JOANNEUM RESEARCH DIGITAL

Institute for
Digital Technologies

Semantic Interpretation of 3D Point Clouds

DI Dr. Olaf Kähler

Introduction

This paper summarizes the application areas and state-of-the-art of 3D point cloud interpretation, in particular algorithms for 3D object detection and 3D semantic segmentation. It starts by highlighting the main advantages of interpreting point cloud data directly, instead of using image interpretations as an indirect helper. Next, it underlines that a wide range of representations for 3D point clouds have been tried and tested to achieve efficient processing in neuronal networks. Finally, two case studies are shown where the algorithms are applied to large-scale railway and road datasets. In both cases, the JOANNEUM RESEARCH workflow for point cloud interpretation proved to be highly successful in detecting and classifying relevant assets.

Motivation

Nowadays it is common to record 3D digital models of critical assets and in particular infrastructure. With recent advances in recording and data processing technologies, such geometric Digital Twins provide representations with unprecedented detail and accuracy. Given sufficient data recording and storage capacities, these geometric Digital Twins can be pushed even to large scale applications, such as digital representations of entire motorways, railway lines or similar structures.

The enabling technology for such processes is mobile mapping. Today, a variety of sensor systems is available on the market, which can be mounted on UAVs, cars, boats, or can be carried by hand. They typically deliver high precision 3D point clouds, often already colorized, and additionally high-resolution image data. Both are georeferenced and thereby allow deriving the absolute location of individual measurements up to a precision of few centimetres.

However, while the purely geometric representation forms a solid basis for manual data exploration and planning, additional semantic information is required to enable automated functions like search, data filtering and data linking. This level of semantic enrichment is crucial for establishing large-scale digital solutions for asset management, predictive maintenance and simulations. Methods from Machine Learning and Artificial Intelligence are ideally suited to automate the generation of this semantic information, even with respect to very specific application requirements. In Machine Learning models, a distinction is made between methods for object detection and methods for semantic segmentation. Object detection has the goal of coarsely detecting an object in a dataset, with high efficiency. Semantic segmentation intends to associate each individual measurement (3D point or image pixel) to a specific class such as “rail” or “mast”. In both cases, the unique labels can be further linked to data like BIM, plans, component lists or legal documents.

How to perform object detection or semantic segmentation on Digital Twin models from mobile mapping?

One option is to apply 2D image-processing on the available data. There are well-known image-based object detectors and segmentation algorithms available, which can be applied with comparably little effort. However, taking the next step from an image-based object detection to a precise geo-location is more complex. It either requires the triangulation of very precise object locations using an equally precise camera calibration, or searching for all 3D-points, which can be associated to a detected object in the image. Moreover, image-based detection and segmentation are highly sensitive to changes of perspective. An AI-algorithm trained to detect vehicles from a side-view will inevitably fail on images taken from a bird's eye view.

Algorithms operating directly on 3D point clouds offer a superior alternative with several advantages. First, they require only the 3D dataset to operate. This typically makes up only a small fraction of the data volume, compared to an equivalent image dataset. Second, results on the 3D dataset give the desired geo-referenced information directly, without the need for further processing. Third, 3D point clouds from modern sensors differ very little between different recording platforms. Therefore, the methodology and trained models for processing e.g. train-based scans directly transfer to UAV-based scans. Last but not least, 3D point cloud data naturally carries spatial information about e.g. neighbourhoods and foreground vs. background objects, that is more difficult to infer from RGB and depth images.

3D Data Interpretation

At a first glance, establishing the semantic interpretation algorithms purely based on 3D point clouds seems somewhat more complicated: image-based computer vision tasks are by and large solved with convolutional neuronal networks, or massively profit in other ways from the efficiency arising from densely sampled neighbourhoods in pixel space. Such technologies do not easily transfer to 3D point clouds, as these are neither densely sampled, nor arranged in a fixed grid by default. Therefore, a different machine-learning technology must be applied, especially in terms of a suitable 3D scene representation. This report shows that the relevant technologies not only exist in the literature, but that they are mature and ready for application in industrial settings, even at large scale.

3D Semantic Interpretation Workflows

For 2D images, object detection is a fundamental and widely applicable computer vision task. Whether it is for counting vehicles in traffic applications, focusing on faces in photography or finding objects in warehouse shelves, object detection is about identifying and localizing individual instances of known object categories. In contrast, differentiating vegetation vs. roads, finding outlines of a tray or identifying warehouse shelves requires semantic segmentation approaches due to the lack of defining shapes and extents in the relevant structures. For 2D images, semantic segmentation assigns a unique class label to each pixel in an image.

The same fundamental goals are relevant for interpreting 3D point clouds. Accordingly, 3D object detection algorithms have been developed for identifying vehicles and pedestrians in autonomous driving tasks, objects in a tray for bin-picking or catenary masts along railway lines for surveying applications. For such tasks, oriented bounding boxes around the objects convey all relevant information for further processing. Likewise, semantic segmentation algorithms for 3D point clouds identify individual 3D points belonging to vegetation or ground, to railheads or catenary wires in scans from mobile mapping data, or the exact shape of clothes for bin picking applications.

Depending on the exact recording conditions, the 2D approaches can have advantages when it comes to identifying very small details, as the resolution of mobile mapping systems is somewhat coarser. Apart from that, i.e. whenever the relevant structures are represented in the point cloud, algorithms working directly on the 3D data offer a range of advantages. The most obvious advantage of 3D interpretation workflows is their ability to predict 3D positions directly and without requiring an aggregation step for potentially conflicting detections in individual images. This completely avoids an error prone additional processing step, for which standardized off-the-shelf solutions do not exist. Additionally, the depth information, which is lost in projective 2D images, offers valuable additional cues for delineating the outlines of a wide range of objects. E.g., the outlines of traffic signs next to a road are directly available in 3D point cloud information, whereas in 2D images they have to be extracted carefully. This can have a significant impact on the overall reliability and accuracy of the respective algorithms. The 3D representation also offers a significant degree of invariance with respect to recording techniques. While separate detection models are required for handling airborne and ground-based 2D image data, a single, combined model is often able to handle 3D data from a wide range of recording platforms, which reduces model maintenance effort and improves generalization of the learned representations. Last but not least, while 2D images offer high resolution and potentially multiple observations of the same objects, this redundant information is inherently condensed to the essential minimum in 3D point cloud representations. Throughout the entire process of recording, transferring and interpreting data, the reduced data volume can lead to significant simplifications of all steps.

Processing methods for 3D data

When applying machine learning algorithms to 3D point clouds, one of the first and most urgent questions is about the representation of 3D data and point neighbourhood relations in the algorithms. While traditional, hand-crafted feature extraction algorithms exist, it is obvious that the algorithms benefit from automatically optimized features, as 2D deep learning creates them for images. However, 3D point clouds are neither densely sampled, nor arranged in a fixed grid like image pixels, and processing such data in a computationally efficient way is not trivial. Recent, ground-breaking algorithms in the literature allow handling this huge amount of data.

Training Data

The exact amount of training data required for a specific application is hard to quantify, because inherent object properties such as their shape, uniqueness and variability play an important role in defining the complexity of a data interpretation task. For example, using 3D data it is relatively easy to detect traffic signs next to a road, because they simply stick out from the ground. More samples are needed to tell apart round traffic signs from triangular ones, and yet more to tell apart individual types of traffic signs. As it will be shown in the case studies below, a few hundred training samples are in many cases sufficient to train a 3D object detector to a decent performance level, and reasonable prototypes require even less. For semantic segmentation it is likewise shown that a few tens of annotated 50x50m tiles are sufficient to reach initial models.

With trained and successfully evaluated models, it still remains an open question how to bring in additional training data during operation. It is hard to guarantee that training and test data used during development will remain representative for a long-time real world deployment. So the question is how to adapt the already trained method efficiently to the data from a novel scenario, which is identified during operation. If sample data from the real world deployment becomes available, the complete dataset can be structured into multiple domains of data. The data available during initial development can be considered as the source domain, and data from real world deployment as target domain. Accordingly, the difference in performance between source and target domains is defined as domain gap, and adjustments to the trained model to improve the performance on the target domain are considered as domain adaptation. With every deployment to a novel domain, such as deployment in a novel country, there may be new domain gaps, but with more and more data available, the unexpected novelty in data will decrease. If additional user input or expert knowledge is available, such as feedback provided from end users of the derived data or knowledge about specific new varieties of objects in a newly recorded dataset, these can be used for more targeted retraining and adaptation steps.

Case Study 1: 3D Detection of Railway Assets

In the following, a practical example of 3D object detection for identifying catenary pylons along railway lines is shown. Although catenary pylons are very outstanding structures along the track, a simple hand-crafted pole-detector will fail to discriminate between pylons, lighting poles, construction elements in stations and trees. As a consequence, training a machine-learning method is reasonable and gives a productivity boost in comparison to manual selection of all pylons, as well as to filtering false positives of a pole detector.

Dataset

Two different datasets consisting of aggregated, high density point cloud data from a moving line-scan LiDAR were recorded. The first dataset is a point cloud as from SoA mobile mapping systems, in this case the Leica Pegasus system mounted on a train going along a section of railway line. The second dataset was created by a Riegl LiDAR system mounted on a multicopter drone as a representative of aerial LiDAR data. The datasets were recorded on different railway lines and at different times. In both cases, the data pre-processing with all subsequent registration and georeferencing steps have been carried out in accordance with industry standards. The dataset recorded from the perspective of a train is spanning a total of 18.3km, and the aerial dataset extends over 6.5km.

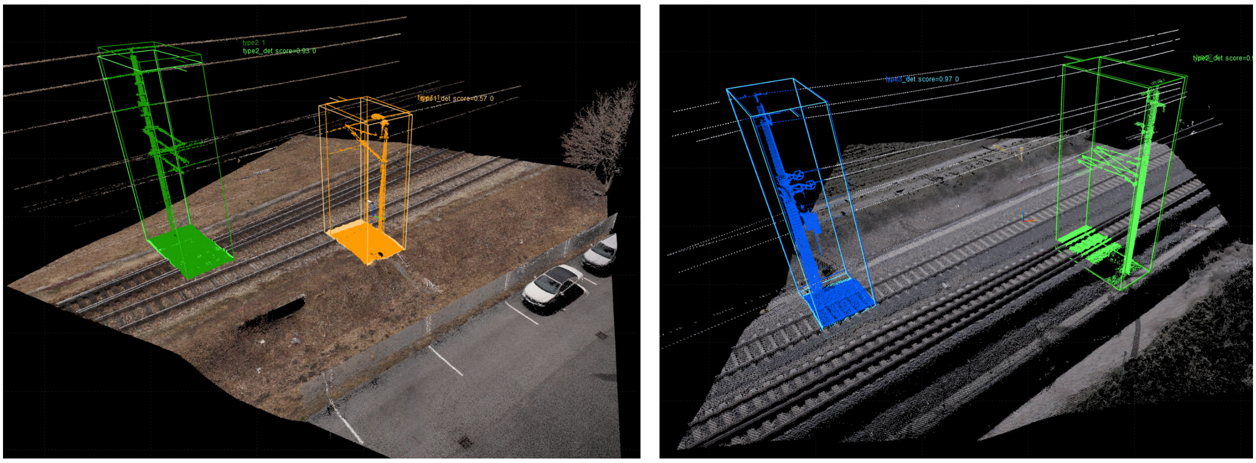

Figure 1: Samples of the dataset for detection of catenary masts along railway lines, along with the detected bounding boxes.

The JOANNEUM RESEARCH workflow is designed to handle virtually unlimited dataset sizes. It follows a partitioning strategy along a given track to allow for a seamless and efficient object detection. The three most frequently occurring catenary mast types were annotated, as shown in Figure 1. The annotation format, consist of boxes with the parameters (x, y, z, dx, dy, dz, orientation), which can be comfortably specified in a 3D labelling environment. In the train dataset this resulted in 4089 mast examples and in the smaller drone dataset, there were 675 mast examples. The labelled data was split into fixed, non-overlapping parts of 75% for training and 25% for evaluation to ensure that no evaluation data is used for the training.

Training and Evaluation

Given sufficient data as stated above, the JOANNEUM RESEARCH 3D object detection workflow provides excellent detection results. Its computational efficiency supports the overall scalability of the method and allows to train new models at fast turnover rates. On a single GPU, the inference for one tile takes about 1 second or less and the training process typically finishes in a few hours. During training, an automated data augmentation is considered as key for improving the performance of a detector, especially on small datasets with few labels.

Results

Table 1 shows the results of various training and evaluation setups. As an evaluation metric the mean average precision metric was used, which is standard for both 2D and 3D object detection tasks. From the results, it becomes clear that nearly perfect 3D object detection is possible for this scenario and dataset, as long as the training data is representative of the evaluation data. If there is a domain shift, this gap can be bridged by adding additional samples of the target domain into the training process.

As one scenario of domain shift changes in recording technology were analysed. If the 3D object detector is trained exclusively on data recorded with the train-based mobile mapping system, then it still generalizes fairly well to aerial data, and the performance drops only from 99.26% to 95.40%. With only 10 additional training samples from the target domain, the performance recovers to 98.61%. This is a confirmation that the dense, aggregated 3D point cloud representations are inherently robust and in many cases, the same relevant details can be spotted regardless of the recording technology.

In a second scenario, changes in object appearance were investigated. It was assumed that during training, only objects of one certain variety (that is, Type 1) is encountered, and during evaluation new varieties (that is, Type 2 and Type 3) are encountered, which were not anticipated. In this case, the domain gap is significantly larger, i.e. the performance drops from 99.78% to merely 39.12%. Retraining with additional samples becomes necessary, and after adding 100 samples of the new varieties, the performance increases to 74.70%, being quite reasonable. This drop in performance is significantly lower, if the intersection-over-union threshold is set to 0.5, indicating that it is mostly the exact prediction of size and location that needs the additional training. In this particular case, additional samples are selected at random, and thus, more sophisticated selection strategies could potentially achieve higher performances with fewer extra annotations.

Training Dataset (# samples)		Evaluation Performance (mAP)	
Train Dataset	UAV Dataset	Train Dataset	UAV Dataset
1710	-	99.26%	95.40%
1710	10	98.95%	98.61%
1710	50	98.78%	100.00%
1710	495	99.41%	100.00%

Training Dataset (# samples)		Evaluation Performance (mAP)	
Typ1	Typ2/3	Typ1	Typ2/3
1422	-	99.78%	39.12%
1422	20	99.54%	48.82%
1422	100	99.78%	74.70%
1422	500	99.85%	97.31%
1422	784	99.70%	97.55%

Table 1: Evaluation of 3D object detection for different numbers of point cloud samples in the source and target domains. In the left table, source and target domains are separated by recording technology, in the right table by variants of catenary masts. The mean average precision (mAP) metric is shown for an intersection-over-union threshold of 0.7.

Overall, a complete database of catenary pylons could be established with less than one hour of processing time. The redundancy offered by small overlap areas allows the system to automatically correct all mistakes near tile boundaries. Manual work is reduced to adding and correcting only two pylons afterwards. The effort to create the pylon detector was approximately 50 hours of manual work including annotation effort and 16 hours of processing time on a standard PC. It is important to note that the created detector can directly be applied to the entire national rail network and easily adapted to international rail networks.

Case Study 2: 3D Semantic Segmentation of Asphalted Areas

In the following, a system for automatically identifying asphalted areas is investigated. Individual 3D points have to be assigned either to the road surface or background classes. This is a first step of an automated pipeline for road quality assessment, where particularly the cross section of the road surface is assessed in later processing steps to identify rutting and to ensure the cross slope supports the specified water runoff. Manually identifying the exact outlines of the asphalted areas is very time-consuming and not feasible for larger sections of road. Hence, an automated solution based on 3D semantic segmentation is introduced.

Dataset

A large stretch of cross-country roads was recorded with a car-mounted mobile mapping system from Riegl. The moving LiDAR line scanner data was pre-processed, geo-referenced and aggregated into a high-density point cloud. To reduce the effort, only 2.2 km of road, covered by about 340 million points, were manually annotated and only the two classes “asphalt” and “others” were used to evaluate this use case.

To process the data, a tiling strategy was applied, allowing the JOANNEM RESEARCH workflow to operate on virtually unlimited datasets. The dataset was split to use 70% of the annotated data for training and the remaining 30% for evaluation purposes, again ensuring distinct datasets with no overlap.

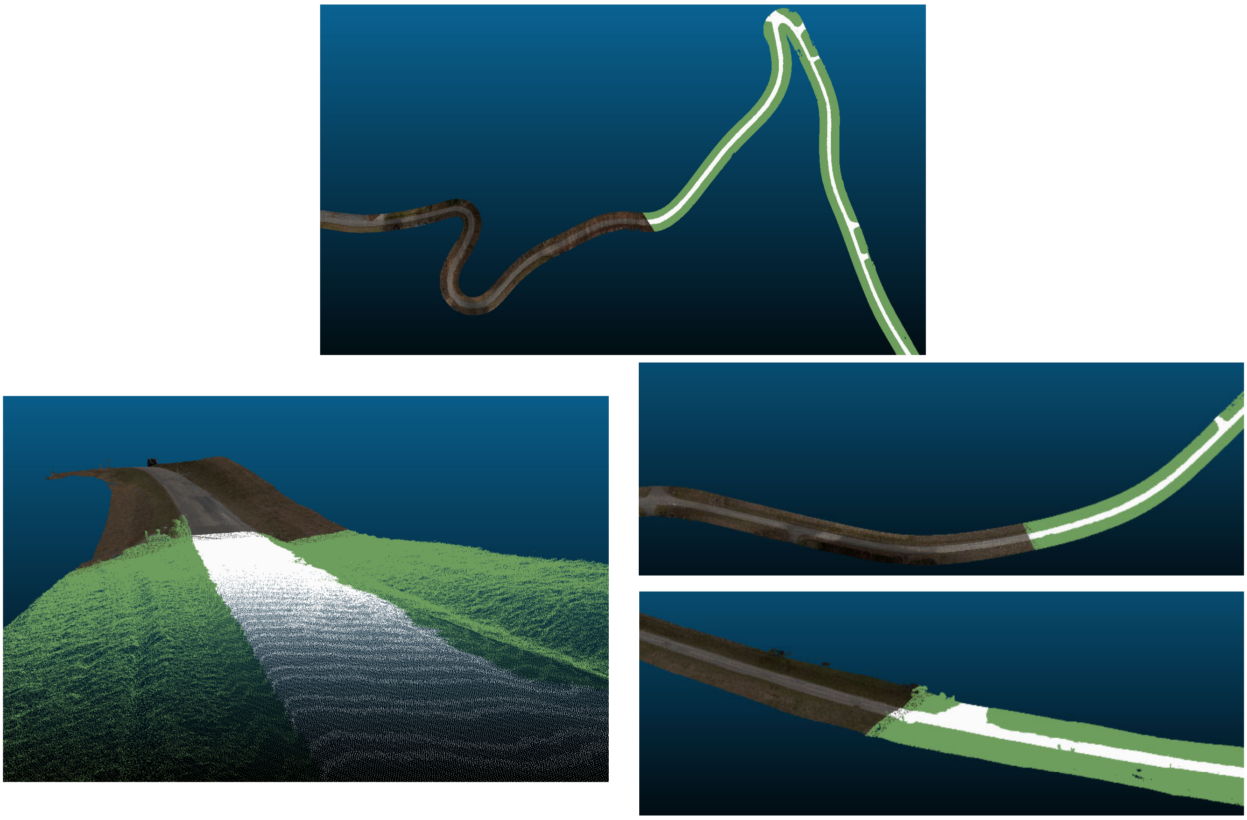

Figure 2: Sample images of the semantic segmentation of 3D point clouds, which assigns a label of "asphalt" or "background" to individual points.

Training and evaluation

Using a single GPU, the training time for the classifier in this workflow is approximately 2-3 hours with the mentioned dataset as input. This allows for efficient training of a range of models and optimizing of internal hyper-parameters of the network. To evaluate the quality of the segmentation results, the industry-standard mean intersection-over-union (mIoU) metric is used. In the IoU metric, the area of agreement between predictions and ground-truth is measured and false positives and false negatives are equally accounted for as misclassifications. The mean IoU over all relevant classes is then computed as mIoU.

Results

Thanks to the rapid turnover times, an evaluation of network architectures and hyper-parameters was performed. In these experiments, it was confirmed that modern transformer based network architectures show improved performance over sparse convolutional network architectures at comparable computational cost.

The final model achieves a mIoU score of 92.44% on the evaluation set. An in depth evaluation reveals that the majority of the remaining errors occur at the edges of the road. The lower left of Figure 2, suggests that the model, following the details of the ragged edges of the asphalt, is even superior to human polyline annotations. In any case, the manual effort for annotating the selected 2.2 km of road was about 8 hours, whereas the automatic segmentation workflow finishes in less than 30 minutes.

Conclusions

Data interpretation workflows for 3D point clouds, such as the JOANNEUM RESEARCH workflow used in the two case studies above, are able to produce near-perfect results, when carefully trained and updated via domain adaptation. The established workflow thereby greatly increases the productivity of an otherwise manual or semi-automated digitization procedure. The benefits over manual workflows were described in detail, and significant reductions of manual labour could be achieved.

Both case studies underline that the technologies for semantic interpretation of 3D point clouds are mature and ready for industrial use. With the experience gathered in the development of the workflow, transferring these methods to novel scenarios is an efficient process with only little manual effort.

The learning-based algorithms require an initial annotation phase, but it is precisely this step that ensures that very specific interpretation algorithms can be created by providing suitable data. Only little data is required for reaching high quality prototypes, and with more data becoming available, these models can subsequently be refined and optimized for specific needs. Processing times for training and inference show that modest computational resources are needed. The presented methods already scale well and will further benefit when parallelized.

DI Dr. Olaf Kähler is a key researcher at the JOANNEUM RESEARCH DIGITAL institute, where he works in the *Intelligent Vision Applications* research group. The DIGITAL institute is dedicated to developing innovative digital technologies that work reliably even under demanding conditions. The aim is to use modern sensor systems to optimise processes, use resources efficiently, increase safety and address social challenges.

DI Dr. Kähler's research focuses on machine vision technologies in combination with artificial intelligence methods. His scientific interests include 3D computer vision, machine learning and the analysis of complex scenes.