

AN EVALUATION FRAMEWORK FOR BENCHMARKING INDOOR MODELLING METHODS

K. Khoshelham¹, H. Tran¹, L. Díaz-Vilariño², M. Peter³, Z. Kang⁴, D. Acharya¹

¹ Dept. of Infrastructure Engineering, University of Melbourne, Parkville 3010 Australia - (k.khoshelham, ha.tran, debaditya.acharya)@unimelb.edu.au

² Applied Geotechnologies Group, Dept. of Natural Resources and Environmental Engineering, University of Vigo, Spain - lucia@uvigo.es

³ Dept. of Earth Observation Science, Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, The Netherlands - m.s.peter@utwente.nl

⁴ Dept. of Remote Sensing and Geo-Information Engineering, School of Land Science and Technology, China University of Geosciences, Beijing 100083, China - zzkang@cugb.edu.cn

Commission IV, WG IV/5

KEY WORDS: 3D reconstruction, Automation, Point cloud, BIM, Quality, Evaluation, Performance, Indoor navigation.

ABSTRACT:

Despite recent progress in the development of methods for automated reconstruction of indoor models, a comparative performance evaluation of these methods is not available due to the lack of publicly available benchmark datasets and a common evaluation framework. The ISPRS Benchmark on Indoor Modelling is an effort to enable comparison and benchmarking of indoor modelling methods by providing a benchmark dataset and a comprehensive evaluation framework. In this paper, we propose a framework for the evaluation of indoor modelling methods, and discuss various quality aspects of the reconstruction methods as well as the reconstructed models. We discuss the challenges in quantitative quality evaluation of indoor models through comparison with a reference model, and propose suitable measures and methods for comparing an automatically reconstructed indoor model with a reference.

1. INTRODUCTION

Up-to-date 3D models of indoor environments play an important role in a variety of applications ranging from navigation assistance and emergency response to planning structural repairs, refurbishment and retrofitting. Manual generation of 3D indoor models is a tedious, slow and expensive process. To make this process more effective and efficient, a number of methods have been developed for automated generation of indoor models from raw data such as point clouds and images (Gunduz et al., 2016; Pătrăucean et al., 2015; Tang et al., 2010).

A major issue in the adoption of indoor modelling methods in practical applications is the lack of a common evaluation framework for the comparison and benchmarking of the performance of these methods. In the literature, different methods have been evaluated using different datasets and based on different evaluation criteria. These criteria focus mainly on the quality of the resulting model. Qualitative evaluation by visual inspection has been the basis for the evaluation of indoor modelling methods in several works (Becker et al., 2015; Khoshelham and Díaz-Vilariño, 2014; Mura et al., 2016; Ochmann et al., 2016; Tran et al., 2017; Xiao and Furukawa, 2014). Quantitative measures derived from a comparison of the model with the data, e.g. a point cloud, have been used in several other works (Macher et al., 2017; Tran et al., 2018; Valero et al., 2012). Comparison with ground truth or a reference model has also been used in a few works for quantitative evaluation of automatically generated indoor models (Díaz-Vilariño et al., 2015; Oesau et al., 2014; Thomson and Boehm, 2015; Xiong et al., 2013).

This heterogeneity in the evaluation methods and criteria makes it difficult to compare and benchmark the performance of

different indoor modelling algorithms. In addition, existing evaluation methods usually focus on one quality aspect, e.g., geometric accuracy, while ignoring other important aspects such as the correctness and completeness of the reconstructed elements.

The ISPRS Benchmark on Indoor Modelling is an effort to enable comparison and benchmarking of indoor modelling methods by providing a benchmark dataset and a comprehensive evaluation framework. In this paper, we focus on the evaluation issue and propose a framework, which includes not only the quality of the reconstructed model but also the level of automation and the computational complexity of the reconstruction method. We discuss the challenges in quantitative quality evaluation of indoor models through comparison with a reference model, and propose suitable measures and methods for comparing an automatically reconstructed indoor model with a reference.

The paper proceeds with a brief introduction of the ISPRS Benchmark on Indoor Modelling in Section 2, followed by an overview of the proposed framework for the evaluation of indoor modelling methods in Section 3. Section 4 presents the method for measuring the geometric quality of an indoor model through comparison with a reference. Experiments and results are presented in Section 5. A summary and concluding remarks are given in Section 6.

2. THE ISPRS BENCHMARK ON INDOOR MODELLING

The ISPRS Benchmark on Indoor Modelling was proposed in 2017 with the aim of stimulating and promoting research on automated indoor modelling from point clouds by providing a

benchmark dataset and a framework for the evaluation and comparison of indoor modelling methods.

The project team collected five point clouds captured by different sensors in five indoor environments representing different levels of complexity. The dataset was made publicly available via the website of the ISPRS Working Group IV/5¹. From each point cloud a 3D model was created manually to serve as reference for the evaluation of automatically reconstructed models. Figure 1 shows the five point clouds and the corresponding reference models. A detailed description of the benchmark dataset, sensor specifications, and reference models, is provided in (Khoshelham et al., 2017).

The ISPRS website for the benchmark dataset was set up in September 2017. Since then, the dataset has been downloaded by 70 researchers from 16 countries. Figure 2 shows the download statistics of the benchmark dataset.

3. A FRAMEWORK FOR THE EVALUATION OF INDOOR MODELLING METHODS

We consider three main aspects to be crucially important in the evaluation of an indoor modelling method: level of automation, computational complexity, and the quality of the generated model.

Level of automation describes the amount of manual interaction and intervention that is needed from a human expert in the automated reconstruction process. It might vary depending on the complexity of the dataset, and is therefore difficult to measure quantitatively. Hence, we assess the level of automation using the following qualitative terms: interactive, semi-automated, and fully automated. An interactive method is one that requires a significant amount of interaction by a human expert in the reconstruction process, whereas semi-automated and fully automated methods require little and no interaction respectively.

Computational complexity determines the time efficiency of the method for automated generation of an indoor model from a point cloud. The computation time of an indoor modelling algorithm varies with the size of the point cloud, complexity of the environment, and hardware specifications. In computer science the Big O notation is used to describe the computation time of an algorithm independent of hardware specifications. However, indoor modelling methods typically consist of several modules with different computational complexities. Therefore, we use computation time per million points on a standard CPU to measure the computational complexity of indoor modelling methods.

The quality of the generated model is perhaps the most important indicator for the benchmarking of indoor modelling methods. An indoor model consists of three main components: geometric elements, semantic attributes, and spaces with topological relations between them. For the latter two, we propose a qualitative assessment in which a panel of experts inspects the indoor model and checks the presence and correctness of semantic attributes and topological relations between the spaces. For the geometric elements, the evaluation has often been done by comparing the model with the input point cloud data. Computing the distances between the points and the geometric elements of the model provides an indication of the fidelity of the model to the data and therefore the accuracy of the model. This

approach is, however, data-dependent, and the quality measure might reflect the quality of the data rather than that of the model. To overcome this issue, we propose to evaluate the geometric quality of an indoor model through a comparison with a reference model. The comparison of two indoor models, however, faces a few challenges which we will discuss in the following section.

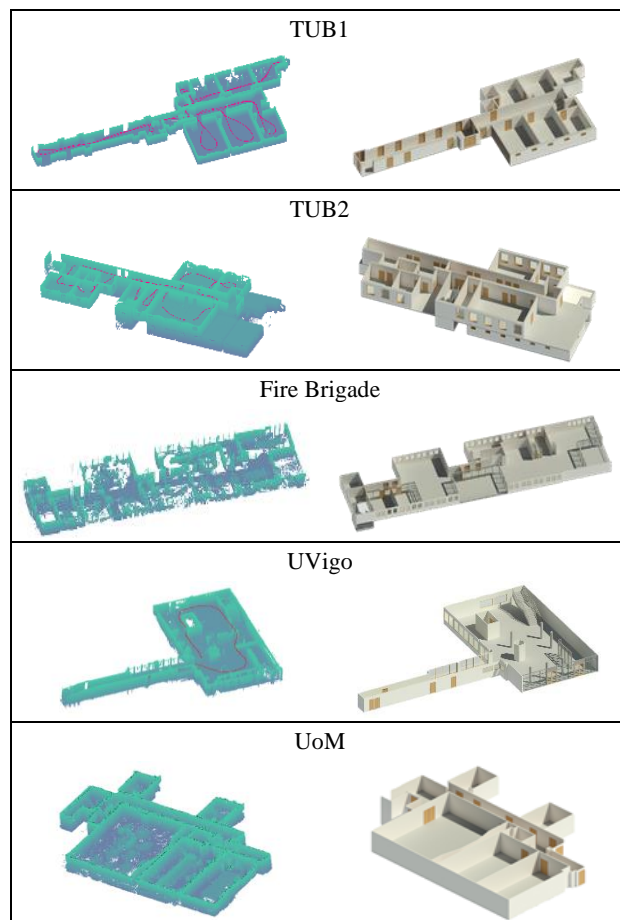


Figure 1. The benchmark point clouds and the corresponding reference models.

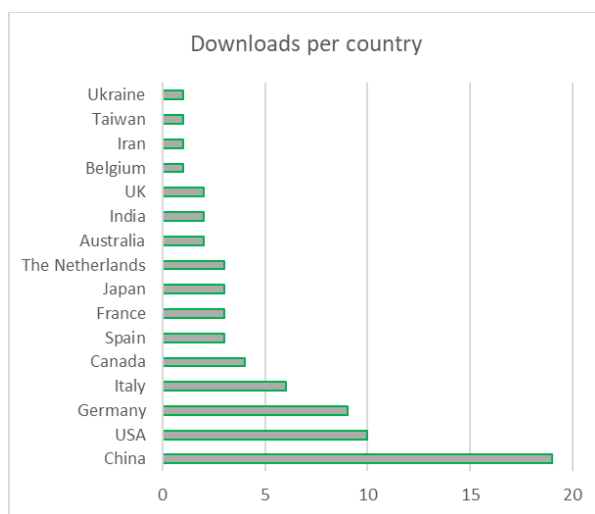


Figure 2. Download statistics of the benchmark dataset.

¹ <http://www2.isprs.org/commissions/comm4/wg5/benchmark-on-indoor-modelling.html>

4. MEASURING GEOMETRIC QUALITY OF INDOOR MODELS

Evaluating the geometric quality of an automatically generated indoor model, hereafter referred to as source, by comparison with a reference model faces the following challenges:

- i) There is no single commonly accepted standard for the representation of geometric elements in indoor models. While in the IFC standard geometric elements are represented as volumetric solids, in the CityGML standard these can be modelled as surfaces. Thus, the method for the geometric comparison of a source model with a reference should be applicable to both surface-based and volumetric models. Figure 3 shows an example of a surface-based model and a volumetric model of the same environment.
- ii) In any indoor environment there are elements that cannot be observed by the sensor and are therefore missing in the data. In the model these are often reconstructed based on the interpretation of the observed elements in the data. For example, when a wall is observed only from one side, the thickness of the wall may be interpreted from the other walls in the environment, which have been observed from both sides. The challenge is that these interpretations might be different in the source model and the reference. Thus, a reliable geometric quality evaluation requires that such interpreted elements are marked and excluded from the comparison. Figure 4 shows an example where the interpreted elements are different in the source model and the reference.
- iii) Measuring the correctness and completeness of geometric elements in the source requires a one-to-one correspondence between the source and reference elements. Such correspondence is generally not available. Figure 5 shows an example, where the walls of a room are modelled with different solids in the source and the reference.

To overcome the above challenges, we propose a method for evaluating the geometric quality of an indoor model, which takes into account interpreted surfaces, does not require one-to-one correspondence between the elements, and is applicable to both surface-based and volumetric models.

The proposed method measures the geometric quality of an indoor model based on three criteria: Completeness, Correctness, and Accuracy. These measures are computed through a comparison of the source model with the reference. The following sub-sections describe the method for computing the above measures.

4.1 Completeness

Completeness measures the extent to which the geometric elements within the reference are reconstructed in the source. It is measured based on the area of intersection between the source S and the reference R . Once interpreted surfaces are marked and excluded from the reference, a buffer b is created around each surface in R , and the area of intersection between the buffer and each surface in S is computed. The intersection areas are computed and summed over all surfaces S^i and R^j , thereby providing independence from one-to-one correspondence between the source and reference elements. Since the resulting completeness value varies with the size of the buffer, we define it as a function of the buffer size b :

$$M_{Comp}(b) = \frac{\sum_{i=1}^n \sum_{j=1}^m |S^i \cap b(R^j)|}{\sum_{j=1}^m |R^j|} \quad (1)$$

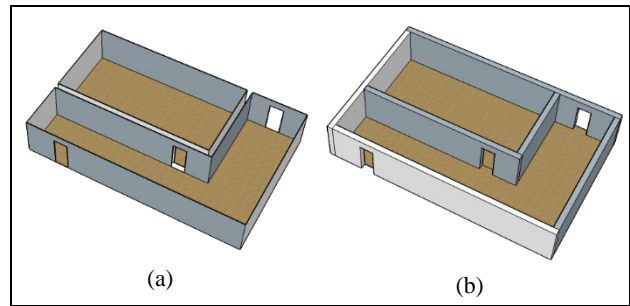


Figure 3. Different representations of a 3D indoor model: (a) surface-based representation; (b) volumetric representation.

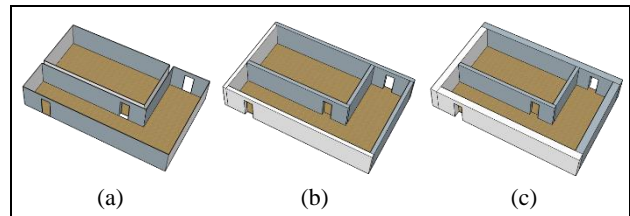


Figure 4. Different interpretations of unobservable elements in 3D indoor models: (a) observed wall surfaces; (b) source model with thin interpreted outer walls; (c) reference model with thick interpreted outer walls.

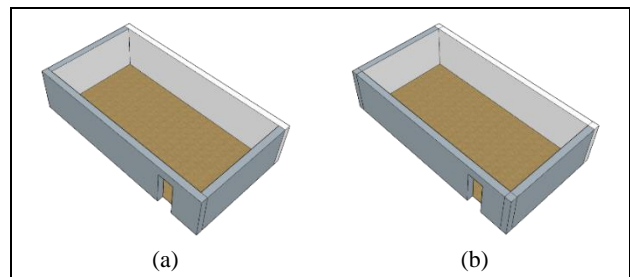


Figure 5. Example of lack of one-to-one correspondence between the source and the reference elements: (a) source model with four wall elements; (b) reference model with eight wall elements.

where m and n denote the number of surfaces in R and S respectively. To avoid the influence of irrelevant surfaces that might fall inside a buffer, only surfaces that are parallel up to a predefined threshold are used in each instance of intersection computation.

4.2 Correctness

Correctness measures the extent to which the geometric elements within the source are present in the reference. Similar to completeness, it is measured by computing the area of intersection between the source and the reference summed over all surfaces S^i and R^j . The correctness metric is also defined as a function of the buffer size b created around each observable reference surface:

$$M_{Corr}(b) = \frac{\sum_{i=1}^n \sum_{j=1}^m |S^i \cap b(R^j)|}{\sum_{i=1}^n |S^i|} \quad (2)$$

Similar to completeness, only surfaces that are parallel up to a predefined threshold are used in each instance of intersection computation.

4.3 Accuracy

Accuracy is measured based on the geometric distance between the source and the reference elements. Specifically, it is defined as the median of unsigned distances between the vertices of the source and the closest observable surfaces of the reference. Following Lehtola et al. (2017), we use a cut-off distance r to avoid the influence of incompleteness or incorrectness of the source model on the accuracy:

$$M_{Acc}(r) = Med\|\pi_j^T p_i\|, \text{ if } \|\pi_j^T p_i\| \leq r \quad (3)$$

where $\|\pi_j^T p_i\|$ is the perpendicular distance between a vertex point p_i in the source and the corresponding surface plane π_j in the reference (Khoshelham, 2015, 2016), and r is the cut-off value, distances beyond which are excluded from the median calculation. The vertex-surface correspondence is established based on the smallest distance and the condition that the perpendicular projection of the vertex on the surface falls within the surface boundary defined by its vertices (Oude Elberink and Khoshelham, 2015; Oude Elberink et al., 2013).

The above measures can be computed for different geometric elements, e.g., walls, doors, or windows, separately. It is also worth noting that these measures are by definition relative. They describe the completeness, correctness, and accuracy of a source model relative to a reference. For example, a source model with a high completeness rate is complete only up to the level of detail of the reference model. Therefore, a complete model may not necessarily be detailed, if the level of detail of the reference model is low.

5. EXPERIMENTS

The geometric quality measures were computed for a set of 3D models reconstructed automatically from the benchmark dataset using the shape grammar approach described in (Khoshelham and Díaz-Vilariño, 2014) and (Tran et al., 2018). Figure 6 shows the reconstructed source models and the corresponding reference models in which each surface is marked as either interpreted (dark grey) or observed (light grey and yellow). The source models contain the main structural elements, i.e., walls, floors and ceilings. The reference models were generated manually by a human expert and included walls, floors and ceilings, but also doors, windows, stairs, and columns. Since the latter elements were not reconstructed in the source models, they were excluded from the evaluation.

Figure 7 and Figure 8 present respectively the completeness and the correctness of the source models plotted against the buffer size ranging from 1 cm to 15 cm. All source models seem to have higher completeness rates while the correctness values are significantly lower. This can be explained by the presence of many interpreted surfaces in the source models (e.g., outer walls), which are considered as incorrect since in the reference models these interpreted surfaces were marked and excluded.

The completeness curves in Figure 7 show an increase of the completeness rates with increasing buffer size for all models except for UVigo, which has a high completeness rate even at small buffer sizes. This is observed also in the correctness curves shown in Figure 8, where correctness values increase with the buffer size except for the UVigo model. This can be due to the high reconstruction accuracy of the UVigo model compared to the other models.

Figure 9 shows the accuracy of the source models, and confirms that the UVigo model has been reconstructed with the highest accuracy indicated by a median vertex-surface distance of 0.5 cm for all cut-off distances smaller than 11 cm. At larger cut-off distances the accuracy is affected by the incorrect or incomplete surfaces.

To compare the geometric quality of different models or the performance of different methods one can compare the quality measures at a selected buffer size and cut-off distance. Table 1 shows a comparison of the quality measures for the buffer size and cut-off distance of 10 cm. It can be seen that TUB1 is less complete than the other models, while all models have a relatively low correctness rate. In terms of accuracy, UVigo is the most accurately reconstructed model with a median vertex-surface distance of 0.51 cm.

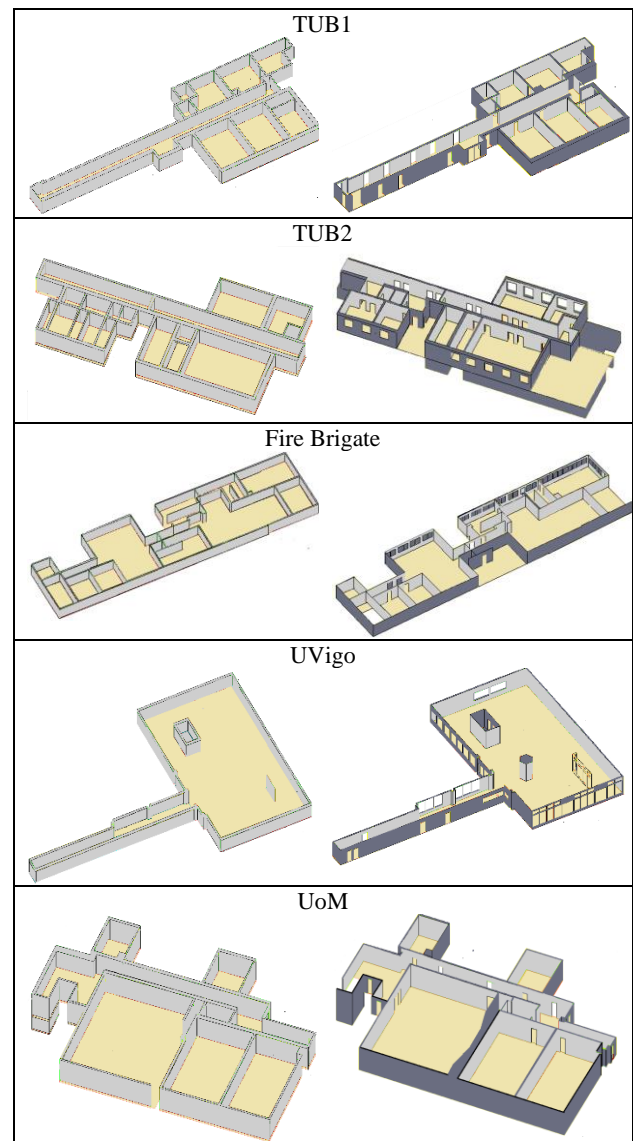


Figure 6. The reconstructed source models (left) and the corresponding reference models (right). Interpreted surfaces in the reference models are marked with dark grey colour, while light grey and yellow colours represent observed walls and floors respectively. The ceilings are removed for better visualization of the interior spaces.

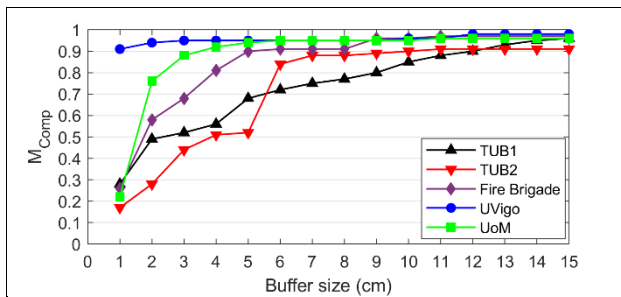


Figure 7. The completeness of the source models plotted against increasing buffer size.

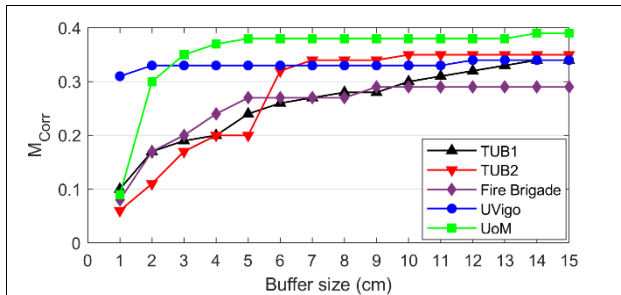


Figure 8. The correctness of the source models plotted against increasing buffer size.

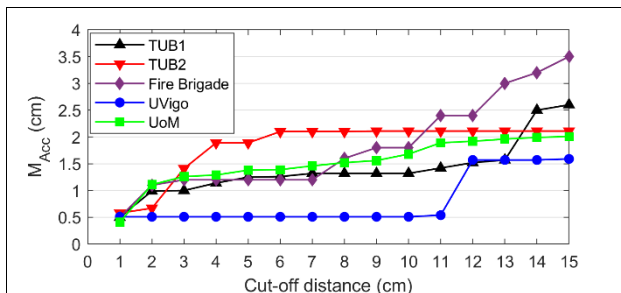


Figure 9. The accuracy of the source models plotted against increasing cut-off distance.

Measure →	Completeness	Correctness	Accuracy (cm)
Model ↓	@ b = 10 cm	@ b = 10 cm	@ r = 10 cm
TUB1	0.85	0.30	1.32
TUB2	0.90	0.35	2.11
Fire Brigade	0.96	0.29	1.80
UVigo	0.96	0.33	0.51
UoM	0.95	0.38	1.68

Table 1. Geometric quality measures for the five reconstructed models at the buffer size and cut-off distance of 10 cm.

6. CONCLUSIONS

This paper presented an evaluation framework for the benchmarking of indoor modelling methods. We discussed various quality aspects of indoor models, and challenges in measuring the geometric quality of a reconstructed model through comparison with a reference model. We proposed a method for evaluating the geometric quality of an indoor model, which takes into account interpreted surfaces, does not require one-to-one correspondence between the source and reference elements, and is applicable to both surface-based and volumetric models. The results of experiments with models reconstructed from the benchmark dataset demonstrated the potential of this

method for automatic evaluation and comparison of the geometric quality of indoor models.

ACKNOWLEDGEMENTS

This work was supported by the ISPRS Scientific Initiatives 2017. The TUB1 and TUB2 datasets were provided by Prof Markus Gerke, Technical University of Braunschweig (Germany), in collaboration with Viametris and Laser Scanning Europe. The Fire Brigade dataset was provided by Prof Sisi Zlatanova, Delft University of Technology (The Netherlands), and was acquired for the project SIMS3D funded by the Netherlands Technology Foundation (STW). The UVigo dataset was provided by Prof Pedro Arias, University of Vigo (Spain), and was acquired for the project ENGINENCY funded by the program H2020-FTIPilot-2015-1. The UoM dataset was provided by Dr Ebadat Ghanbari Parmehr, RMIT University (Australia). The authors acknowledge the financial support from the University of Melbourne through Melbourne International Fee Remission and Melbourne International Research Scholarship. The support from Xunta de Galicia through the human resources grant ED481B 2016/079-0 is also gratefully acknowledged.

REFERENCES

- Becker, S., Peter, M., Fritsch, D., 2015. Grammar-Supported 3D Indoor Reconstruction From Point Clouds for "As-Built" BIM. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* II-3/W4, 17-24.
- Díaz-Vilariño, L., Khoshelham, K., Martínez-Sánchez, J., Arias, P., 2015. 3D Modeling of Building Indoor Spaces and Closed Doors from Imagery and Point Clouds. *Sensors* 15, 3491-3512.
- Gunduz, M., Isikdag, U., Basaraner, M., 2016. A Review of Recent Research in Indoor Modelling & Mapping. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* XLI-B4, 289-294.
- Khoshelham, K., 2015. Direct 6-DoF Pose Estimation from Point-Plane Correspondences. *Digital Image Computing: Techniques and Applications (DICTA)*, 2015 International Conference on, pp. 1-6.
- Khoshelham, K., 2016. Closed-form solutions for estimating a rigid motion from plane correspondences extracted from point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing* 114, 78-91.
- Khoshelham, K., Díaz-Vilariño, L., 2014. 3D modeling of interior spaces: learning the language of indoor architecture. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XL-5, 321-326.
- Khoshelham, K., Vilariño, L.D., Peter, M., Kang, Z., Acharya, D., 2017. The ISPRS Benchmark on Indoor Modelling. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2/W7, 367-372.
- Lehtola, V., Kaartinen, H., Nüchter, A., Kajaluoto, R., Kukko, A., Litkey, P., Honkavaara, E., Rosnell, T., Vaaja, M., Virtanen, J.-P., Kurkela, M., El Issaoui, A., Zhu, L., Jaakkola, A., Hyypä, J., 2017. Comparison of the Selected State-Of-The-Art 3D Indoor Scanning and Point Cloud Generation Methods. *Remote Sensing* 9, 796.

- Macher, H., Landes, T., Grussenmeyer, P., 2017. From Point Clouds to Building Information Models: 3D Semi-Automatic Reconstruction of Indoors of Existing Buildings. *Applied Sciences* 7, 1030.
- Mura, C., Mattausch, O., Pajarola, R., 2016. Piecewise-planar Reconstruction of Multi-room Interiors with Arbitrary Wall Arrangements. *Computer Graphics Forum* 35, 179-188.
- Ochmann, S., Vock, R., Wessel, R., Klein, R., 2016. Automatic reconstruction of parametric building models from indoor point clouds. *Computers & Graphics* 54, 94-103.
- Oesau, S., Lafarge, F., Alliez, P., 2014. Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing* 90, 68-82.
- Oude Elberink, S., Khoshelham, K., 2015. Automatic Extraction of Railroad Centerlines from Mobile Laser Scanning Data. *Remote Sensing* 7, 5565-5583.
- Oude Elberink, S., Khoshelham, K., Arastounia, M., Diaz Benito, D., 2013. Rail Track Detection and Modelling in Mobile Laser Scanner Data. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* II-5/W2, 223-228.
- Pătrăucean, V., Armeni, I., Nahangi, M., Yeung, J., Brilakis, I., Haas, C., 2015. State of research in automatic as-built modelling. *Advanced Engineering Informatics* 29, 162-171.
- Tang, P., Huber, D., Akinci, B., Lipman, R., Lytle, A., 2010. Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in Construction* 19, 829-843.
- Thomson, C., Boehm, J., 2015. Automatic Geometry Generation from Point Clouds for BIM. *Remote Sensing* 7, 11753.
- Tran, H., Khoshelham, K., Kealy, A., Díaz-Vilariño, L., 2017. Extracting Topological Relations Between Indoor Spaces From Point Clouds. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* IV-2/W4, 401-406.
- Tran, H., Khoshelham, K., Kealy, A., Díaz-Vilariño, L., 2018. Shape Grammar Approach to 3D Modelling of Indoor Environments Using Point Clouds. *Journal of Computing in Civil Engineering*, In Press.
- Valero, E., Adán, A., Cerrada, C., 2012. Automatic Method for Building Indoor Boundary Models from Dense Point Clouds Collected by Laser Scanners. *Sensors* 12, 16099-16115.
- Xiao, J., Furukawa, Y., 2014. Reconstructing the World's Museums. *Int. J. Comput. Vision* 110, 243-258.
- Xiong, X., Adan, A., Akinci, B., Huber, D., 2013. Automatic creation of semantically rich 3D building models from laser scanner data. *Automation in Construction* 31, 325-337.