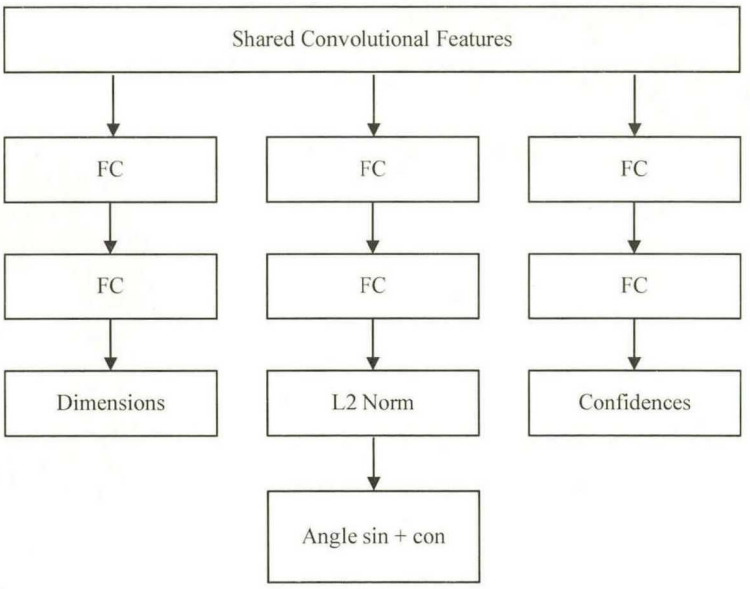


图 4 回归网络分支图

2.3 激光和单目相机相结合的 3D 目标检测方法

文献[15]提出一种多视图目标检测网络,使用雷达和 RGB 图像来生成特征。该算法主要包括两部分:RPN 网络和二级检测网络。RPN 网络可以在高分辨率特征图上执行多模型特征融合,从而对道路上的多类目标生成可靠的 3D 候选目标;二级检测网络执行精确的 3D 边框的定向、回归和分类,以及预测目标在三维空间中的尺寸、方向和分类。AVOD 网络也可用于对小尺寸目标的监测和定位,AVOD 网络中的 RPN 结构将图像俯视特征图的全分辨率的特征要素作为输入,即使是较小尺寸的目标也能产生高召回率的候选框。此外,特征提取器可以提高全分辨率下特征图的获取效率,这对于小尺寸目标的定位精度非常有帮助。候选框的生成过程为:1)利用特征提取器从鸟瞰图和图像中生成特征图;2)RPN 结构利用这两个特征图生成未定向的候选区域;3)检测网络利用候选区域进行维度细化、方向估计和分类。网络整体结构如图 5 所示。

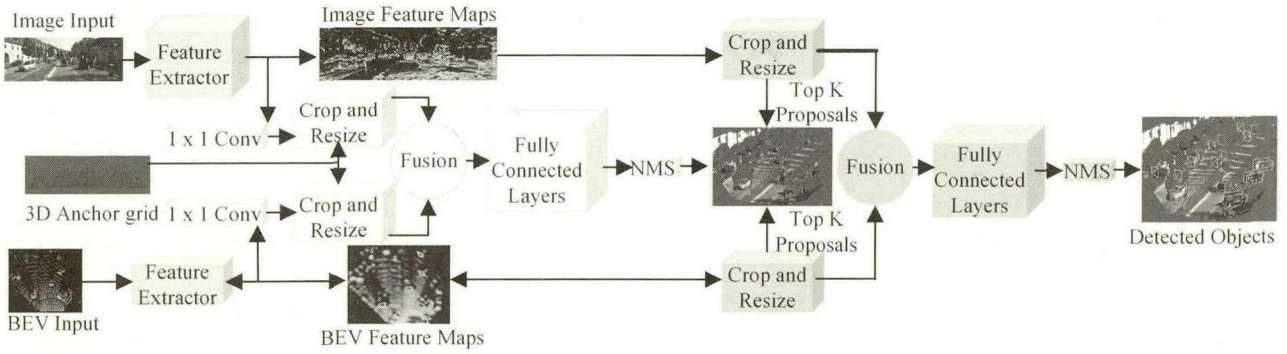


图 5　AVOD 网络结构图

2.4　基于双目视觉的 3D 目标检测方法

文献[16]提出一种基于双目视觉的 3D 目标检测方法。该方法是一种充分利用立体图像中稀疏、密集、语义和几何信息的自动驾驶三维目标检测方法，其同时检测和关联对象在左图像和右图像中的特征。他们在立体区域建议网络之后添加额外的分支来预测稀疏的关键点、视点和对象尺寸，将这些关键点、视点和对象尺寸与二维左、右框相结合来计算粗糙的三维对象边界框。然后，通过使用左、右感兴趣区域的光度校准来恢复精确的三维边界框。该方法不需要深度数据和三维位置信息，但优于所有现有的完全基于图像的方法。在具有挑战性的 KITTI 数据集上的实验表明，在 3D 检测和 3D 定位任务上，他们的方法比最先进的基于立体信息的方法快 30%。

该方法同时检测和关联对象在左、右图像中的特征信息，使用立体的 R-CNN 网络来输出相应的左右图像区域建议。其将 ROI Align[17] 分别应用于左、右特征图后，将左、右感兴趣区域连接起来，对目标类别进行分类，并在立体回归中回归精确的二维立体框、视点和尺寸。通过预测左侧感兴趣区域的对象关键点，将其用于 3D 框估计的稀疏约束。他们用目标在左、右二维图像中框的位置和关键点来建立三维框对应顶点之间的投影关系，以确保三维定位性能的关键组件是密集的三维框。我们认为三维物体定位是一个学习辅助几何问题，而不是一个端到端的回归问题。我们不直接使用深度信息，而是直接使用对象属性，将对象的感兴趣区域视为一个整体，而不是独立的像素。对于规则形状的对象，在给定粗略的三维边界框的情况下，可以推断出每个像素与三维中心之间的深度关系。他们根据对象与 3D 结构中心的深度关系，将左侧感兴趣区域中的密集像素映射到右侧图像，以找到最佳的中心深度，最大限度地减少误差。因此，整个对象的感兴趣区域形成了三维对象深度估计的密集约束。

3　进一步研究需要解决的问题

综上所述，采用纯视觉的单目相机 3D 目标检测方法在准确率上离预期还有较大差距，但是它的一大优点就是成本比较低，下一步研究计划可以考虑引入使用深度神经网络结合稀疏激光点云生成稠密点云对检测结果进行修正，从而提高 3D 检测的准确率。除此之外，目前大部分的方法都是采用“一步法”进行 3D 目标的姿态回归，后续可以考虑使用“两步法”来解决，将检测阶段分为两个或多个阶段进行处理，并且可以结合传统的图像处理方法，比如加入分割的信息等，以进一步提升检测的精度，还可以考虑使用更多的几何约束等。另外，考虑到目前 3D 目标检测的标注数据较少，可以考虑引入非监督学习的方法进行训练。

结束语　目前 3D 目标检测发展还不是很成熟，虽然检测方法很多，但是综合精度、速度、成本还没有达到一个权衡，精度高的方法如雷达，相比其他方法有很高的精度，但是雷达的成本相对来说非常高，然而，成本最低的单目相机的方法受限于 2D 的 RGB 图像，导致精度不高。所以，在 3D 目标检测方面，还有很大的发展空间，相信未来一定会取得更大的进展，从而推动整个行业的进步。

参 考 文 献

[1] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? the kitti vision benchmark suite [C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 3354-3361.

[2] XIANG Y, MOTTAGHI R, SAVARESE S. Beyond pascal: A benchmark for 3d object detection in the wild [C]// IEEE Winter Conference on Applications of Computer Vision. IEEE, 2014: 75-82.

[3] CHABOT F, CHAOUCH M, RABARISOA J, et al. Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2040-2049.

[4] CHEN X, KUNDU K, ZHANG Z, et al. Monocular 3d object detection for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2147-2156.

[5] CHEN X, KUNDU K, ZHU Y, et al. 3d object proposals for accurate object class detection[C]// Advances in Neural Information Processing Systems. 2015: 424-432.

[6] CHEN X, MA H, WAN J, et al. Multi-view 3d object detection network for autonomous driving[C]// IEEE CVPR. 2017, 1(2): 3.

[7] HUANG Q, WANG W, NEUMANN U. Recurrent Slice Networks for 3D Segmentation of Point Clouds [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2626-2635.

[8] LIANG M, YANG B, WANG S, et al. Deep continuous fusion for multi-sensor 3d object detection[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2018: 641-656.

[9] QI C R, LIU W, WU C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 918-927.

[10] ZHOU Y, TUZEL O. Voxelnet: End-to-end learning for point cloud based 3d object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4490-4499.

[11] QI C R, LITANY O, HE K, et al. Deep Hough Voting for 3D Object Detection in Point Clouds[J]. arXiv: 1904. 09664, 2019.

[12] SONG S, LICHTENBERG S P, XIAO J. Sun rgb-d: A rgb-d scene understanding benchmark suite[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 567-576.

[13] DAI A, CHANG A X, SAVVA M, et al. Scannet: Richly-annotated 3d reconstructions of indoor scenes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5828-5839.

[14] MOUSAVIAN A, ANGUELOV D, FLYNN J, et al. 3d bounding box estimation using deep learning and geometry[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7074-7082.

[15] KU J, MOZIFIAN M, LEE J, et al. Joint 3d proposal generation and object detection from view aggregation [C]// International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018: 1-8.

[16] LI P, CHEN X, SHEN S. Stereo r-cnn based 3d object detection for autonomous driving[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7644-7652.

[17] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.