3D INDOOR MAPPING WITH THE MICROSOFT HOLOLENS: QUALITATIVE AND QUANTITATIVE EVALUATION BY MEANS OF GEOMETRIC FEATURES

Martin Weinmann[1,*], Miriam Amelie Jäger[1], Sven Wursthorn[1], Boris Jutzi[1], Michael Weinmann[2], Patrick Hübner[1]

[1] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany - (martin.weinmann, sven.wursthorn, boris.jutzi, patrick.huebner)@kit.edu, miriam.jaeger@student.kit.edu
[2] Institute of Computer Science II, University of Bonn, Bonn, Germany - mw@cs.uni-bonn.de

Commission I, ICWG I/IV

KEY WORDS: 3D, Indoor Mapping, HoloLens, Feature Extraction, Classification, Semantic Segmentation, Evaluation

ABSTRACT:

3D indoor mapping and scene understanding have seen tremendous progress in recent years due to the rapid development of sensor systems, reconstruction techniques and semantic segmentation approaches. However, the quality of the acquired data strongly influences the accuracy of both reconstruction and segmentation. In this paper, we direct our attention to the evaluation of the mapping capabilities of the Microsoft HoloLens in comparison to high-quality TLS systems with respect to 3D indoor mapping, feature extraction and semantic segmentation. We demonstrate how a set of rather interpretable low-level geometric features and the resulting semantic segmentation achieved with a Random Forest classifier applied on these features are affected by the quality of the acquired data. The achieved results indicate that, while allowing for a fast acquisition of room geometries, the HoloLens provides data with sufficient accuracy for a wide range of applications.

1. INTRODUCTION

Rapid 3D mapping and scene understanding for indoor environments have seen tremendous progress in recent years, enabling a rich diversity of applications including scene modeling, navigation and perception assistance, and future use cases like telepresence. Besides 3D reconstruction based on RGB imagery (Remondino et al., 2017; Stathopoulou et al., 2019), RGB-D data (Zollhöfer et al., 2018) or data acquired via mobile indoor mapping systems (Lehtola et al., 2017; Chen et al., 2018; Nocerino et al., 2017; Masiero et al., 2018), there has also been an increasing interest in augmenting the acquired 3D data with virtual contents or semantics. In this regard, mobile Augmented Reality (AR) devices like the Microsoft HoloLens allow for the in-situ visualization of virtual contents (*e.g.*, Building Information Modelling (BIM) data or information directly derived from the acquired data) which, in turn, facilitates numerous applications addressing facility management, cultural heritage documentation or educational services.

While the HoloLens has recently been evaluated regarding its capabilities as an AR device (Liu et al., 2018) and regarding the spatial stability of holograms (Vassallo et al., 2017), there have also been first investigations on the spatial accuracy of triangle meshes acquired by the HoloLens in comparison to ground truth data acquired with a terrestrial laser scanning system (Khoshelham et al., 2019; Hübner et al., 2019). However, to the best of our knowledge, the impact of the quality of the acquired data on the extraction of geometric features and thus on the results of semantic segmentation (Weinmann, 2016; Poux, Billen, 2019) still remains an open issue, although it has recently been proven that the robustness of such geometric features is strongly influenced by such cues (Dittrich et al., 2017).

In this paper, we address 3D indoor mapping with the Microsoft HoloLens (Version 1) with a particular focus on a quantitative and qualitative evaluation by means of geometric features. We use a set of rather interpretable low-level geometric 3D and 2D features (Weinmann, 2016; Weinmann et al., 2017), which are extracted from the local neighborhood of each query point and concatenated to define the respective feature vector. The latter, in turn, serves as input to a classifier, for which we use a Random Forest classifier (Breiman, 2001) in the scope of our work. We compare the behavior and expressiveness of the involved features to their counterparts extracted from downsampled TLS data, and we analyze the impact of different feature sets on the classification results.

This paper is organized as follows. We first briefly discuss related work with respect to sensor systems and recent progress in indoor mapping in Section 2. Subsequently, we explain the applied methodology in Section 3. In Section 4, we present and compare the results achieved for an indoor environment that has been acquired with the Microsoft HoloLens and a TLS system in two independent scan campaigns. These results are discussed in Section 5. Finally, a summary and concluding remarks as well as suggestions for future work are provided in Section 6.

2. RELATED WORK

In the following, we briefly summarize related work with respect to sensor systems (Section 2.1) and recent progress in indoor mapping (Section 2.2).

2.1 Sensor Systems

For highly accurate geometry acquisition within indoor environments, Terrestrial Laser Scanning (TLS) systems are typically used. While the quality of a range measurement generally depends on a variety of influencing factors (Soudarissanane et al., 2011; Weinmann, 2016), remaining errors often tend to be negligible and are mainly caused by either (i) the characteristics of the observed scene in terms of object materials,

* Corresponding author

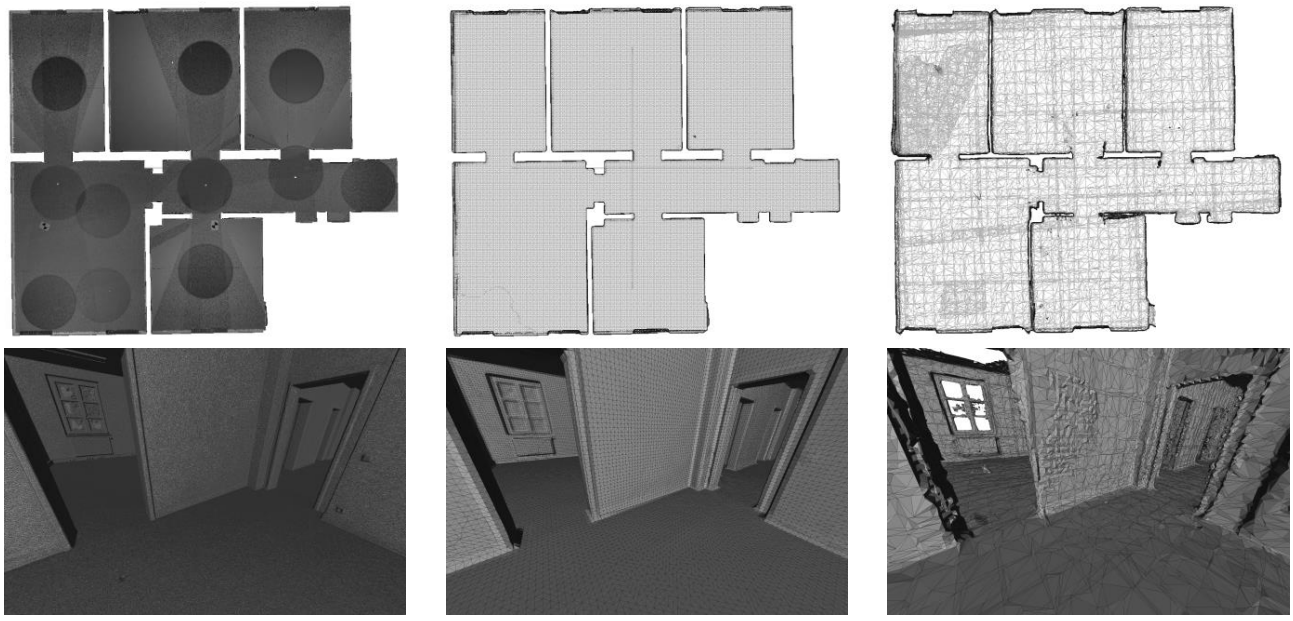

Figure 1. The considered scene representing an empty apartment in nadir view (top row) and in detailed oblique view (bottom row): raw TLS data (left), TLS data downsampled via a voxel-grid filter using a voxel size of 3 cm (center), and HoloLens data (right).

surface reflectivity, surface roughness, etc. or (ii) the scanning geometry (*i.e.*, the relative distance and orientation of object surfaces with respect to the used scanning device). However, a TLS system is rather expensive, and a single scan may typically not be sufficient to achieve a full coverage of the considered indoor scene. Hence, several scans have to be acquired and transformed into a common coordinate system, as indicated in Figure 1. In this context, different side constraints such as range constraints and/or incidence angle constraints may be taken into account (Soudarissanane, Lindenbergh, 2011) and, if done manually using artificial markers, this process may be laborious and time-consuming.

To facilitate indoor scene acquisition, a diversity of Mobile Laser Scanning (MLS) systems or Mobile Mapping Systems (MMSs) have been presented with only little loss in measurement accuracy. In this regard, commonly used systems are represented by trolley-based systems (*e.g.*, the NavVis mobile mapping system[1]), UAV-based systems (Hillemann et al., 2019), backpack-based systems (Nüchter et al., 2015; Blaser et al., 2018) or hand-held systems (*e.g.*, the Leica BLK2GO[2]). However, such systems tend to be rather expensive due to the involved laser scanning device(s) and/or the involved multi-camera system. Furthermore, trolley-based systems encounter challenges in stairways, while UAV-based systems require an expert to fly the sensor platform and backpack-based systems have a significant weight. Thus, applicability for the end-user is typically reduced.

To address the required expenses, low-cost RGB-D cameras (*e.g.*, the Microsoft Kinect or the Intel RealSense) have been presented which can be used as a hand-held device for scene acquisition. Such RGB-D cameras allow for scene capture with high frame rates and are therefore often suitable for acquiring both static and dynamic scenes. Among a diversity of approaches, KinectFusion (Izadi et al., 2011) and respective improvements (Nießner et al., 2013; Kähler et al., 2016; Dai et al., 2017b; Stotko et al., 2019) have become popular methods for fast scene reconstruction. For a detailed survey on 3D scene acquisition with RGB-D cameras, we refer to (Zollhöfer et al., 2018). Due to the focus on the low-cost constraint, however, such systems tend to reveal limited capabilities regarding the accuracy of geometry acquisition. In particular, errors are caused by sensor noise, limited resolution and misalignments due to drift (Zollhöfer et al., 2018).

Providing a trade-off between accurate scene acquisition and low-cost solution, a popular device is given with the Microsoft HoloLens[3] representing a mobile, head-worn AR device. The HoloLens provides the capability to map its direct environment in real-time in the form of triangle meshes and to simultaneously localize itself within the acquired meshes. The latter is achieved based on four gray-scale tracking cameras, while the 3D mapping relies on a time-of-flight (ToF) range camera operating to distances of up to about 3.5 m. Besides these mapping capabilities, the HoloLens is capable of augmenting the physical environment of the user with virtual content. This AR capability of the device could be used to guide the user by providing information about where to look in order to have a scene coverage as high as possible. This makes the Microsoft HoloLens rather easy-to-use for non-expert end-users.

2.2 Indoor Mapping

Besides geometry acquisition as possible with various sensor systems described in the previous section, indoor mapping may also address further tasks, such as the acquisition of the given room topology as well as Building Information Modeling (BIM) (Tran et al., 2017; Nikoohemat et al., 2019; Ochmann et al., 2019), and semantic segmentation (Armeni et al., 2016; Engelmann et al., 2017; Poux et al., 2018; Poux, Billen, 2019). The latter can be done on point-level (*i.e.*, each point is assigned a class label indicating one of the defined object categories) and on instance-level (*i.e.*, each point is assigned a semantic class label indicating one of the defined object categories and an instance label indicating the respective object in the scene), and by using traditional approaches relying on the use of hand-crafted features or by using modern deep learning techniques.

[1] https://www.navvis.com/m6
[2] https://blk2go.com
[3] https://www.microsoft.com/en-us/hololens

To foster research on indoor scene reconstruction and understanding, a variety of datasets have been presented:

- The dataset released with the ISPRS Benchmark on Indoor Modelling (Khoshelham et al., 2017) contains five indoor scenes, each captured with a different sensor.
- The Stanford 2D-3D-Semantics Dataset (Armeni et al., 2017) contains six large-scale indoor areas with both semantic and geometric annotations.
- The ScanNet dataset (Dai et al., 2017a) represents a large-scale RGB-D video dataset containing more than 1.5k indoor scenes annotated with respect to camera poses, surface reconstructions, and semantic segmentation.
- The Matterport3D dataset (Chang et al., 2017) is a large-scale RGB-D dataset containing 90 indoor scenes and more than 2000 rooms annotated with respect to surface reconstruction, camera poses, and both 2D and 3D semantic segmentations suitable for several scene understanding tasks.
- The House3D dataset (Wu et al., 2018) contains more than 45k human-designed, visually realistic 3D indoor scenes characterized by a diversity of 3D objects, textures and scene layouts.
- The Replica Dataset (Straub et al., 2019) contains high-quality reconstructions of a variety of indoor scenes, whereby the focus was set on obtaining visually, geometrically, and semantically realistic models of the world.

While all these datasets have been created with a focus on the mapping and/or modeling of large-scale indoor scenes, none of them contains data for the same scene, but acquired with different sensor systems.

3. METHODOLOGY

For evaluating the influence of the quality of the acquired data on the expressiveness of geometric features and on the accuracy of semantic segmentation, we focus on a traditional workflow. To describe each 3D point via geometric features, characteristics of the spatial arrangement of neighboring points have to be encoded appropriately. Accordingly, we first need to recover the local neighborhood for each point of the point cloud (Section 3.1) in order to encode the local 3D structure via geometric features (Section 3.2). The derived encoding then serves as input for classification (Section 3.3).

3.1 Recovery of Local Neighborhoods

To recover the local neighborhood for each point $\mathbf{X}_i$ of the point cloud, we focus on local neighborhoods with a locally-adaptive neighborhood size (Weinmann, 2016). Instead of relying on an identical scale parameter (represented by the number k of nearest neighbors) that needs to be determined once for the considered dataset, this definition relies on the idea that the selection of an optimal neighborhood size parameterized by $k_i = k_{i,\text{opt}}$ might depend on the local 3D structure and thus, to some degree, the considered classification task. To achieve such a local adaptation, we focus on eigenentropy-based scale selection (Weinmann, 2016) that has proven beneficial compared to dimensionality-based scale selection (Demantké et al., 2011).

The main idea of eigenentropy-based scale selection (Weinmann, 2016) consists in the consideration of different values of the scale parameter to derive multiple neighborhoods for each 3D point and selecting the value of the scale parameter that corresponds to the minimal disorder of 3D points across the considered local neighborhoods. More specifically, for different values of the scale parameter (here: k), the 3D coordinates of a query point $\mathbf{X}_i = \mathbf{X}_{i,0}$ and its k nearest neighbors $\mathbf{X}_{i,j}$ with $j = 1, \ldots, k$ are used to calculate the 3D structure tensor

$$\mathbf{S}_i = \frac{1}{k+1} \sum_{j=0}^{k} \left(\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i \right) \left(\mathbf{X}_{i,j} - \bar{\mathbf{X}}_i \right)^T \tag{1}$$

representing a 3D covariance matrix for the local barycenter

$$\bar{\mathbf{X}}_i = \frac{1}{k+1} \sum_{j=0}^{k} \mathbf{X}_{i,j}. \tag{2}$$

Thus, the three eigenvalues of $\mathbf{S}_i$ exist, are non-negative and indicate the dispersion magnitude along their corresponding eigenvectors (Dittrich et al., 2017). Normalizing these eigenvalues by their sum yields normalized eigenvectors $\lambda_{i,1}$, $\lambda_{i,2}$ and $\lambda_{i,3}$. Without loss of generality, we assume that $\lambda_{i,1} \geq \lambda_{i,2} \geq \lambda_{i,3} \geq 0$, and that these normalized eigenvalues can be expressed as a function of the neighborhood size: $\lambda_{i,j} = f(k)$. The optimal neighborhood size $k_{i,\text{opt}}$ minimizes the eigenentropy E_i (*i.e.*, a measure for the disorder of neighboring 3D points) according to

$$k_{i,\text{opt}} = \underset{k \in \mathcal{K}}{\arg\min}\, E_i = \underset{k \in \mathcal{K}}{\arg\min} \left(-\sum_{j=1}^{3} \lambda_{i,j}(k) \ln \lambda_{i,j}(k) \right). \tag{3}$$

In accordance with related work (Demantké et al., 2011; Weinmann, 2016), we consider scale parameters within the interval $\mathcal{K} = [k_{\text{min}}, k_{\text{max}}]$, whereby we consider relevant statistics to start with a minimum number of $k_{\text{min}} = 10$ neighboring points and, in order to limit the computational burden, we select the upper boundary as $k_{\text{max}} = 100$.

3.2 Extraction of Geometric Features

To encode characteristics of the spatial arrangement of points within the local neighborhood of a query point $\mathbf{X}_i$, we consider the corresponding 3D structure tensor $\mathbf{S}_i$ and its normalized eigenvectors $\lambda_{i,1}$, $\lambda_{i,2}$ and $\lambda_{i,3}$. The latter, in turn, are used to derive the dimensionality features of linearity L_i, planarity P_i and sphericity S_i as well as further eigenvalue-based features represented by omnivariance O_i, anisotropy A_i, eigenentropy E_i and change of curvature C_i (West et al., 2004; Pauly et al., 2003):

$$L_i = \frac{\lambda_{i,1} - \lambda_{i,2}}{\lambda_{i,1}} \tag{4}$$

$$P_i = \frac{\lambda_{i,2} - \lambda_{i,3}}{\lambda_{i,1}} \tag{5}$$

$$S_i = \frac{\lambda_{i,3}}{\lambda_{i,1}} \tag{6}$$

$$O_i = \sqrt[3]{\prod_{j=1}^{3} \lambda_{i,j}} \tag{7}$$

$$A_i = \frac{\lambda_{i,1} - \lambda_{i,3}}{\lambda_{i,1}} \tag{8}$$

$$E_i = -\sum_{j=1}^{3} \lambda_{i,j} \ln \lambda_{i,j} \tag{9}$$

$$C_i = \frac{\lambda_{i,3}}{\sum_{j=1}^{3} \lambda_{i,j}} \quad (10)$$

Furthermore, we follow (Weinmann, 2016) and consider geometric features represented by the absolute height H_i of the query point $\mathbf{X}_i$, the radius R_i of the local neighborhood, the local point density ρ_i with

$$\rho_i = \frac{k+1}{\frac{4}{3}\pi R_i^3}, \quad (11)$$

the verticality V_i with

$$V_i = 1 - |n_z| \quad (12)$$

where n_z is the vertical component of the local normal vector, and the maximum difference ΔH_i as well as the standard deviation $\sigma_{H,i}$ of the height values of all points within the local neighborhood.

Finally, we take into account that indoor environments reveal many vertical structures. Accordingly, we apply a 2D projection of the point $\mathbf{X}_i$ and its k_i nearest neighbors onto a horizontal plane (Weinmann, 2016). Based on the 2D coordinates of these projections, we derive the 2D structure tensor in analogy to the 3D structure tensor. The 2D structure tensor, in turn, has two eigenvalues ξ_1 and ξ_2 with $\xi_1 \geq \xi_2 \geq 0$ from which we derive their sum $\Sigma_{\xi,i}$ as well as their ratio $R_{\xi,i}$. Besides these eigenvalue-based 2D features, we also consider the radius r_i and the point density ζ_i on the basis of the 2D projections.

For our framework, we consider different feature sets as input for a subsequent classification. These feature sets comprise the set $\mathcal{S}_{\text{3D,EV}}$ including all eigenvalue-based 3D features, the set $\mathcal{S}_{\text{3D,other}}$ including all other geometric 3D features, the set $\mathcal{S}_{\text{3D,all}}$ including all 3D features, the set $\mathcal{S}_{\text{2D,all}}$ including all 2D features, and the set $\mathcal{S}_{\text{all}}$ including all 3D and 2D features:

$$\mathcal{S}_{\text{3D,EV}} = \{L_i, P_i, S_i, O_i, A_i, E_i, C_i\} \quad (13)$$

$$\mathcal{S}_{\text{3D,other}} = \{H_i, R_i, \rho_i, V_i, \Delta H_i, \sigma_{H,i}\} \quad (14)$$

$$\mathcal{S}_{\text{3D,all}} = \{L_i, P_i, S_i, O_i, A_i, E_i, C_i, H_i, R_i, \rho_i, V_i, \Delta H_i, \sigma_{H,i}\} \quad (15)$$

$$\mathcal{S}_{\text{2D,all}} = \{\Sigma_{\xi,i}, R_{\xi,i}, r_i, \zeta_i\} \quad (16)$$

$$\mathcal{S}_{\text{all}} = \mathcal{S}_{\text{3D,all}} \cup \mathcal{S}_{\text{2D,all}} \quad (17)$$

3.3 Supervised Classification

To assign an appropriate class label to a query point, in the scope of this paper, we focus on only considering the corresponding feature vector resulting from the concatenation of all extracted features, while we consider both smooth labeling techniques (Schindler, 2012) and structured regularization techniques (Landrieu et al., 2017) as subject of future work.

Given a set of representative training data, we focus on supervised classification based on a Random Forest classifier (Breiman, 2001). This classifier is a representative of discriminative classification approaches searching for the best separation of data points, independent of underlying probability density functions. More specifically, a Random Forest classifier is generated via a strategic combination of a set of weak learners represented by decision trees. These decision trees are trained on different, randomly chosen subsets of the given training data (Breiman, 1996). When training a single decision tree, the focus is set on a successive splitting of the data into smaller subsets based on specific homogeneity criteria until the resulting subset at a leaf node is as pure as possible. Since all decision trees are trained on independent, randomly different subsets of the given training data, their hypotheses for new unseen data to be classified can be considered as de-correlated. Thus, taking the majority vote across all these hypotheses represents a reasonable class prediction with improved generalization and robustness (Criminisi, Shotton, 2013).

To select the internal settings of the Random Forest (*e.g.*, the number of involved decision trees), we perform a grid search on a suitable raster during the training process. Given the hypotheses of the involved decision trees also allows interpreting the output of the Random Forest as a soft assignment indicating the probabilities with which a query point $\mathbf{X}_i$ is associated to each of the given classes. Such a soft assignment thus represents a measure of confidence with respect to the assigned class label.

4. EXPERIMENTAL RESULTS

In the following, we first focus on the two involved sensor systems represented by the Microsoft HoloLens and a Leica HDS6000 (Section 4.1), and we then describe the acquired datasets (Section 4.2). Subsequently, we demonstrate the impact of the quality of the acquired data on the extraction of geometric features (Section 4.3). Finally, we present the classification results achieved when using different feature sets as input for classification (Section 4.4).

4.1 Microsoft HoloLens vs. Leica HDS6000

The Microsoft HoloLens is equipped with a variety of sensors. Among these sensors, a video camera is used to allow recording screenshot videos and pictures, in which the physical environment can be augmented with virtual contents. In addition, there are four gray-scale tracking cameras for a robust self-localization. Two of these are oriented to the front in a stereo configuration with large overlap, while the other two are oriented to the right and left with nearly no overlap to the center pair. Furthermore, the HoloLens contains a time-of-flight (ToF) depth sensing camera providing images with pixel-wise range measurements, whereby range images can be queried in two different modes for the range from $0\,\text{m}$ to $0.8\,\text{m}$ ("short throw" mode) and the range from $0.8\,\text{m}$ to about $3.5\,\text{m}$ ("long throw" mode). The respective field-of-view of these sensors is illustrated in Figure 2. More detailed specifications can be accessed via the Microsoft Windows 10 SDK for the device.

The Leica HDS6000 is a standard phase-based terrestrial laser scanner with survey-grade accuracy (within a few mm range) and a field-of-view of $360^\circ \times 155^\circ$. To obtain complete scene coverage, several scans have to be taken from different positions and, as the data acquired with each scan refers to the local coordinate system of the scanner, all acquired scans have to be transferred into a common reference coordinate system. This process is referred to as point cloud registration.

4.2 Datasets

The considered scene is represented by an empty apartment consisting of five rooms of different size and one central hallway as shown in Figure 1.

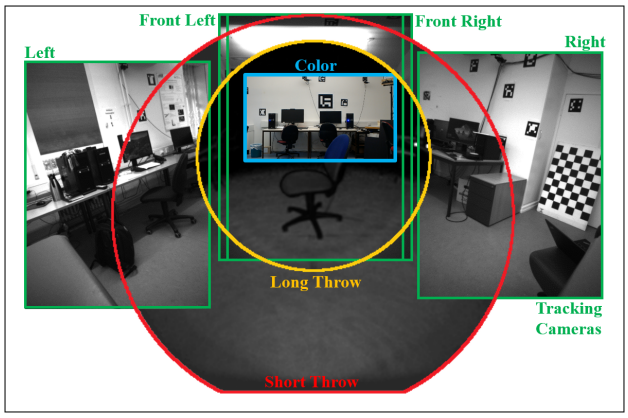

Figure 2. Overlay of the images recorded by the different sensors of the Microsoft HoloLens (Hübner et al., 2020).

For HoloLens-based scene acquisition, an operator wearing the device went through the apartment and achieved a rapid and comfortable mapping of the indoor scene within a few minutes. To create the triangle mesh, the commercially available SpaceCatcher HoloLens App[4] was used, since this allowed directly visualizing the triangle meshes for the operator while they were recorded. The resulting mesh is visualized in the right part of Figure 1 and contains 105,200 points.

For TLS-based scene acquisition, 11 scans were taken from the positions indicated with a circle in the left part of Figure 1 and registered by using artificial planar and spherical markers placed in the apartment to establish correspondence. Subsequently, the complete point cloud was manually cleaned, downsampled via a voxel-grid filter using a voxel size of 3 cm, and finally meshed via Poisson Surface Reconstruction (Kazhdan et al., 2006). The resulting mesh is visualized in the center part of Figure 1 and contains 178,322 points.

4.3 Feature Extraction Results

We use eigenentropy-based scale selection to derive locally-adaptive neighborhoods (*i.e.*, local neighborhoods whose size is optimized for each query point individually; see Section 3.1). Based on these neighborhoods, geometric features are extracted (see Section 3.2). The behavior of the neighborhood size and the different features across the complete mesh is visualized in Figures 3 and 4 for the HoloLens dataset and the TLS dataset, respectively.

4.4 Classification Results

For classification, we focus on a rather simple scenario with the three classes "Ceiling", "Floor" and "Wall" in the scope of this work. The ground truth labeling obtained via manual annotation is visualized in Figures 3 and 4 for the HoloLens dataset and the downsampled TLS dataset, respectively.

For training, we take into account that an imbalanced amount of training examples across different classes may have a detrimental effect on the generalization capability of the classifier. Hence, we randomly select 1000 points per class for training and all remaining points for performance evaluation. The latter is carried out based on commonly used evaluation metrics: Overall Accuracy (OA), κ-Index and class-wise F_1-scores.

The classification results achieved when using different feature sets as input for the classifier are provided in Tables 1 and 2 for the HoloLens dataset and the downsampled TLS dataset, respectively. Visualizations corresponding to these results are provided in Figure 5.

Feature Set	OA	κ	F_1(C)	F_1(F)	F_1(W)
$\mathcal{S}_{\text{3D,EV}}$	59.36	28.23	43.50	40.53	71.03
$\mathcal{S}_{\text{3D,other}}$	93.28	87.36	90.84	91.57	94.62
$\mathcal{S}_{\text{3D,all}}$	93.26	87.25	90.51	91.94	94.63
$\mathcal{S}_{\text{2D,all}}$	73.08	52.65	60.60	51.56	84.12
$\mathcal{S}_{\text{all}}$	93.36	87.46	90.69	92.02	94.70

Table 1. Results (in %) achieved for the classification of the HoloLens dataset when using different feature sets as input for the classifier: Overall Accuracy (OA), κ-Index and class-wise F_1-scores (C: Ceiling; F: Floor; W: Wall).

Feature Set	OA	κ	F_1(C)	F_1(F)	F_1(W)
$\mathcal{S}_{\text{3D,EV}}$	57.30	30.36	37.42	60.71	65.65
$\mathcal{S}_{\text{3D,other}}$	98.60	97.47	98.42	97.99	98.86
$\mathcal{S}_{\text{3D,all}}$	98.44	97.17	97.88	98.11	98.72
$\mathcal{S}_{\text{2D,all}}$	82.12	67.56	63.40	50.61	97.42
$\mathcal{S}_{\text{all}}$	98.10	96.57	97.58	97.57	98.44

Table 2. Results (in %) achieved for the classification of the downsampled TLS dataset when using different feature sets as input for the classifier: Overall Accuracy (OA), κ-Index and class-wise F_1-scores (C: Ceiling; F: Floor; W: Wall).

5. DISCUSSION

The provided visualizations reveal that the accuracy of the acquired HoloLens dataset is worse compared to the accuracy of the downsampled TLS dataset. However, the HoloLens allows for a fast acquisition of the room geometry, and the accuracy is still sufficient as initialization to apply voxel representations or plane fitting techniques for creating a 3D model of the indoor scene. The accuracy might also still be sufficient to have a fast guess about the area and volume of the apartment, two cues important when calculating the rent for or the costs of the apartment.

The visualizations in Figures 3 and 4 indicate the flexibility of the neighborhood size varying between 10 and 100 nearest neighbors for the query points. Furthermore, they allow for reasoning about expressive features (*e.g.*, the height H_i, the verticality V_i, or the ratio $R_{\xi,i}$ of the eigenvalues of the 2D structure tensor) and less-expressive features (*e.g.*, the radii R_i and r_i of the local neighborhood in 3D and 2D, the local point densities ρ_i and ζ_i in 3D and 2D, or the sum $\Sigma_{\xi,i}$ of the eigenvalues of the 2D structure tensor) with respect to the considered classification task. Of course, some features might be less suitable if there are more classes with a higher similarity or more complex indoor scenes (*e.g.*, scenes covering different floors and/or also containing room inventory).

Among the feature sets, the set $\mathcal{S}_{\text{3D,EV}}$ including all eigenvalue-based 3D features is not suitable to achieve appropriate classification results (Figures 3 and 4 and Tables 1 and 2). The reason for this is that the respective features describe local characteristics around the query point with respect to the principal axes of the 3D ellipsoid spanned by the neighboring points, while the absolute orientation with respect to horizontal and vertical directions is not taken into account. Also the set $\mathcal{S}_{\text{2D,all}}$ including all 2D features does not allow appropriately separating the defined classes, since the 2D features r_i, ζ_i and $\Sigma_{\xi,i}$ are less expressive and only $R_{\xi,i}$ is expressive allowing to separate vertical from horizontal planes, which is not sufficient for separating the defined classes. The other feature sets $\mathcal{S}_{\text{3D,other}}$, $\mathcal{S}_{\text{3D,all}}$

[4] `http://spacecatcher.madeinholo.com`

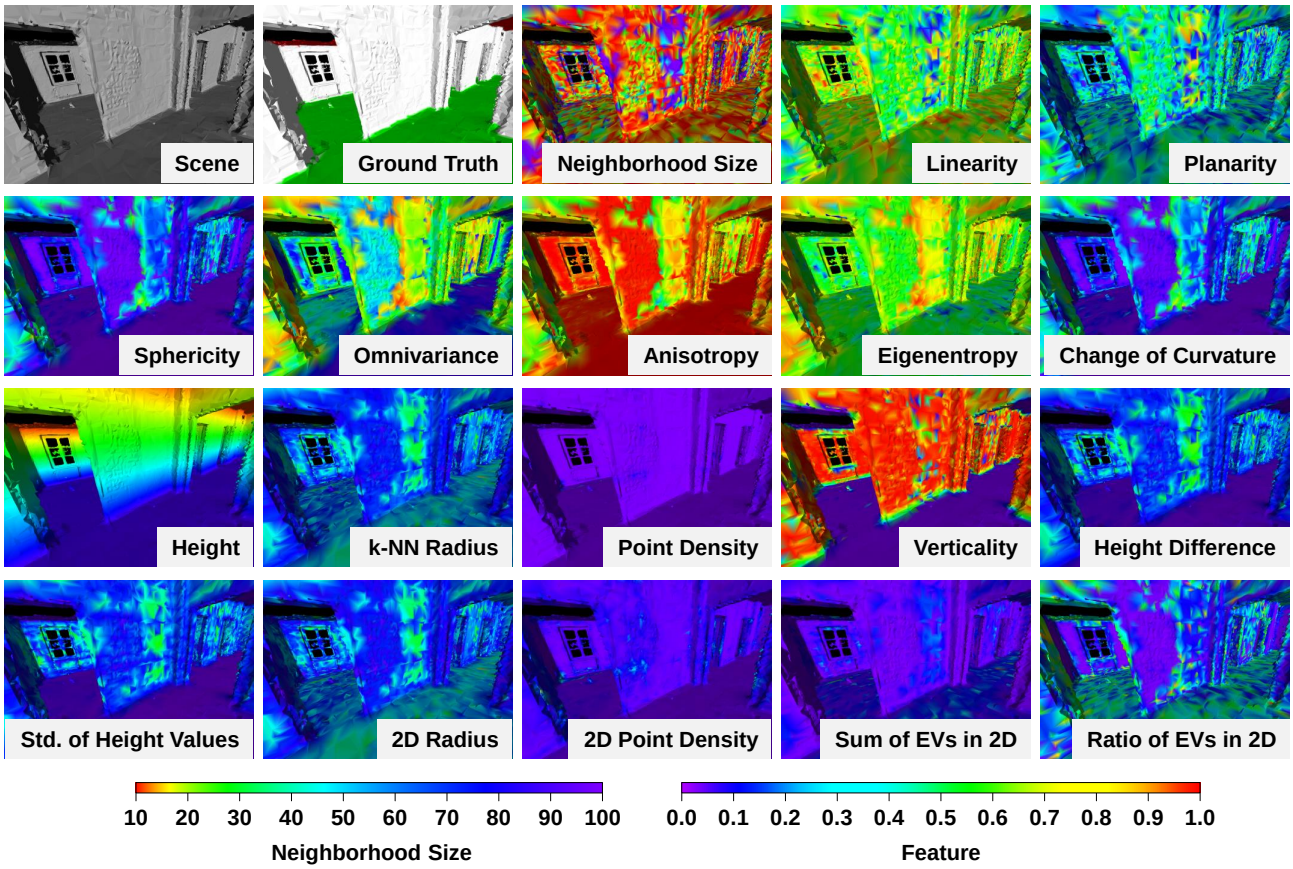

Figure 3. Visualization of the scene, the ground truth labeling, the neighborhood size, and the 17 considered features (scaled to $[0, 1]$) for the HoloLens dataset.

and $\mathcal{S}_{\text{all}}$ lead to classification results of almost the same quality, since they have several features in common that are highly relevant for the considered classification task (*e.g.*, the height H_i, the verticality V_i, and the maximum difference ΔH_i and standard deviation $\sigma_{H,i}$ of height values).

A comparison of the classification results achieved for the HoloLens dataset and for the downsampled TLS dataset (Tables 1 and 2 and Figure 5) reveals a decrease in OA when using the HoloLens for data acquisition. This decrease in OA is about 5 %, when considering the more suitable feature sets $\mathcal{S}_{\text{3D,other}}$, $\mathcal{S}_{\text{3D,all}}$ and $\mathcal{S}_{\text{all}}$.

6. CONCLUSIONS

In this paper, we have focused on rapid 3D mapping and scene understanding for indoor environments. In particular, we have addressed 3D indoor mapping with the Microsoft HoloLens with a particular focus on a quantitative and qualitative evaluation by means of geometric features. Considering an indoor scene acquired with either a HoloLens or a TLS system (Leica HDS6000), we have extracted a set of rather interpretable low-level geometric 3D and 2D features and provided these features as input for a Random Forest classifier. We have analyzed the impact of the quality of the acquired point cloud data on the behavior and expressiveness of the interpretable geometric features and on the classification with respect to three classes ("Ceiling", "Floor" and "Wall"). Furthermore, we have evaluated the impact of different feature sets on the classification results.

In future work, we plan to increase the number of considered classes and the complexity of the considered scene (*e.g.*, by considering indoor scenes covering different floors and also containing room inventory). Furthermore, we aim at guiding the user during the acquisition regarding scene completion and densification of sparsely reconstructed areas.

REFERENCES

Armeni, I., Sax, S., Zamir, A. R., Savarese, S., 2017. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105v2*.

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S., 2016. 3d semantic parsing of large-scale indoor spaces. *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543.

Blaser, S., Cavegn, S., Nebiker, S., 2018. Development of a portable high performance mobile mapping system using the robot operating system. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, IV-1, 13–20.

Breiman, L., 1996. Bagging predictors. *Mach. Learn.*, 24(2), 123–140.

Breiman, L., 2001. Random forests. *Mach. Learn.*, 45(1), 5–32.

Chang, A., Dai, A., Funkhouser, T., Halber, M., Nießner, M., Savva, M., Song, S., Zeng, A., Zhang, Y., 2017. Matterport3D: learning from rgb-d data in indoor environments. *Proc. 2017 International Conference on 3D Vision*, 667–676.

Chen, Y., Tang, J., Jiang, C., Zhu, L., Lehtomäki, M., Kaartinen, H., Kaijaluoto, R., Wang, Y., Hyyppä, J., Hyyppä, H., Zhou, H., Pei, L., Chen, R., 2018. The accuracy comparison of three simultaneous localization and mapping (SLAM)-based indoor mapping technologies. *Sensors*, 18(10), 3228:1–3228:25.

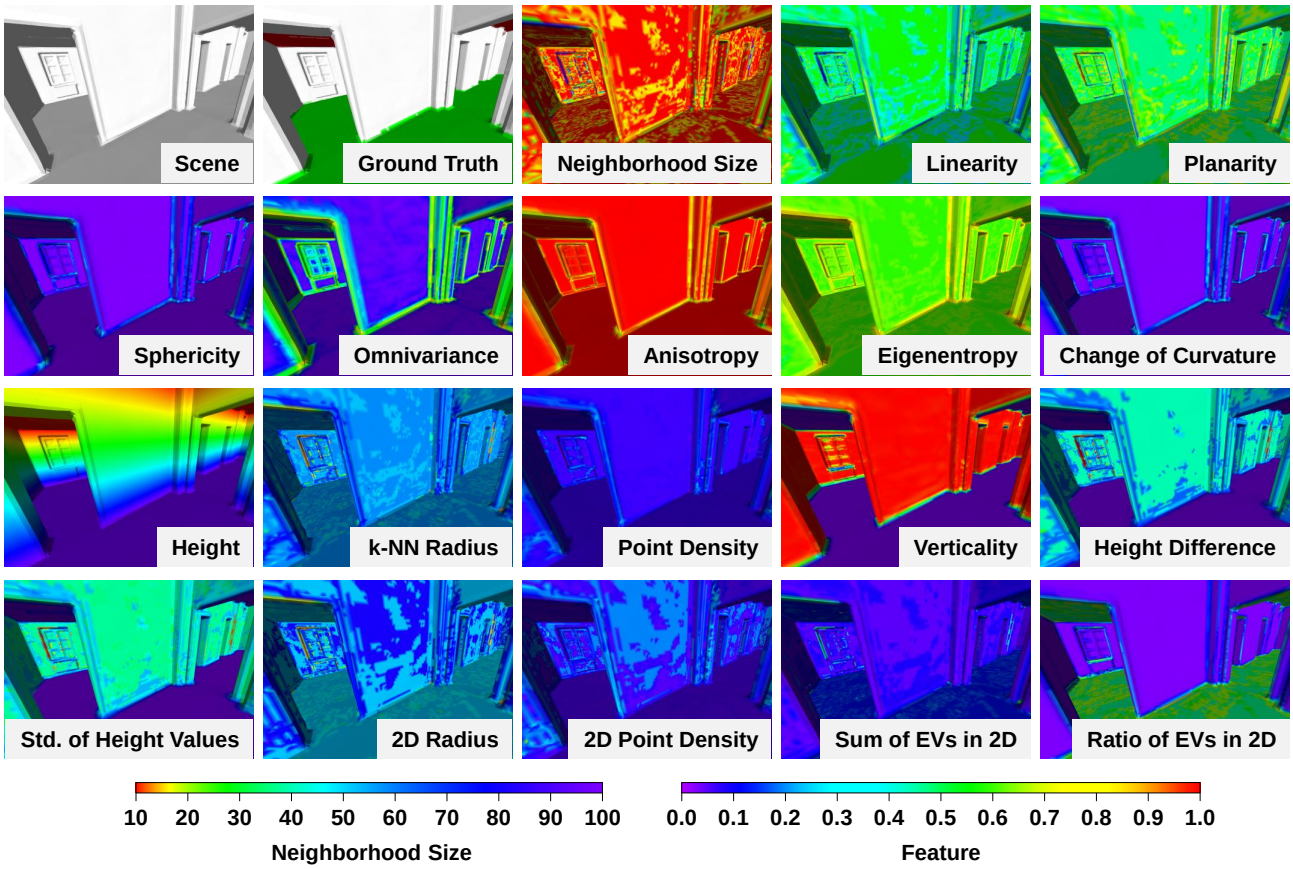

Figure 4. Visualization of the scene, the ground truth labeling, the neighborhood size, and the 17 considered features (scaled to $[0, 1]$) for the TLS dataset.

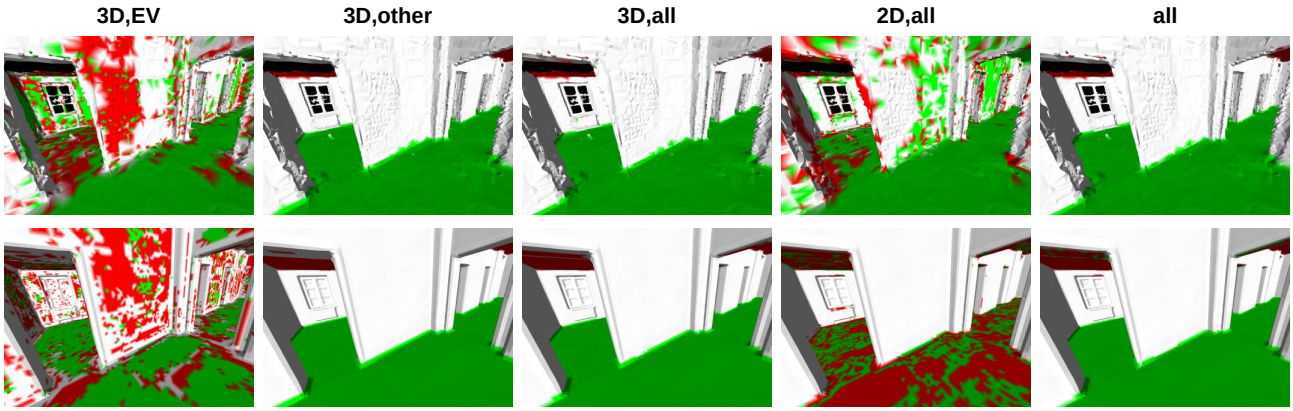

Figure 5. Classification results achieved using the set $\mathcal{S}_{\text{3D,EV}}$ including all eigenvalue-based 3D features, the set $\mathcal{S}_{\text{3D,other}}$ including all other geometric 3D features, the set $\mathcal{S}_{\text{3D,all}}$ including all 3D features, the set $\mathcal{S}_{\text{2D,all}}$ including all 2D features, and the set $\mathcal{S}_{\text{all}}$ including all 3D and 2D features extracted from the HoloLens dataset (top row) and from the downsampled TLS dataset (bottom row): the class "Ceiling" is indicated in red, the class "Floor" in green and the class "Wall" in white.

Criminisi, A., Shotton, J., 2013. *Decision forests for computer vision and medical image analysis*. Advances in Computer Vision and Pattern Recognition, Springer, London, UK.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017a. ScanNet: richly-annotated 3d reconstructions of indoor scenes. *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2432–2443.

Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C., 2017b. BundleFusion: real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(3), 24:1–24:18.

Demantké, J., Mallet, C., David, N., Vallet, B., 2011. Dimensionality based scale selection in 3d lidar point clouds. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XXXVIII-5/W12, 97–102.

Dittrich, A., Weinmann, M., Hinz, S., 2017. Analytical and numerical investigations on the accuracy and robustness of geometric features extracted from 3d point cloud data. *ISPRS J. Photogramm. Remote Sens.*, 126, 195–208.

Engelmann, F., Kontogianni, T., Hermans, A., Leibe, B., 2017. Exploring spatial context for 3d semantic segmentation of point clouds. *Proc. 2017 IEEE International Conference on Computer Vision Workshops*, 716–724.

Hillemann, M., Weinmann, M., Mueller, M. S., Jutzi, B., 2019. Automatic extrinsic self-calibration of mobile mapping systems

based on geometric 3d features. *Remote Sens.*, 11(16), 1955:1–1955:29.

Hübner, P., Clintworth, K., Liu, Q., Weinmann, M., Wursthorn, S., 2020. Evaluation of HoloLens tracking and depth sensing for indoor mapping applications. *Sensors*, 20(4), 1021:1–1021:24.

Hübner, P., Landgraf, S., Weinmann, M., Wursthorn, S., 2019. Evaluation of the Microsoft HoloLens for the mapping of indoor building environments. *Proc. Dreiländertagung der DGPF, der OVG und der SGPF*, 44–53.

Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., Fitzgibbon, A., 2011. KinectFusion: real-time 3d reconstruction and interaction using a moving depth camera. *Proc. 24th Annual ACM Symposium on User Interface Software and Technology*, 559–568.

Kähler, O., Prisacariu, V. A., Murray, D. W., 2016. Real-time large-scale dense 3d reconstruction with loop closure. *Proc. European Conference on Computer Vision*, 500–516.

Kazhdan, M., Bolitho, M., Hoppe, H., 2006. Poisson surface reconstruction. *Proc. Fourth Eurographics Symposium on Geometry Processing*, 61–70.

Khoshelham, K., Díaz Vilariño, L., Peter, M., Kang, Z., Acharya, D., 2017. The ISPRS benchmark on indoor modelling. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W7, 367–372.

Khoshelham, K., Tran, H., Acharya, D., 2019. Indoor mapping eyewear: geometric evaluation of spatial mapping capability of HoloLens. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W13, 805–810.

Landrieu, L., Raguet, H., Vallet, B., Mallet, C., Weinmann, M., 2017. A structured regularization framework for spatially smoothing semantic labelings of 3d point clouds. *ISPRS J. Photogramm. Remote Sens.*, 132, 102–118.

Lehtola, V. V., Kaartinen, H., Nüchter, A., Kaijaluoto, R., Kukko, A., Litkey, P., Honkavaara, E., Rosnell, T., Vaaja, M. T., Virtanen, J.-P., Kurkela, M., El Issaoui, A., Zhu, L., Jaakkola, A., Hyyppä, J., 2017. Comparison of the selected state-of-the-art 3d indoor scanning and point cloud generation methods. *Remote Sens.*, 9(8), 796:1–796:26.

Liu, Y., Dong, H., Zhang, L., Saddik, A. E., 2018. Technical evaluation of HoloLens for multimedia: a first look. *IEEE MultiMedia*, 25(4), 8–18.

Masiero, A., Fissore, F., Guarnieri, A., Pirotti, F., Visintini, D., Vettore, A., 2018. Performance evaluation of two indoor mapping systems: low-cost UWB-aided photogrammetry and backpack laser scanning. *Appl. Sci.*, 8(3), 416:1–416:19.

Nießner, M., Zollhöfer, M., Izadi, S., Stamminger, M., 2013. Real-time 3d reconstruction at scale using voxel hashing. *ACM Trans. Graph.*, 32(6), 169:1–169:11.

Nikoohemat, S., Diakité, A., Zlatanova, S., Vosselman, G., 2019. Indoor 3d modeling and flexible space subdivision from point clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, IV-2/W5, 285–292.

Nocerino, E., Menna, F., Remondino, F., Toschi, I., Rodríguez-Gonzálvez, P., 2017. Investigation of indoor and outdoor performance of two portable mobile mapping systems. *Proc. SPIE*, 10332, 125–139.

Nüchter, A., Borrmann, D., Koch, P., Kühn, M., May, S., 2015. A man-portable, IMU-free mobile mapping system. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, II-3/W5, 17–23.

Ochmann, S., Vock, R., Klein, R., 2019. Automatic reconstruction of fully volumetric 3d building models from oriented point clouds. *ISPRS J. Photogramm. Remote Sens.*, 151, 251–262.

Pauly, M., Keiser, R., Gross, M., 2003. Multi-scale feature extraction on point-sampled surfaces. *Comput. Graph. Forum*, 22(3), 81–89.

Poux, F., Billen, R., 2019. Voxel-based 3d point cloud semantic segmentation: unsupervised geometric and relationship featuring vs deep learning methods. *ISPRS Int. J. Geo-Inf.*, 8(5), 213:1–213:34.

Poux, F., Neuville, R., Nys, G.-A., Billen, R., 2018. 3d point cloud semantic modelling: integrated framework for indoor spaces and furniture. *Remote Sens.*, 10(9), 1412:1–1412:35.

Remondino, F., Nocerino, E., Toschi, I., Menna, F., 2017. A critical review of automated photogrammetric processing of large datasets. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W5, 591–599.

Schindler, K., 2012. An overview and comparison of smooth labeling methods for land-cover classification. *IEEE Trans. Geosci. Remote Sens.*, 50(11), 4534–4545.

Soudarissanane, S., Lindenbergh, R., 2011. Optimizing terrestrial laser scanning measurement set-up. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XXXVIII-5/W12, 127–132.

Soudarissanane, S., Lindenbergh, R., Menenti, M., Teunissen, P., 2011. Scanning geometry: influencing factor on the quality of terrestrial laser scanning points. *ISPRS J. Photogramm. Remote Sens.*, 66(4), 389–399.

Stathopoulou, E.-K., Welponer, M., Remondino, F., 2019. Open-source image-based 3d reconstruction pipelines: review, comparison and evaluation. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W17, 331–338.

Stotko, P., Krumpen, S., Weinmann, M., Klein, R., 2019. Efficient 3d reconstruction and streaming for group-scale multi-client live telepresence. *Proc. 2019 IEEE International Symposium on Mixed and Augmented Reality*, 19–25.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., Mur-Artal, R., Ren, C., Verma, S., Clarkson, A., Yan, M., Budge, B., Yan, Y., Pan, X., Yon, J., Zou, Y., Leon, K., Carter, N., Briales, J., Gillingham, T., Mueggler, E., Pesqueira, L., Savva, M., Batra, D., Strasdat, H. M., De Nardi, R., Goesele, M., Lovegrove, S., Newcombe, R., 2019. The Replica Dataset: a digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797v1*.

Tran, H., Khoshelham, K., Kealya, A., Díaz-Vilariño, L., 2017. Extracting topological relations between indoor spaces from point clouds. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, IV-2/W4, 401–406.

Vassallo, R., Rankin, A., Chen, E. C. S., Peters, T. M., 2017. Hologram stability evaluation for Microsoft HoloLens. *Proc. SPIE*, 10136, 295–300.

Weinmann, M., 2016. *Reconstruction and analysis of 3d scenes – From irregularly distributed 3d points to object classes*. Springer, Cham, Switzerland.

Weinmann, M., Jutzi, B., Mallet, C., Weinmann, M., 2017. Geometric features and their relevance for 3d point cloud classification. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, IV-1/W1, 157–164.

West, K. F., Webb, B. N., Lersch, J. R., Pothier, S., Triscari, J. M., Iverson, A. E., 2004. Context-driven automated target detection in 3-d data. *Proc. SPIE*, 5426, 133–143.

Wu, Y., Wu, Y., Gkioxari, G., Tian, Y., 2018. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209v2*.

Zollhöfer, M., Stotko, P., Görlitz, A., Theobalt, C., Nießner, M., Klein, R., Kolb, A., 2018. State of the art on 3d reconstruction with RGB-D cameras. *Comput. Graph. Forum*, 37(2), 625–652.