Review and Systematization of Solutions for 3D Object Detection

Jonas Friederich[1], Patrick Zschech[1]

[1] Technische Universität Dresden, Business Intelligence Research, Dresden, Germany
{jonas.friederich,patrick.zschech}@tu-dresden.de

Abstract. Since 2017 there has been an exponential growth in scientific publications regarding the field of 3D object detection (3DOD). On the one hand, this growth can be explained by the strong demand for autonomous vehicles, and on the other hand, by the wide availability of 3D sensors. Due to the strong heterogeneity of developed approaches, this paper aims to identify, analyze and systematize publications that propose end-to-end solutions for 3DOD towards the goal to provide a structured framework which can guide future development, evaluation and application activities. To carry out the research, a systematic literature review is combined with a taxonomy development approach. The resulting framework consists of six dimensions, covering the addressed domains, applied datasets, sensor properties, data representation formats, modeling techniques, and evaluation criteria. The taxonomy can help researchers and practitioners to get a quick overview about the field by decomposing 3DOD solutions into more manageable pieces.

Keywords: Data Science, Computer Vision, 3D Object Detection, Taxonomy.

1 Introduction

The visual system allows humans to identify objects in an image and determine where they are by simply glancing at them [1]. The interdisciplinary research field of computer vision seeks to mimic this specific ability and refers to it as object detection [2]. Object detection includes the parametrization of a bounding box containing the recognized and classified object [3, 4]. Most research in that area has focused on two-dimensional (2D) object detection based on widely available Red-Green-Blue (RGB) or greyscale images. However, this completely leaves out the third dimension and only partly imitates the human visual system. In fact, many modern applications like robot assistants or autonomous vehicles are highly dependent on depth and surface data to securely navigate in three-dimensional (3D) environments [5].

With the rising availability of mobile 3D sensors like RGB-Depth (RGB-D) cameras and LiDAR (light detection and ranging) sensors, more depth data can be captured and processed [5]. This allowed great progress in the field of 3D object detection (3DOD) and brought forth a variety of heterogeneous methods and solutions used for this task [6]. Simultaneously, this heterogeneity results in an increasing difficulty to compare the multifaceted approaches with each other and derive findings

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_r2-friedrich

for archetypal application scenarios. Consequently, it requires a systematization that structures the field and provides an orientation for researchers and practitioners. As of today, there is only one publication so far aiming to arrange and classify a representative set of approaches for 3DOD with a particular focus on the domain of autonomous driving [6].

Therefore, we aim to complement and extend this work by conducting an exhaustive domain-independent review, as it can help developers and operators in information systems (IS), data science and other related fields to gain a quick overview about different design options within the area of 3DOD. More specifically, we pursue the goal to identify, analyze and systematize characteristic components of 3DOD end-to-end solutions to increase transparency and guide the selection of different components during the development and evaluation of (novel) solutions and their embedding into analytical IS. Thus, the following two research questions are focused in this paper:
RQ1: *Which publications deal with the development of 3DOD methods/solutions?*
RQ2: *How can the identified results be systematized within a structured framework?*

To answer our research questions, we first apply a systematic literature review [7]. Subsequently, we use the identified articles to develop a taxonomy according to the guidelines of Nickerson et al. [8]. Following this line, our paper is organized as follows: In the next section, we briefly describe the background of 3DOD and refer to related work. We then describe our research method in detail and subsequently present the results of the literature review and the taxonomic framework. Finally, we discuss our findings, draw a conclusion and give an outlook for future research.

2 Conceptual Background and Related Work

To emulate the human visual system has become a challenging task within the last decades. As a result, the scientific community tries to make computers gain the same high-level understanding from digital images or videos as humans within the research field of computer vision [9]. A core task arising from this trend is object detection, which is the fusion of object recognition and localization [2]. For this task, different kinds of machine learning (ML) algorithms can be applied, whereas recent efforts are increasingly directed towards neural networks with ever deeper network architectures. This allows them to be fed with high-dimensional input data and then automatically discover internal structures and representations that are needed for detection tasks [10].

3D vision aims to extend the previously discussed concepts of computer vision by adding data of the third dimension. This leads on the one hand to six possible degrees of freedom (6DoF) instead of three (i.e., surge, heave, sway, yaw, pitch and roll), and on the other, to an accompanying increase in the number of scenery configurations. While methods in 2D space are good for simple visual tasks, more sophisticated approaches are needed to improve, for instance, autonomous driving applications or complex automated productions lines supported by robots [11]. To capture 3D scenes, commonly used monocular cameras are no longer sufficient. Therefore, special

https://doi.org/10.30844/wi_2020_r2-friedrich

sensors have been developed to capture depth information. RGB-D cameras like Microsoft's Kinect use stereo vision [12], while LiDAR sensors like Velodyne's HDL-64E use laser beams to infer depth information [13]. The data acquired by these 3D sensors can be converted to a more generic structure, the point cloud. Formally spoken, it is a set of points of a vector space that has an unorganized spatial structure [14]. The point cloud is described by the contained points, which are each described by their coordinates x, y and z in a 3D coordinate system. Besides these spatial coordinates, they can also contain additional parameters like RGB-color intensity or distance to ground plane.

Utilizing the input acquired by sensors, the idea of 3DOD is to output 3D bounding boxes and the corresponding class labels for all relevant objects within the sensors field of view. 3D bounding boxes are rectangular cuboids in the 3D space. To ensure relevancy, their size should be minimal, while still containing all relevant parts of an object. A 3D bounding box usually gets parameterized as (x, y, z, h, w, l, c), where (x, y, z) represent the 3D coordinates of the bounding box center, (h, w, l) refer to the height, width and length of the box and c stands for the class of the box [15].

So far, only one article by Arnold et al. [6] has been published dealing with the review and systematization of 3DOD solutions. In their survey, the authors examine exclusively DL-based methods focused at the domain of autonomous driving scenarios. Thus, contributions that cover the relevant domain of indoor applications as well as non-DL-based approaches are currently lacking. To fill this gap, we conduct the following exhaustive review and develop a taxonomy to systematize the emerging field.

3 Research Method

3.1 Systematic Literature Review

To identify relevant studies dealing with 3DOD methods and solutions, we carried out a systematic literature review [7]. Specifically, we applied a database search using the following common libraries for IS and ML research: *IEEE Xplore, ScienceDirect, ACM Digital Library, SpringerLink, EBSCOhost* and *arXiv*. As search terms, we combined the keywords '3D', 'object' and 'detection' and applied them with alternative spellings (e.g., 'three dimensional') and related terms (e.g., 'recognition', 'localization', 'classification'). However, to ensure relevancy, the related terms were exclusively applied in combination with the term 'detection'. This led to a total of 705 items which had to be further reduced by appropriate filter criteria (day of search: 2019-07-15).

In a **first stage**, inclusion and exclusion criteria were applied after reading the titles, summaries and conclusions of each article [16]. The criteria were derived from the review scope based on the research questions and were formulated as decision statements. Thus, inclusion criteria covered that the publications are written in German or English and that the focus of the publication is on 3DOD methods and solutions, whereas the exclusion criteria assured to remove duplicates and that the

https://doi.org/10.30844/wi_2020_r2-friedrich

publication is not a survey or an introduction of a new dataset. By applying these criteria, 560 results could be excluded from further processing.

In a **second stage**, more fine-grained criteria were defined and all full texts were screened to assign the remaining 146 results to the classes *relevant*, *supporting* and *irrelevant*. The *relevant* class covered the most important articles and served as the main sources, as they describe end-to-end solutions for 3DOD (83 items), that is solutions that describe components along the entire 3DOD processing pipeline. *Supporting* articles contain important partial information alongside this pipeline and concentrate on specific components, such as data preprocessing, feature engineering or specific network architectures (40 items). They were used to back up certain statements and ideas presented within the relevant articles. The *irrelevant* class turned out to be unimportant for this work when screening the full texts (23 items). Consequently, the 83 relevant studies were used for the subsequent step of the taxonomy development (cf. Appendix, Table 2).

3.2 Taxonomy Development

For the systematization of the identified results, we chose a taxonomy development approach. Generally, taxonomies serve as a viable tool for organizing knowledge in a structured manner and manifesting descriptive theories [17]. For this purpose, they enable researchers to study the relationship among concepts and help to analyze and understand complex domains [18].

Particularly, to carry out the taxonomy development, we applied the method proposed by Nickerson et al. [8] as it provides systematic guidance. It basically consists of three steps: i) determining a meta-characteristic, ii) specifying ending conditions, and iii) identifying dimensions and characteristics of the taxonomy.

The meta-characteristic is the root element, as it serves as a foundation for the choice of all other characteristics. Thus, it was defined in accordance with our research goal *to identify characteristic components of 3DOD solutions*. The specification of ending conditions, on the other hand, is required due to the iterative method character. For this purpose, certain subjective criteria must be fulfilled, e.g., that a taxonomy is sufficiently robust to contain enough dimensions and characteristics to separate between the objects of interest, while it is sufficiently concise to not exceed the cognitive load of the taxonomy user [8]. Moreover, the method requires the specification of objective ending conditions, e.g., that every characteristic within its dimension is unique and not repeated. At this point, we adopted the following four criteria for our approach: i) all objects were examined, ii) at least one object can be assigned for each characteristic across all dimensions, iii) no new dimensions or characteristics were added in the last iteration, and iv) no dimensions or characteristics were modified in the last iteration.

The actual step of identifying dimensions and characteristics can then be carried out either with an *empirical-to-conceptual (E2C)* or a *conceptual-to-empirical (C2E)* path. We applied a combination of both paths by running several iterations until all ending conditions were met. More specifically, we followed the suggestion from Zschech [19] to consider principal phases of procedure models from the field of data

https://doi.org/10.30844/wi_2020_r2-friedrich

mining such as CRISP-DM when extracting taxonomic dimensions and characteristics. Such procedure models are basically divided into aspects related to *domain properties*, *data properties*, *data preparation steps*, *modeling techniques* and *evaluation criteria* [20]. Thus, this distinction serves as a helpful orientation, since any kind of predictive modeling or supervised ML solution can be well described and structured on this basis.

Accordingly, in a **first iteration**, all 83 documented solutions were differentiated according to the addressed domain objects by applying an E2C-path. This was done to support the understanding of business goals and addressed sectors. While most articles specifically focus on one particular domain, some approaches also provide *comprehensive* solutions to cover a broader variety. In the **next two iterations**, we extracted relevant data properties of the developed solutions, which could be organized within the two dimensions *dataset* and *sensor*. For the first dimension, we applied an E2C-path classifying empirically observed datasets. For the sensor dimension, on the other hand, we applied the categorization suggested by Arnold et al. [6] distinguishing between *monocular cameras, stereo cameras and LiDAR sensors*, which could also be confirmed empirically. In a **fourth iteration**, we considered all data representation alternatives by applying a C2E-path based on the understanding of Arnold et al. [6]. Thus, methods for 3DOD solutions can either be based on *monocular images*, *point clouds* or a *fusion* of these modalities. Moreover, point clouds could further be subdivided into three sub-categories, which will be discussed in the result section. In a **fifth iteration**, we empirically extracted different modeling techniques reflecting the role of deep neural networks vs. handcrafted feature modeling. We then proceeded in the **sixth iteration** to extract different types of evaluation criteria for 3DOD model assessment based on another E2C-path. Finally, in a **last iteration**, all studies were screened again and since no more modifications occurred, all ending conditions were met to complete the taxonomy development process.

4 Results

4.1 Quantitative Overview

A quantitative analysis of the identified literature corpus reveals that the 83 studies have been published between 2012 and 2019 (cf. Figure 1). While the majority of studies (45 publications) are part of conference proceedings, only 8 studies have been published in journals. The remaining 30 studies are currently only available as preprints on arXiv, waiting for their approval through peer review. Moreover, it is noticeable that the number of studies tremendously increased within the last six years with a remarkable jump from 9 articles in 2017 to 23 articles in 2018 and up to 36 articles in 2019, which illustrates the growing interest and rising importance of the emerging field.

https://doi.org/10.30844/wi_2020_r2-friedrich

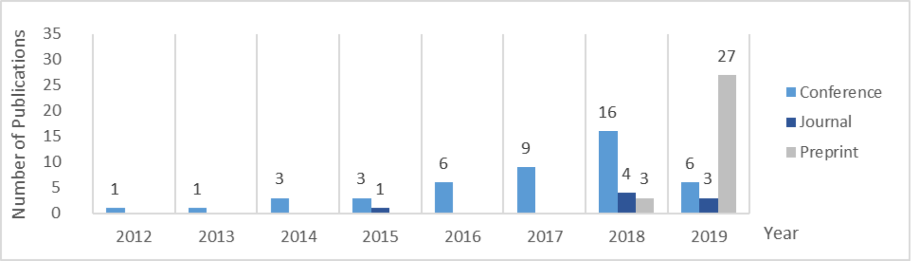

Figure 1. Number of Publications Proposing Solutions for 3DOD

4.2 Systematization of 3DOD Solutions via Taxonomic Framework

In this section, we systematize the field of 3DOD according to our derived taxonomy. We briefly describe each characteristic in each dimension and refer to selected examples. Additionally, in Table 2 of the Appendix, we list all 83 examined studies with their corresponding characteristics. To find the full references for each study, please refer to the following link: https://www.researchgate.net/publication/337802121

Table 1. Taxonomy for 3D Object Detection Solutions

<table>
<tr><th>Dimension</th><th colspan="11">Characteristics</th></tr>
<tr><td>Domain</td><td colspan="3">Indoor</td><td colspan="4">Autonomous Driving</td><td colspan="4">Comprehensive</td></tr>
<tr><td>Dataset</td><td colspan="2">NYUv2</td><td colspan="3">SUN RGB-D</td><td colspan="3">LineMOD</td><td colspan="3">KITTI</td></tr>
<tr><td>Sensor Data</td><td colspan="3">Mono</td><td colspan="4">Stereo</td><td colspan="4">LiDAR</td></tr>
<tr><td rowspan="2">Data Preparation (Representation)</td><td rowspan="2">Monocular</td><td colspan="8">Point Cloud</td><td colspan="2" rowspan="2">Fusion</td></tr>
<tr><td colspan="3">Projection</td><td colspan="2">Volumetric</td><td colspan="3">PointNet</td></tr>
<tr><td rowspan="2">Modeling</td><td colspan="5">Handcrafted Feature Engineering</td><td colspan="6">Deep Learning</td></tr>
<tr><td colspan="2">Keypoint Features</td><td colspan="3">HOG/COG Feat.</td><td colspan="2">Two-Stage</td><td colspan="3">One-Stage</td><td>FE Only</td></tr>
<tr><td>Evaluation</td><td colspan="3">Precision-based</td><td colspan="4">Time-based</td><td colspan="4">Memory-based</td></tr>
</table>

Domain: Current applications for 3DOD can generally be divided into the two domains of *autonomous driving* and *indoor applications,* with the first category being the more frequently studied domain (54 vs. 24 studies). Moreover, another 5 studies also provide *comprehensive* solutions that do not focus on a specific scenario. A fundamental difference between indoor applications and autonomous driving (AD) is that objects in environments like apartments or offices are often arranged one above the other. Based on this fact, inter-object relationships can be learned to consider spatial co-occurrences [21]. Moreover, methods for holistic scene understanding are applied to enable a better communication between, e.g., service robots and humans [22]. Challenges for indoor applications are that scenes are often cluttered and that many objects occlude one another [23]. AD applications, by contrast, are subject to long distances to potential objects and difficult weather conditions, including snow, rain and fog [6]. Objects also occlude one another, but due to the observation that objects like cars, pedestrians and traffic lights are unlikely to be positioned on top of

https://doi.org/10.30844/wi_2020_r2-friedrich

each other, techniques such as the bird's-eye view projection of a scene can efficiently compensate for this effect [24, 25].

Datatset: The second dimension refers to applied datasets that are used to train and, in most cases, evaluate 3DOD solutions. Here, it can basically be distinguished between three datasets representing indoor scenes, i.e., *NYUv2* [26], *SUN RGB-D* [27] and *LineMOD* [28], whereas the *KITTI* dataset [29] focuses on street environments. The goal of NYUv2 and SUN RGB-D is to encourage methods focused on total scene understanding. The datasets got recorded using four different RGB-D sensors to ensure the generalizability of applied methods for different sensors. Even though, SUN RGB-D inherited the 1449 labeled RGB-D frames from the NYUv2 dataset, it is still occasionally used by nowadays contributions. SUN RGB-D consists of 10.335 RGB-D images that are labelled with 146.617 2D polygons and 64.595 3D bounding boxes with accurate object orientation measures. The LineMOD dataset focuses on individual objects in cluttered environments. It consists of 13 RGB-scenes of an object of interest and the corresponding 6DoF ground truth labels as well as computer-aided design (CAD) models. KITTI consists of stereo images, LiDAR point clouds and GPS coordinates, all synchronized in time. Recorded scenes range from highways, complex urban areas and narrow country roads. For 3DOD, KITTI provides 7.481 training and 7.518 test frames, including sensor calibration information and annotated 3D bounding boxes around the objects of interest. KITTI is by far the most frequent used dataset with 59 studies. Of the three datasets that address indoor applications, SUN RGB-D is the most commonly used dataset followed by NYUv2 and LineMOD (15/8/4 studies).

Sensor: The next dimension covers the specific sensor data selected from the datasets, which are used for subsequent modeling. Here, either conventional *monocular* sensor data is used (45 studies) or depth data from *stereo* (30 studies) or *LiDAR* sensors (34 studies). As it can be noticed by the numbers, some solutions are even based on multiple sensors for the purpose to improve detection results (e.g., [5]). While LiDAR sensors generally have the advantage to provide highly precise depth information, simpler image sensors are significantly cheaper, they can capture scenes up to 100 meters, and they are often already in use in operating environments [30, 31].

Data Representation: Based on the way data is represented, 3DOD solutions can be subdivided into three basic approaches: *Monocular-based*, *point-cloud-based* and *fusion-based* methods [6]. Monocular-based methods utilize RGB images acquired by monocular cameras to predict 3D bounding boxes. Since depth data is not available, most approaches first detect 2D candidates before predicting a 3D bounding box representing the object. This is done using either neural networks (e.g., [32]), geometric constraints (e.g., [33]) or 3D model matching (e.g., [34]). Point-cloud-based methods are based on point clouds that are either generated by LiDAR sensors or stereo cameras. They can be further subdivided into *projection*, *volumetric* and *PointNet* methods [6]. Some approaches project point clouds onto depth maps (also called front-view projection) (e.g., [35]), while others project them onto the ground plane using bird's-eye projection techniques (e.g., [24]). Volumetric methods first encode point clouds to a sparse, volumetric voxel grid before processing them (e.g.,

https://doi.org/10.30844/wi_2020_r2-friedrich

[36]). Approaches utilizing the PointNet-architecture [37] process raw point clouds without the need of projection or voxelization (e.g., [38]). Fusion-based methods combine two or more sensor inputs to improve the overall performance of 3DOD (e.g., [5]). Most approaches can further be classified as either *early fusion*, *late fusion* or *deep fusion* [15]. In early fusion, the different sensor inputs are combined at the beginning of the pipeline. This results in a new representation that is dependent on all inputs. When performing late fusion, the sensor inputs are processed independently up until the last stage in the pipeline. This results in a complete or partial autonomy of the particular input channels. Accordingly, deep fusion allows an interaction of the input modalities at several stages within the architecture. This enables the exchange and adjustment of features from different input types resulting in a better model generalizability [15]. Even though, monocular-based methods utilize no depth information, they are commonly applied (21 studies). Considering the different point-cloud-based methods, projection and volumetric approaches are used far more often than PointNet-based methods (19/14/5 studies). However, fusion-based methods often use point cloud data processed by PointNets in combination with monocular or even other point-cloud-based inputs (24 studies).

Modeling: In the modeling step, the input data in their respective representation format are used together with annotated label information (i.e., 3D bounding boxes) to train a 3DOD model in a supervised learning manner. For this purpose, different methods are applicable, where it can basically be distinguished between approaches based on *handcrafted feature engineering* and *deep learning (DL)*. For the first category, it is necessary to first define features and then use algorithms like support vector machines for the classification task. Handcrafted 3DOD features are either based on *keypoints* (i.e., remarkable points that best describe an object) or *histograms of oriented gradients (HOG)* in 2D spaces and its counterpart *clouds of oriented gradients (COG)* in 3D spaces, where the appearance and the shape of objects can be represented by the distribution of the local intensity of edges and corners. Both groups are severely underrepresented (2 vs. 5 studies) and mainly applied by older studies. DL techniques, on the other hand, can perform object detection without manually defining specific features in advance due to their ability of automatic feature extraction (FE). They are based on different kinds of convolutional neural networks (CNN), where the complete 3DOD pipeline is either organized as a *one-stage* (16 studies) or a *two-stage* (52 studies) architecture. The latter consists of an additional stage where possible object regions are proposed to reduce the search space in an image before the actual detection takes place. This can increase the accuracy, while simultaneously being more time consuming. Apart from both variants, there is also a minority of 2 studies that exploits the strengths of CNNs only for the dedicated task of *feature extraction*.

Evaluation: The last dimension comprises different categories of evaluation criteria. Since most datasets extracted within the *dataset* dimension primarily serve as benchmark instances, the majority of 3DOD solutions (82 studies) is evaluated on the basis of *precision-based metrics*, such as the average precision as introduced at the Pascal VOC Challenge [39]. Another large proportion of articles (52 studies) is also assessed via *time-based evaluations*. Of particular interest here is the inference time,

https://doi.org/10.30844/wi_2020_r2-friedrich

indicating how long a trained model requires to detect the objects of interest. A last evaluation category refers to the required memory usage of the trained model, which plays an increasing role in the context of mobile applications [40]. However, only one study explicitly measured required memory usage [41].

5 Conclusion, Discussion and Outlook

In this paper, we developed a systematization for 3DOD end-to-end solutions by applying a taxonomy approach to reach better transparency and decompose complex solutions into more manageable pieces. Reflecting the results of this research, we contribute to the research field of 3DOD in several ways: First, we extended the survey of Arnold et al. [6] by a more recent scope with a broader domain focus and by a comprehensive categorization of all identified 3DOD solutions using a classification matrix (cf. Appendix). This helps the community to get an overview of current trends, where it becomes apparent, for example, that DL methods have surpassed methods that rely on hand-crafted features. Second, by applying a taxonomic approach, it was possible to decompose the complexity of methods for 3DOD to a certain degree, which is necessary for future efforts to compare novel solutions on a more fine-grained basis. By employing the method of Nickerson et al. [8], a number of useful dimensions and characteristics could be extracted, including addressed *domain*, *dataset, sensor,* the overall *data representation, modeling techniques* and *evaluation criteria*. Thus, the taxonomy delivers an overview about different design options and provides a setting to position individual configurations of novel solutions on a more comparable basis.

Our work has also some limitations. Due to the recent public interest in autonomous vehicles and indoor service robots, 3DOD studies have drastically increased in the last two years. Therefore, many studies have not been peer-reviewed yet and thus are only available as preprints, making their final validity somewhat doubtful. Another limitation is the organization of several 3DOD solution components within a simplified framework, which was also noted by previous taxonomy developers in other analytical scenarios (e.g., [19]). Here, we were faced with a fine-grained diversity of different design options, particularly for neural network architectures based on varying network topologies, filter kernels, activation and loss functions, etc., which were difficult to be transferred into a flat taxonomy structure. At this point, it seems more reasonable to allow hierarchical, tree-like categorizations or even create sub-taxonomies for several dimensions. Nevertheless, we are still confident that we reached a suitable level of abstraction which can currently help researchers and practitioners to get a quick overview about the field and organize 3DOD solutions within a structured framework. A last limitation concerns the external evaluation of the taxonomic structure. For this purpose, we plan to conduct interviews with experts from industry to further refine or extend the taxonomy with additional insights. As such, it is also conceivable to consider other sources of interest, such as patents or existing products on the market.

https://doi.org/10.30844/wi_2020_r2-friedrich

In future research, the results will be used to carry out a cluster analysis on the 83 classified studies to identify recurring patterns. Thus, it is intended to find archetypal solutions based on commonly applied combinations and derive prescriptive knowledge towards the creation of preconfigured templates. Moreover, we will use the framework to conduct systematic benchmarking studies following the idea of Zschech et al. [18], where the taxonomic elements serve as evaluation options to be iteratively modified under ceteris paribus conditions. In this way, it is planned to establish a better understanding to what extent certain components affect the results of 3DOD solutions.

References

1. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You Only Look Once: Unified, Real-Time Object Detection. In: arXiv:1506.02640 [cs]. pp. 779–788 (2016).
2. Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X.: Object Detection With Deep Learning: A Review. IEEE Transactions on Neural Networks and Learning Systems. 1–21 (2019).
3. Heinrich, K., Zschech, P., Möller, B., Breithaupt, L., Maresch, J.: Objekterkennung im Weinanbau – Eine Fallstudie zur Unterstützung von Winzertätigkeiten mithilfe von Deep Learning. HMD Praxis der Wirtschaftsinformatik. 56, 964–985 (2019).
4. Heinrich, K., Roth, A., Zschech, P.: Everything Counts: A Taxonomy of Deep Learning Approaches for Object Counting. In: European Conference on Information Systems. Stockholm-Uppsala, Sweden (2019).
5. Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J.: Frustum PointNets for 3D Object Detection from RGB-D Data. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 918–927 (2018).
6. Arnold, E., Al-Jarrah, O.Y., Dianati, M., Fallah, S., Oxtoby, D., Mouzakitis, A.: A Survey on 3D Object Detection Methods for Autonomous Driving Applications. IEEE Transactions on Intelligent Transportation Systems. 1–14 (2019).
7. vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R., Cleven, A.: Reconstructing the Giant: On the Importance of Rigour in Documenting The Literature Search Process. In: European Conference on Information Systems. Verona, Italy (2009).
8. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in information systems. European Journal of Information Systems. 22, 336–359 (2013).
9. Szeliski, R.: Computer Vision: Algorithms and Applications. Springer Science & Business Media (2010).
10. Heinrich, K., Zschech, P., Skouti, T., Griebenow, J., Riechert, S.: Demystifying the Black Box: A Classification Scheme for Interpretation and Visualization of Deep Intelligent Systems. In: Americas Conference on Information Systems. Cancún, Mexico (2019).
11. Davies, E.R.: Computer and Machine Vision: Theory, Algorithms, Practicalities. Elsevier, Amsterdam; Boston (2012).
12. Microsoft: Kinect, https://developer.microsoft.com/en-us/windows/kinect, last accessed 2019/10/10.
13. Velodyne: HDL-64E, https://velodynelidar.com/hdl-64e.html, last accessed 2019/10/10.
14. Otepka, J., Ghuffar, S., Waldhauser, C., Hochreiter, R., Pfeifer, N.: Georeferenced Point Clouds: A Survey of Features and Point Cloud Management. ISPRS Int. J. Geo-Information. 2, 1038–1065 (2013).

https://doi.org/10.30844/wi_2020_r2-friedrich

15. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3D Object Detection Network for Autonomous Driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6526–6534 (2017).
16. Kitchenham, B., Charters, S.: Guidelines for performing Systematic Literature Reviews in Software Engineering. (2007).
17. Gregor, S.: The Nature of Theory in Information Systems. MIS Quarterly. 30, 611–642 (2006).
18. Zschech, P., Bernien, J., Heinrich, K.: Towards a Taxonomic Benchmarking Framework for Predictive Maintenance: The Case of NASA's Turbofan Degradation. In: International Conference on Information Systems. Munich, Germany (2019).
19. Zschech, P.: A Taxonomy of Recurring Data Analysis Problems in Maintenance Analytics. In: European Conference on Information Systems. Portsmouth, UK (2018).
20. Kurgan, L.A., Musilek, P.: A Survey of Knowledge Discovery and Data Mining Process Models. The Knowledge Engineering Review. 21, 1–24 (2006).
21. Lin, D., Fidler, S., Urtasun, R.: Holistic Scene Understanding for 3D Object Detection with RGBD Cameras. In: 2013 IEEE International Conference on Computer Vision. pp. 1417–1424 (2013).
22. Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.-C.: Cooperative Holistic Scene Understanding: Unifying 3D Object, Layout, and Camera Pose Estimation. In: International Conference on Neural Information Processing Systems. pp. 206–217 (2018).
23. Ren, Z., Sudderth, E.B.: Clouds of Oriented Gradients for 3D Detection of Objects, Surfaces, and Indoor Scene Layouts. arXiv:1906.04725 [cs]. 1–14 (2019).
24. Beltrán, J., Guindel, C., Moreno, F.M., Cruzado, D., García, F., Escalera, A.D.L.: BirdNet: A 3D Object Detection Framework from LiDAR Information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC). pp. 3517–3523 (2018).
25. Wang, Z., Zhan, W., Tomizuka, M.: Fusing Bird's Eye View LIDAR Point Cloud and Front View Camera Image for 3D Object Detection. In: 2018 IEEE Intelligent Vehicles Symposium (IV). pp. 1–6 (2018).
26. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C. (eds.) Computer Vision – ECCV 2012. pp. 746–760. Springer Berlin Heidelberg (2012).
27. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 567–576. IEEE, Boston, MA, USA (2015).
28. Hinterstoisser, S., Holzer, S., Cagniart, C., Ilic, S., Konolige, K., Navab, N., Lepetit, V.: Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In: 2011 International Conference on Computer Vision. pp. 858–865. IEEE, Barcelona, Spain (2011).
29. Geiger, A., Lenz, P., Urtasun, R.: Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012).
30. Weng, X., Kitani, K.: Monocular 3D Object Detection with Pseudo-LiDAR Point Cloud. arXiv:1903.09847 [cs]. (2019).
31. Jörgensen, E., Zach, C., Kahl, F.: Monocular 3D Object Detection and Box Fitting Trained End-to-End Using Intersection-over-Union Loss. arXiv:1906.08070 [cs]. 1–10 (2019).
32. Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., Urtasun, R.: Monocular 3D Object Detection for Autonomous Driving. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2147–2156 (2016).

https://doi.org/10.30844/wi_2020_r2-friedrich

33. Mousavian, A., Anguelov, D., Flynn, J., Košecká, J.: 3D Bounding Box Estimation Using Deep Learning and Geometry. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5632–5640 (2017).
34. Zakharov, S., Shugurov, I., Ilic, S.: DPOD: 6D Pose Object Detector and Refiner. arXiv:1902.11020 [cs]. (2019).
35. Qin, Z., Wang, J., Lu, Y.: Triangulation Learning Network: from Monocular to Stereo 3D Object Detection. arXiv:1906.01193 [cs]. 1–19 (2019).
36. Sun, H., Meng, Z., Du, X., Ang, M.H.: A 3D Convolutional Neural Network Towards Real-Time Amodal 3D Object Detection. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8331–8338 (2018).
37. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: arXiv:1612.00593 [cs]. pp. 652–660 (2017).
38. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep Hough Voting for 3D Object Detection in Point Clouds. arXiv:1904.09664 [cs]. (2019).
39. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. International Journal of Computer Vision. 88, 303–338 (2010).
40. Heinrich, K., Janiesch, C., Möller, B., Zschech, P.: Is Bigger Always Better? Lessons Learnt from the Evolution of Deep Learning Architectures for Image Classification. In: Pre-ICIS SIGDSA Symposium. Munich, Germany (2019).
41. Maisano, R., Tomaselli, V., Capra, A., Longo, F., Puliafito, A.: Reducing Complexity of 3D Indoor Object Detection. In: 2018 IEEE 4th International Forum on Research and Technology for Society and Industry (RTSI). pp. 1–6 (2018).

Appendix

Table 2. Application of the Taxonomy on Reviewed Studies

				Domain			Dataset				Sensor			Data Representation					Modeling					Evaluation		
ID	Authors	Ref.	Pseudonym (* given by the authors of this paper)	Indoor	Autonom. Driv.	Comprehensive	SUN RGB-D	NYUv2	LineMOD	KITTI	Mono	Stereo	LiDAR	Monocular	PC-Projection	PC-Volumetric	PC-PointNet	Fusion	DL-Two-Stage	DL-One-Stage	DL-FE-Only	Keypoint Feat.	HOG/COG Feat.	Precision	Time	Memory
1	Fidler et al. (2012)	[42]	3DCuboidDPM*			x				x	x			x									x	x		
2	Lin et al. (2013)	[21]	3D-CPMC*	x				x				x				x								x	x	
3	Song and Xiao (2014)	[43]	Sliding Shapes	x				x				x				x					x			x		
4	Tejani et al. (2014)	[44]	L-C HoughForest*	x					x			x			x							x		x		
5	Teng and Xiao (2014)	[45]	SB-3D*	x								x			x							x		x		
6	Chen et al. (2015)	[46]	3DOP		x					x		x			x				x					x	x	
7	Crivellaro et al. (2015)	[47]	Cluttered3DPose*	x							x			x					x							
8	Geiger and Wang (2015)	[48]	Joint3D*	x				x				x			x									x	x	
9	Zia et al. (2015)	[49]	3DRepresentations*		x					x	x			x										x		
10	Chen et al. (2016)	[32]	Mono3D		x					x	x			x					x					x		
11	Kehl et al. (2016)	[50]	Patch3D*	x					x			x			x					x				x	x	
12	Li (2016)	[51]	3D-FCN*		x					x			x			x				x				x	x	
13	Li et al. (2016)	[52]	VeloFCN		x					x			x		x					x				x		
14	Ren and Sudderth (2016)	[53]	3D-COG 1.0*	x			x					x				x							x	x	x	
15	Song and Xiao (2016)	[54]	DSS	x			x	x				x				x			x					x	x	
16	Chabot et al. (2017)	[55]	Deep MANTA		x					x	x			x					x					x	x	
17	Chen et al. (2017)	[15]	MV3D		x					x	x		x					x	x					x		
18	Deng and Latecki (2017)	[56]	Amodal3D*	x				x			x	x						x	x					x		
19	Engelcke et al. (2017)	[57]	Vote3Deep		x					x			x			x				x				x	x	
20	He et al. (2017)	[58]	3DTemplateMatch*	x					x			x			x									x		
21	Kim and Kang (2017)	[59]	CCD R-FCN*		x					x	x		x					x	x					x		
22	Kim et al. (2017)	[60]	3DMulti-Frame*		x					x			x		x				x					x		
23	Lahoud and Ghanem (2017)	[61]	2D-3D*	x			x				x	x						x	x					x	x	
24	Mousavian et al. (2017)	[33]	Deep3DBox		x					x	x			x					x					x		
25	Beltrán et al. (2018)	[24]	BirdNet		x					x			x		x				x					x	x	
26	Chen et al. (2018)	[62]	TwoStream3D*		x					x		x			x				x					x	x	
27	P. d. l. Garanderie et al. (2018)	[63]	360Panoramic*		x					x	x			x					x					x		
28	Huang et al. (2018)	[22]	3D-OLC*	x			x				x			x					x					x	x	
29	Ku et al. (2018)	[64]	AVOD		x					x	x		x					x	x					x	x	
30	Liang et al. (2018)	[65]	DeepContinousFusion*		x					x	x		x					x	x					x	x	
31	Liu et al. (2018)	[66]	3D SS	x			x					x				x							x	x	x	
32	Maisano et al. (2018)	[41]	MobileNet-3D*	x				x			x	x						x	x					x	x	x
33	Qi et al. (2018)	[5]	Frustum PointNets			x	x			x	x	x	x					x	x					x	x	
34	Ren et al. (2018)	[67]	C3D	x			x				x	x						x	x					x		
35	Ren and Sudderth (2018)	[68]	3D-LSS*	x			x					x				x							x	x	x	
36	Shi et al. (2018)	[69]	PointRCNN		x					x			x				x		x					x	x	
37	Shin et al. (2018)	[70]	RoarNet		x					x	x		x					x	x					x		
38	Sun et al. (2018)	[36]	3D-CNN*	x			x	x				x				x				x				x	x	

https://doi.org/10.30844/wi_2020_r2-friedrich

39	Wang et al. (2018)	[25]	FuseBEV-FV*		x					x	x		x					x		x				x	x	
40	Xu and Chen (2018)	[71]	3D-MLF*		x					x	x			x					x					x	x	
41	Xu et al. (2018)	[72]	PointFusion			x	x			x	x	x	x					x	x					x	x	
42	Yamazaki et al. (2018)	[73]	3D-CORR*	x								x			x									x	x	
43	Yan et al. (2018)	[74]	SECOND		x					x			x			x			x					x	x	
44	B. Yang et al. (2018)	[75]	PIXOR		x					x			x		x					x				x	x	
45	Z. Yang et al. (2018)	[76]	IPOD		x					x	x		x					x	x					x	x	
46	Zeng et al. (2018)	[77]	RT3D		x					x			x		x				x					x	x	
47	Zhou and Tuzel (2018)	[78]	VoxelNet		x					x			x				x			x				x	x	
48	Ali et al. (2019)	[79]	YOLO3D		x					x			x		x					x				x		
49	Barabanau et al. (2019)	[80]	Keypoint3D*		x					x	x			x					x					x		
50	Brazil and Liu (2019)	[81]	M3D-RPN		x					x	x			x					x					x	x	
51	Chen et al. (2019)	[82]	Cooper		x					x	x					x			x					x		
52	Ferguson and Law (2019)	[83]	Object R-CNN*	x								x			x				x					x		
53	Gupta et al. (2019)	[84]	KeypointCBF*		x					x	x			x						x				x		
54	Huang et al. (2019)	[85]	3DRestoration*		x					x	x	x						x	x					x		
55	Jörgensen et al. (2019)	[31]	SS3D		x					x	x			x						x				x	x	
56	Ku et al. (2019)	[86]	MonoPSR		x					x	x			x					x					x	x	
57	B. Li et al. (2019)	[87]	GS3D		x					x	x			x						x				x	x	
58	M. Li et al. (2019)	[88]	Complex-Retina		x					x	x		x					x		x				x	x	
59	P. Li et al. (2019)	[89]	Stereo R-CNN		x					x		x			x				x					x	x	
60	X. Li et al. (2019)	[90]	3DBN		x					x			x			x			x					x	x	
61	Liu et al. (2019)	[91]	FQNet		x					x	x			x					x					x	x	
62	Lu et al. (2019)	[92]	SCANet		x					x	x		x					x	x					x	x	
63	Ma et al. (2019)	[93]	Color-Embedded3DRecon*		x					x	x			x					x					x		
64	Meyer et al. (2019a)	[94]	SensorFusion*		x					x	x		x					x	x					x		
65	Meyer et al. (2019b)	[95]	LaserNet		x					x			x		x					x				x		
66	Naiden et al. (2019)	[96]	ShiftNet		x					x	x			x					x					x	x	
67	Pamplona et al. (2019)	[97]	On-road3D*		x					x			x				x		x					x		
68	Qi et al. (2019)	[38]	VoteNet	x			x					x					x		x					x	x	
69	Qin et al. (2019)	[35]	TLNet		x					x		x			x				x					x		
70	Ren and Sudderth (2019)	[23]	3D-COG 2.0*	x			x					x				x					x		x	x	x	
71	Shi et al. (2019)	[98]	Part-A^2 Net		x					x			x				x		x					x	x	
72	Simon et al. (2019a)	[99]	Complex-YOLO		x					x			x		x				x					x	x	
73	Simon et al. (2019b)	[100]	Complexer-YOLO		x					x	x		x					x		x				x	x	
74	Simonelli et al. (2019)	[101]	MonoDIS		x					x	x			x						x				x	x	
75	Sindagi et al. (2019)	[102]	MVX-Net		x					x	x		x					x		x				x	x	
76	Srivastava et al. (2019)	[103]	3D-GAN*		x					x	x							x	x					x		
77	Tang and Lee (2019)	[104]	Transfer3D*			x	x			x	x	x	x					x	x					x		
78	B. Wang et al. (2019)	[105]	Voxel-FPN		x					x			x			x			x					x	x	
79	L. Wang et al. (2019)	[106]	3D MC-CNN*	x			x	x			x	x						x	x					x	x	
80	Wang and Jia (2019)	[107]	F-ConvNet			x	x			x	x	x	x					x	x					x	x	
81	Weng and Kitani (2019)	[30]	PseudoLidar*		x					x	x			x					x					x		
82	Zakharov et al. (2019)	[34]	DPOD	x					x		x			x						x				x	x	
83	Zhou et al. (2019)	[108]	FVNet		x					x			x					x	x					x	x	
Total Number of Coverage				**24**	**54**	**5**	**15**	**8**	**4**	**59**	**45**	**30**	**34**	**21**	**19**	**14**	**5**	**24**	**52**	**16**	**2**	**2**	**5**	**82**	**52**	**1**

https://doi.org/10.30844/wi_2020_r2-friedrich