THE ISPRS BENCHMARK ON INDOOR MODELLING – PRELIMINARY RESULTS

K. Khoshelham [a], H. Tran [a], D. Acharya [a], L. Díaz Vilariño [b], Z. Kang [c] , S. Dalyot [d]

[a] Dept. of Infrastructure Engineering, The University of Melbourne, Parkville 3010 Australia – {k.khoshelham, ha.tran, debaditya.acharya}@unimelb.edu.au
[b] Applied Geotechnologies Group, School of Industrial Engineering, University of Vigo, Spain – lucia@uvigo.es
[c] 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
[d] Environmental Crowdsourcing Laboratory, Faculty of Civil and Environmental Engineering, Technion - Israel Institute of Technology, Haifa 3200003, Israel – dalyot@technion.ac.il

Commission IV, WG IV/5

KEY WORDS: 3D modelling, Reconstruction, Point cloud, BIM, Quality, Evaluation, Performance, Benchmarking, Automation.

ABSTRACT:

Automated 3D reconstruction of indoor environments from point clouds has been a topic of intensive research in recent years. Different methods developed for the generation of 3D indoor models have achieved promising results on different case studies. However, a comprehensive evaluation and comparison of the performance of these methods has not been available. This paper presents the preliminary results of the ISPRS benchmark on indoor modelling, an initiative of Working Group IV/5 to benchmark the performance of indoor modelling methods using a public dataset and a comprehensive evaluation framework. The performances of the different methods are compared through geometric quality evaluation of the reconstructed models in terms of completeness, correctness, and accuracy of wall elements. The results show that the reconstruction methods generally achieve high completeness but lower correctness for the reconstructed models while accuracies range from 0.5 cm to 6.7 cm.

1. INTRODUCTION

3D models of indoor environments find applications in navigation (Díaz-Vilariño et al., 2016), emergency response (Rueppel and Stuebbe, 2008), and a range of location-based services (Gu et al., 2019). Because the manual creation of such models is a slow and tedious task, in recent years many research efforts have been made to develop methods for automated reconstruction of 3D indoor models. While these methods have shown promising results on different case studies, a comprehensive evaluation and comparison of the performance of these methods has not been available.

The ISPRS benchmark on indoor modelling is an initiative of Working Group IV/5 to fill this gap. Launched in 2017, the project created a public dataset comprising six point clouds of different indoor environments and developed a framework for performance evaluation and comparison of indoor modelling methods. In 2019, the project was further supported by the ISPRS to organise a benchmark test to experimentally evaluate and benchmark the performance of indoor modelling methods. The project team issued a call for participation and received nine submissions. The submitted models were evaluated using the developed framework and the results were published on the ISPRS website for the benchmark test[1].

This paper presents the preliminary results of the benchmark test for the first six submissions. The performances of the reconstruction methods are compared through geometric quality evaluation of the reconstructed models in terms of completeness, correctness, and accuracy. The current evaluation focuses on wall elements only and other structural (e.g., floors and ceilings) and non-structural elements (e.g., doors, windows, and spaces) as well as semantic properties are excluded from the evaluation. The results show that the reconstruction methods participating in the benchmark generally achieve high completeness and relatively lower correctness for the reconstructed models while accuracies range from 0.5 cm to 6.7 cm.

The paper proceeds with an overview of the benchmark dataset in Section 2. The six submissions to the benchmark test are introduced in Section 3. The evaluation method is described in Section 4. The preliminary results of the benchmark test are presented in Section 5. The paper concludes with a summary of the findings in Section 6.

2. BENCHMARK DATASET

The benchmark dataset comprises six point clouds captured in indoor environments representing different levels of complexity. From each point cloud a 3D model was generated manually to serve as reference for the evaluation of the automatically reconstructed 3D models. Figure 1 shows the point clouds and the corresponding reference models. The point clouds were made publicly available via the benchmark website[2] for potential test participants as well as the wider research community to encourage further research on developing indoor modelling methods. The reference models, however, were withheld from the test participants to enable a fair evaluation and comparison of the submitted models without bias.

Table 1 summarises the characteristics of the point clouds and the indoor environments. The dataset represents indoor environments with different levels of complexity, including cluttered, multi-storey, and non-Manhattan-World indoor environments. A detailed description of the benchmark dataset including sensor specifications and the method for the generation of the reference models is provided in (Khoshelham et al., 2017).

[1] http://www2.isprs.org/commissions/comm4/wg5/benchmark-test-on-indoor-modelling.html

[2] http://www2.isprs.org/commissions/comm4/wg5/benchmark-on-indoor-modelling/datasets.html

Dataset	Sensor	Nr of points	Average point spacing (cm)	Sensor trajectory	Clutter	Multi-storey	Manhattan-World
TUB1	Viametris iMS3D	33.6×10^6	0.5	Yes	Low	No	Yes
TUB2	Zeb Revo	21.6×10^6	0.8	Yes	Low	Yes	Yes
Fire Brigade	TLS Leica C10	14.1×10^6	1.1	No	High	No	Yes
UVigo	UVigo Backpack	14.9×10^6	1.0	Yes	Moderate	No	Yes
UoM	Zeb1	13.9×10^6	0.7	No	Moderate	No	Yes*
Grainger Museum	Zeb Revo RT	28.9×10^6	2.9	Yes	High	No	No

* Partial deviation from Manhattan-World configuration.

Table 1. Characteristics of the benchmark dataset and the indoor environments.

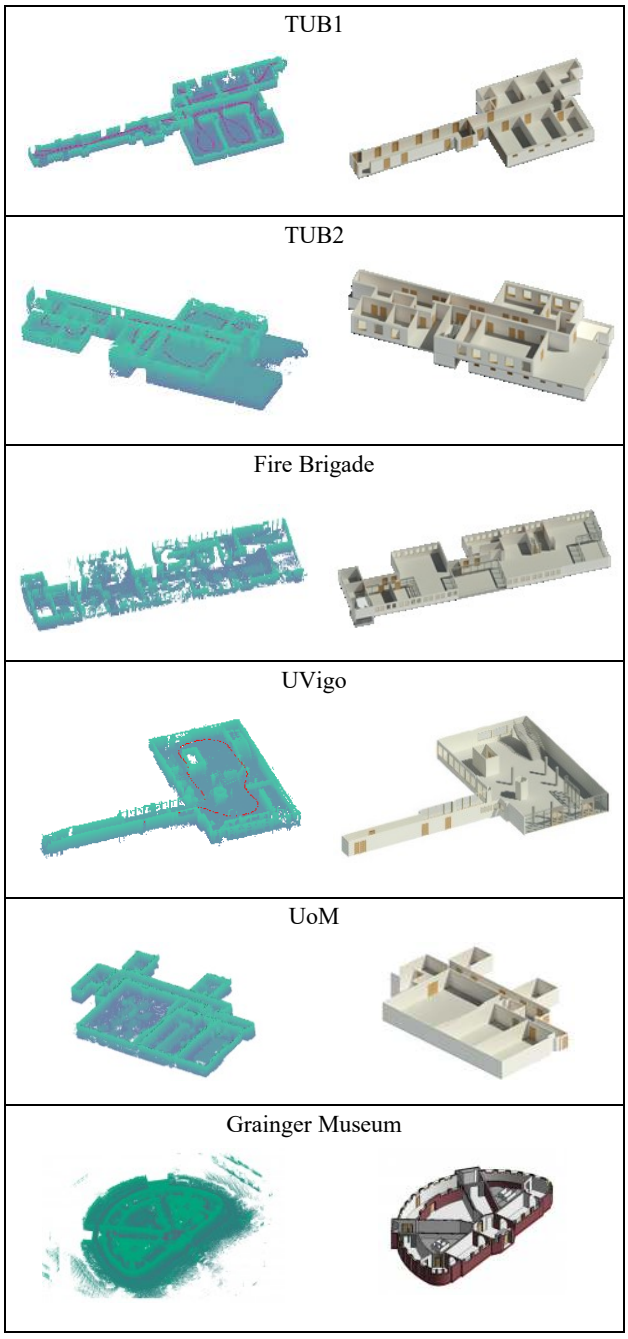

Figure 1. The benchmark point clouds (left) and the corresponding reference models (right).

3. SUBMISSIONS

The first six submissions to the benchmark test were included in the current evaluation. Table 2 provides a summary of the six submissions. Note that Grainger Museum is not reconstructed by any of the participants. This is because this dataset was added to the benchmark after the initial call for participation was issued. Previtali et al. (2018) and Cui et al. (2019) submitted only one model (TUB1 and TUB2 respectively). Tran and Khoshelham (2019) submitted four models (TUB1, Fire Brigade, UVigo, and UoM). Ochmann et al. (2019), Maset et al. (2019), and Tran et al. (2019b) submitted five models (TUB1, TUB2, Fire Brigade, UVigo, and UoM).

Authors	Affiliation	Reconstructed model					
		TUB1	TUB2	Fire Brigade	UVigo	UoM	Grainger Museum
Cui et al.	Shenzhen University	–	✓	–	–	–	–
Ochmann et al.	University of Bonn	✓	✓	✓	✓	✓	–
Maset et al.	Udine University	✓	✓	✓	✓	✓	–
Previtali et al.	Polytechnic University of Milan	✓	–	–	–	–	–
Tran et al.	University of Melbourne	✓	✓	✓	✓	✓	–
Tran & Khoshelham	University of Melbourne	✓	–	✓	✓	✓	–

Table 2. Summary of the submissions included in the current evaluation.

4. EVALUATION METHOD

The geometric quality of the submitted models, hereafter referred to as the source, is evaluated by comparing these with the reference models. The current quality evaluation focuses on the geometry of the reconstructed models and is based on the following measures: completeness, correctness, and accuracy.

Completeness is defined as the proportion of the reference elements reconstructed in the source. It is measured by computing the area of intersection between the source and reference elements within a buffer:

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

where $b(.)$ denotes the buffer with size b created around a visible reference surface R^j, and n, m are the number of surfaces in the source S and in the reference R respectively. The operator $|.|$ denotes the area of the surface and $\mathcal{P}(S^i)$ denotes the orthogonal projection of the source surface S^i on its corresponding reference surface.

Correctness is defined as the proportion of the source elements that are present in the reference. It is also measured by computing the area of intersection between the source and reference elements within a buffer b:

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

Accuracy is defined as the geometric closeness of the source elements to their corresponding reference elements. It is measured by computing the median Euclidean distance between sample points representing the reference model and the closest surfaces in the source model:

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

where ${\pi_j}^T p_i$ is the orthogonal distance between the source point p_i and the reference surface plane π_j both represented by homogeneous coordinates (Khoshelham, 2015, 2016), r is the cut-off distance, and $|.|$ denotes the absolute value. A smaller M_{Acc} indicates higher accuracy of the reconstructed model. A more detailed description of the evaluation method can be found in (Khoshelham et al., 2018) and (Tran et al., 2019a).

5. RESULTS

The above measures of completeness, correctness and accuracy were computed for each of the submitted models but only for wall elements. This choice was made because wall elements were the only elements reconstructed in all submitted models, whereas other elements, such as spaces, doors and windows, were present only in some of the submitted models. The completeness and correctness scores were computed for buffer sizes ranging from 1 cm to 15 cm. The accuracies were also computed for cut-off distances ranging from 1 cm to 15 cm. For the comparison of the results, the completeness and correctness scores at a buffer size of 10 cm, and the accuracies at a cut-off distance of 10 cm were considered.

Table 3 summarises the evaluation results for all submissions. Note that smaller values for accuracy indicate higher accuracy of the reconstructed models, whereas for completeness and correctness larger values indicate more complete and correct models. Overall, the results show that the reconstruction methods achieve higher completeness and lower correctness while accuracies range from 0.5 cm to 6.7 cm. The relatively high completeness and low correctness scores mean that the reconstructed models contain most of the wall elements that are present in the corresponding reference models, but they also include a considerable number of incorrect wall surfaces. This is due to the fact that the reconstructed models contain invisible surfaces, i.e. surfaces that are not present in the point cloud. These surfaces are not necessarily incorrect. For example, the top surface of a wall is not visible in the point cloud but is often included in the reconstructed model for completeness. However, in the reference models these invisible surfaces are marked as such and are excluded from the computation of the correctness measure. This results in a low correctness score for the reconstructed models that include invisible surfaces.

TUB1

Authors	Affiliation	Completeness @ 10 cm	Correctness @ 10 cm	Accuracy @ 10 cm (cm)
Maset et al.	Udine University	0.83	0.48	1.80
Ochhmann et al.	University of Bonn	0.93	0.37	1.82
Previtali et al.	Politecnico di Milano	0.77	0.50	2.23
Tran et al.	University of Melbourne	0.84	0.31	1.34
Tran & Khoshelham	University of Melbourne	0.82	0.78	2.71

TUB2

Authors	Affiliation	Completeness @ 10 cm	Correctness @ 10 cm	Accuracy @ 10 cm (cm)
Cui et al.	Shenzhen University	0.53	0.56	5.99
Maset et al.	Udine University	0.62	0.52	4.16
Ochhmann et al.	University of Bonn	0.83	0.44	2.75
Tran et al.	University of Melbourne	0.88	0.39	2.12

Fire Brigade

Authors	Affiliation	Completeness @ 10 cm	Correctness @ 10 cm	Accuracy @ 10 cm (cm)
Maset et al.	Udine University	0.63	0.36	4.84
Ochhmann et al.	University of Bonn	0.65	0.13	2.79
Tran et al.	University of Melbourne	0.96	0.29	1.41
Tran & Khoshelham	University of Melbourne	0.78	0.35	2.59

UVigo

Authors	Affiliation	Completeness @ 10 cm	Correctness @ 10 cm	Accuracy @ 10 cm (cm)
Maset et al.	Udine University	0.44	0.24	4.63
Ochhmann et al.	University of Bonn	0.55	0.19	4.46
Tran et al.	University of Melbourne	0.89	0.29	0.51
Tran & Khoshelham	University of Melbourne	0.58	0.49	6.66

UoM

Authors	Affiliation	Completeness @ 10 cm	Correctness @ 10 cm	Accuracy @ 10 cm (cm)
Maset et al.	Udine University	0.72	0.52	3.10
Ochhmann et al.	University of Bonn	0.87	0.34	3.06
Tran et al.	University of Melbourne	0.92	0.44	0.92
Tran & Khoshelham	University of Melbourne	0.73	0.77	3.77

Table 3. Evaluation results for the current submissions.

A comparison of the results for the different datasets shows that in general the models reconstructed for TUB1 and UoM are relatively more complete and more correct than those for other datasets. This can be attributed to the lower complexity and perhaps the better data quality of these two datasets. As shown in Table 1 and Figure 1, TUB1 and UoM represent relatively simple environments and the corresponding point clouds have the highest point density among the benchmark datasets.

The completeness scores in Table 3 show the model reconstructed by Ochmann et al. (2019) has the highest completeness for TUB1, while for the other datasets Tran et al. (2019b) achieve higher completeness scores. In particular, the completeness of the model of UVigo reconstructed by Tran et al. is significantly higher than those of the other submissions. This can be more clearly seen in Figure. 2, which shows the completeness of the different models of UVigo for increasing buffer sizes. Figure 3 shows the completeness score computed for individual surfaces of the UVigo model reconstructed by Tran et al. (2019b) visualised on the reference model of UVigo. It can be seen that most surfaces of the reference model indeed have high completeness scores indicating that they are present in the reconstrued model.

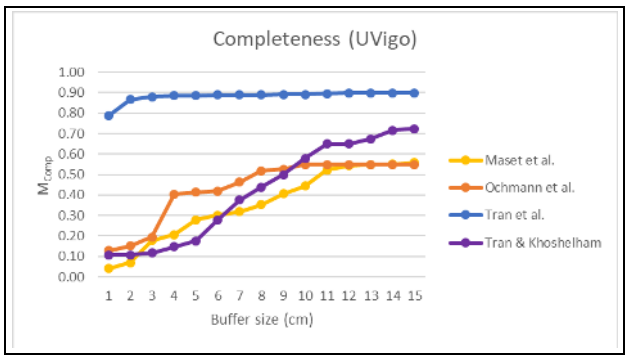

Figure 2. Completeness of submitted models for UVigo dataset.

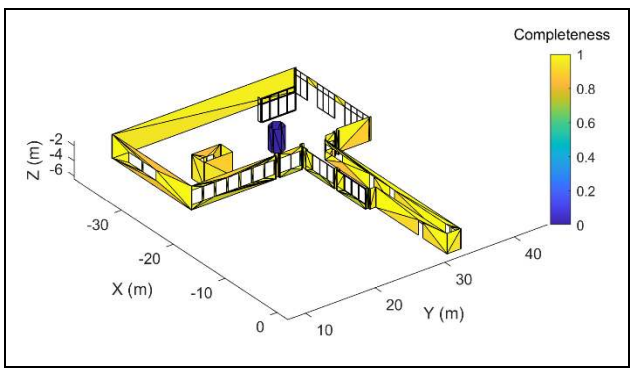

Figure 3. Completeness of individual surfaces of the UVigo model reconstructed by Tran et al. (2019b) shown on the reference model of UVigo.

The correctness scores in Table 3 show that all methods achieve lower correctness scores as compared to completeness scores on almost all datasets. The difference between the correctness and completeness of the reconstructed models is more pronounced for the Fire Brigade and UVigo datasets. For instance, the most correct model of Fire Brigade (Maset et al.), has a correctness score of only 0.36, whereas the most complete model of Fire Brigade (Tran et al.) has a completeness score of 0.96. The Fire Brigade and UVigo datasets are characterised by high levels of clutter, such as furniture and other indoor objects. As many of these objects have planar surfaces they can be incorrectly reconstructed as wall surfaces resulting in lower correctness scores for these two datasets.

A comparison of the performance of different methods in terms of correctness reveals that Tran and Khoshelham (2019) achieve relatively higher correctness on most datasets. This can be seen for instance in the correctness results for TUB1 as shown in Figure 4, which compares the correctness of the submitted models for increasing buffer sizes. The better performance of the method of Tran and Khoshelham (2019) in terms of correctness can be attributed to the consideration of global likelihood and model plausibility in the reconstruction process, where models with low global likelihood and low plausibility, i.e. those with little support from the points in the point cloud, are discarded. As a result, the final reconstructed model has a high global likelihood, which means it contains few invisible and incorrect surfaces resulting in a higher correctness score.

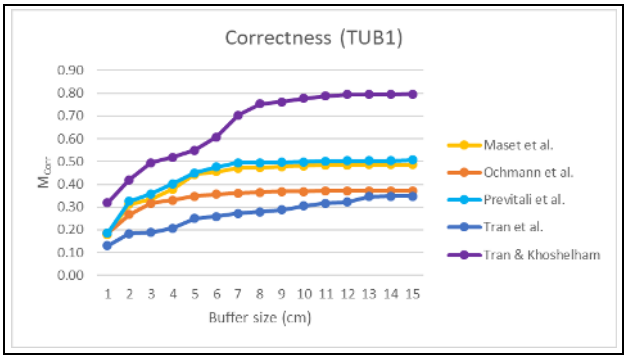

Figure 4. Correctness of submitted models for TUB1 dataset.

The accuracy measures shown in Table 3 reveal that the models reconstructed by Tran et al. (2019b) have the highest accuracy for all datasets. This can be seen for instance in the accuracy results for UoM as shown in Figure 5, which compares the accuracy of the submitted models for increasing cut-off distances. The higher accuracy of the method of Tran et al. can be attributed to the use of point coordinate histograms for the detection of wall surfaces. Unlike the other methods, which detect surfaces in arbitrary orientations, the use of point coordinate histograms results in reconstructed surfaces that are perpendicular to the x, y, z axes of the point cloud. These surfaces are geometrically closer to the reference surfaces because the orthogonality constraint is often applied by the human expert during the manual generation of reference models.

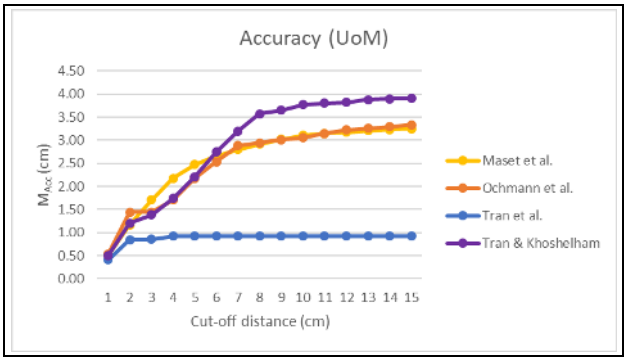

Figure 5. Accuracy of submitted models for UoM dataset.

6. CONCLUSIONS

This paper presented the preliminary results of the ISPRS benchmark test on indoor modelling. The performances of the reconstruction methods developed by the six participants in the benchmark test were compared through geometric quality evaluation of the reconstructed models in terms of completeness, correctness, and accuracy of the wall elements. The results showed that the reconstruction methods generally achieve high completeness and relatively lower correctness for the reconstructed models while accuracies range from 0.5 cm to 6.7 cm. It was also found that the results vary with the complexity of the indoor environment, level of clutter, as well as accuracy and density of the point clouds. A comparison of the reconstruction methods showed that the method of Tran et al. (2019b) achieved higher completeness and accuracy on most datasets, while the method of Tran and Khoshelham (2019) achieved better results in terms of correctness.

At the time of writing, several new submissions were made to the benchmark test. A more complete evaluation and comparison of all submissions will be carried out and the results will be published on the ISPRS website for the benchmark test[3] and in future publications.

ACKNOWLEDGEMENTS

This work was supported by the ISPRS Scientific Initiatives 2019. The authors acknowledge the financial support from the University of Melbourne through Melbourne International Fee Remission and Melbourne International Research Scholarship.

REFERENCES

Cui, Y., Li, Q., Yang, B., Xiao, W., Chen, C., Dong, Z., 2019. Automatic 3-D Reconstruction of Indoor Environment With Mobile Laser Scanning Point Clouds. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 12, 3117-3130.

Díaz-Vilariño, L., Boguslawski, P., Khoshelham, K., Lorenzo, H., Mahdjoubi, L., 2016. Indoor navigation from point clouds: 3D modelling and obstacle detection, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLI-B4. Copernicus Publications, Prague, Czech Republic, pp. 275-281.

Gu, F., Hu, X., Ramezani, M., Acharya, D., Khoshelham, K., Valaee, S., Shang, J., 2019. Indoor Localization Improved by Spatial Context - A Survey. ACM Computing Surveys (CSUR) 52, 64.

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., Tran, H., Díaz-Vilariño, L., Peter, M., Kang, Z., Acharya, D., 2018. An Evaluation Framework for Benchmarking Indoor Modelling Methods, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., Vol. XLII-4. Copernicus Publications, Delft, The Netherlands, pp. 297-302.

Khoshelham, K., Vilariño, L.D., Peter, M., Kang, Z., Acharya, D., 2017. The ISPRS Benchmark on Indoor Modelling, International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences, Vol. XLII-2/W7, pp. 367-372.

Maset, E., Magri, L., Fusiello, A., 2019. Improving Automatic Reconstruction of Interior Walls from Point Cloud Data, Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. Copernicus Publications, pp. 849-855.

Ochmann, S., Vock, R., Klein, R., 2019. Automatic reconstruction of fully volumetric 3D building models from oriented point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 151, 251-262.

Previtali, M., Díaz-Vilariño, L., Scaioni, M., 2018. Indoor Building Reconstruction from Occluded Point Clouds Using Graph-Cut and Ray-Tracing. Applied Sciences 8, 1529.

Rueppel, U., Stuebbe, K.M., 2008. BIM-based indoor-emergency-navigation-system for complex buildings. Tsinghua Science and Technology 13, 362-367.

Tran, H., Khoshelham, K., 2019. A Stochastic Approach to Automated Reconstruction of 3D Models of Interior Spaces from Point Clouds, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. Copernicus Publications, pp. 299-306.

Tran, H., Khoshelham, K., Kealy, A., 2019a. Geometric comparison and quality evaluation of 3D models of indoor environments. ISPRS Journal of Photogrammetry and Remote Sensing 149, 29-39.

Tran, H., Khoshelham, K., Kealy, A., Díaz-Vilariño, L., 2019b. Shape Grammar Approach to 3D Modelling of Indoor Environments Using Point Clouds. Journal of Computing in Civil Engineering 33, 04018055.

[3] http://www2.isprs.org/commissions/comm4/wg5/benchmark-test-on-indoor-modelling.html