https://doi.org/10.2352/ISSN.2470-1173.2020.17.3DMP-036

Quality assessment for 3D reconstruction of building interiors

Umamaheswaran RAMAN KUMAR[1], Inge COUDRON[2], Steven PUTTEMANS[3], Patrick VANDEWALLE[1];
[1]KU Leuven, Belgium [2]Flanders Make, Belgium [3]VLAIO, Belgium

Abstract

Applications ranging from simple visualization to complex design require 3D models of indoor environments. This has given rise to advancements in the field of automated reconstruction of such models. In this paper, we review several state-of-the-art metrics proposed for geometric comparison of 3D models of building interiors. We evaluate their performance on a real-world dataset and propose one tailored metric which can be used to assess the quality of the reconstructed model. In addition, the proposed metric can also be easily visualized to highlight the regions or structures where the reconstruction failed. To demonstrate the versatility of the proposed metric we conducted experiments on various interior models by comparison with ground truth data created by expert Blender artists. The results of the experiments were then used to improve the reconstruction pipeline.

Introduction

With recent advancements in virtual reality applications like virtual tours and interactive interior design, there is an increasing demand for realistic and semantically rich 3D models of indoor environments. Manual generation of these models from scanned point clouds is a time consuming and labor-intensive process. In order to address this demand, various reconstruction techniques have been proposed and developed over the years to automate the model reconstruction pipeline ([1], [2], [3], [4], [5], [6], [7], [8]). However there does not exist a well defined quality metric to assess the generated models. There are methods to evaluate the building facades ([9], [10]) which are usually one exterior side, like building front that faces the street or roof from an aerial view. These methods do not intrinsically work well for building interiors because they have more surfaces and complex structures compared to facades. Also, when comparing different reconstruction models it is necessary to localize the regions where the reconstruction performs badly and these are not visible in a single metric value. Both quantitative and visual feedback are equally important to select the right approach and to suggest areas of improvement in the reconstruction pipeline.

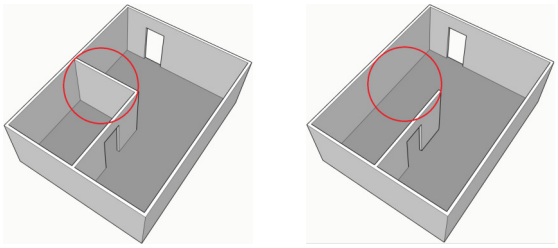
Related work

We review several state-of-the-art metrics ([11], [12]). The 3D models are assessed by comparing the surfaces of the source model, S, with the surfaces of the reference ground truth model, R. There are two main representations for 3D models, namely volumetric and surface representations. The metrics which are chosen work well with both representations as they are applied on surface projections. In the following sub-sections, $P(.)$ denotes the projection of a source volume/surface onto a reference surface and $b(.)$ denotes the buffer range considered around a surface.

Completeness

The completeness metric, M_{Comp} represents how much of the reference model is present in the generated source model. It is calculated as the intersection of the source and reference surfaces divided by the total surface in the reference model (Eqn. 1). The value is in a range $[0,1]$ where 1 denotes a high completeness score.

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



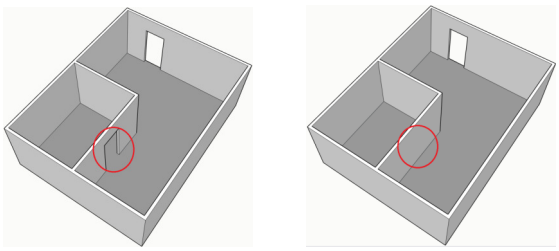
(a) Reference model, R (b) Source model, S

Figure 1. *Source model with missing wall compared to reference model highlighted with red circle gives a completeness score of 0.9634.*

Correctness

The correctness metric, M_{Corr} represents how much of the source model is present in the reference model. It is calculated as the intersection of the source and reference surfaces divided by the total surface in the source model (Eqn. 2). The value is in a range $[0,1]$ where 1 denotes a high correctness score.

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



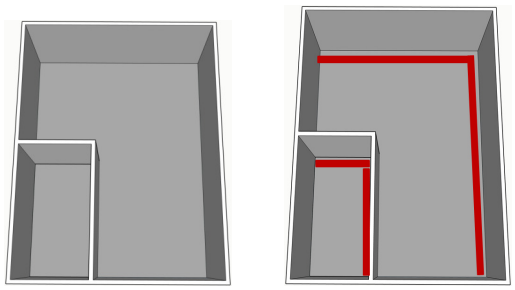
(a) Reference model, R (b) Source model, S

Figure 2. *Source model with missing door compared to reference model highlighted with red circle gives a correctness score of 0.9946.*

Accuracy

The accuracy metric, M_{Acc} represents how close the correctly generated surfaces of the source model represents the reference model. It is calculated as the median of all the distances, lesser than a threshold value r (similar value as buffer size b in most cases), between points p_i representing the uniformly sampled reference model and the closest surface π_j in the source model (Eqn. 3). The value is in a range $[0, r]$ and smaller values represent more accurate models.

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



(a) Reference model, R (b) Source model, S

Figure 3. *Source model with shifted walls compared to reference model highlighted with red lines gives an accuracy score of 5cm.*

Method

The state-of-the-art metrics when used individually to compare models do not serve as a good comparison criterion. We thus smartly combine these metrics to serve the specific application domain of building interior reconstruction. This section explains the various approaches tested to arrive at the final metric solution and evaluates the results of the ISPRS (International Society for Photogrammetry and Remote Sensing) benchmark on indoor modeling [13] compared to the state-of-the-art metrics.

Metrics weighting

Each of the described metrics targets a very specific aspect of the reconstruction process. Considering the reconstruction as a binary classification problem where every voxel is classified as wall or empty space, Table 1 shows the regions of the confusion matrix where each of the metrics work well and where they suffer.

Table 1: Metrics working regions in confusion matrix

	M_{Comp}	M_{Corr}	M_{Acc}
1. True positive	✓	✗	✓
2. True negative	✗	✓	✓
3. False positive	✗	✓	✗
4. False negative	✓	✗	✗

The completeness metric is mainly used to identify true positives and false negatives whereas the correctness metric is used to identify true negatives and false positives. The accuracy metric is a median value and therefore, it is not a very good discriminator as the value is not continuous and can have abrupt changes. However, for building reconstruction we deal mainly with planar surfaces with less discontinuities. In such cases, it can help identify better models among the models having good completeness and correctness scores. The final inference metric M_{Inf} expressed in percentage (%) has two terms, a function of completeness and correctness scores $f(M_{Comp}, M_{Corr})$ with a weight of 90% and a normalised accuracy score M_{Acc_norm} with a weight of 10%:

$$M_{Inf} = (90 \times f(M_{Comp}, M_{Corr})) + (10 \times M_{Acc\_norm}) \tag{4}$$

The formulation of $f(M_{Comp}, M_{Corr})$ and M_{Acc_norm} are explained in the following subsections.

Function of completeness and correctness

In order to have a good reconstruction result it is necessary to formulate a function that satisfies the following conditions:

- $f(M_{Comp}, M_{Corr}) \leq min(M_{Comp}, M_{Corr})$. This makes sure that the reconstruction is considered to be good only if both metric scores are high. A high value of completeness and low value of correctness means there are too many false positives and a low value of completeness and high value of correctness means there are too few true positives.
- $f(M_{Comp}, M_{Corr})$ increases linearly when either M_{Comp} or M_{Corr} is a constant. This condition is required to have a steady increase in the function and without any bias.

Eqns. (5), (6) and (7) shows the different formulations to combine the completeness and correctness metric scores.

Arithmetic mean (AM)

The arithmetic mean is defined as the sum of the metric scores divided by the number of metrics:

$$M_{AM} = \frac{M_{Comp} + M_{Corr}}{2} \tag{5}$$

Harmonic mean (HM)

The harmonic mean is defined as the inverse of the sum of the inverses of the metric scores:

$$M_{HM} = \frac{1}{\frac{1}{M_{Comp}} + \frac{1}{M_{Corr}}} \tag{6}$$

It is used to find the true average in case of ratios when the numerators are equal as opposed to the arithmetic mean where the denominators are considered to be equal.

Area under curve (AUC)

Area under curve for two values is considered to be a simple rectangle and so the function is a simple multiplication of the metric scores:

$$M_{AUC} = M_{Comp} \times M_{Corr} \tag{7}$$

Figure 4 shows the plot for different functions with completeness metric as a constant 0.6 and correctness metric varying between $0-1$. It can be observed that the arithmetic mean (Eqn. 5) fails as it cannot aggravate the impact of small values whereas the harmonic mean (Eqn. 6) fairly succeeds to have an

Table 2: ISPRS benchmark results of the TUB1 model with state-of-the-art metrics (M_{Comp}, M_{Corr}, M_{Acc}), different formulations (M_{AM}, M_{HM}, M_{AUC}) for combining completeness and correctness metric scores and final inference metric M_{Inf} (@10cm buffer size and accuracy threshold).

Authors	M_{Comp}	M_{Corr}	M_{Acc}	M_{AM}	M_{HM}	M_{AUC}	M_{Inf}
1. Ochmann et al. [4]	0.93	0.36	1.79	0.65	0.52	0.33	32.88
2. Tran and Khoshelham [5]	0.91	0.84	5.66	0.88	0.87	0.76	72.11
3. Tran et al. [6]	0.85	0.30	1.34	0.58	0.44	0.26	25.16
4. Maset et al. [7]	0.83	0.47	1.80	0.65	0.60	0.39	38.31
5. Previtali et al. [8]	0.78	0.49	2.22	0.64	0.60	0.38	37.37

impact on small values but fails because it is not a linear function. The area under the curve (Eqn. 7) on the other hand satisfies both conditions. Table 2 shows the state-of-the-art metrics (columns M_{Comp}, M_{Corr}, M_{Acc}) and the different formulations (columns M_{AM}, M_{HM}, M_{AUC}) for the ISPRS benchmark results of the TUB1 model. It can be observed that though all the three formulations can distinguish the best performing method, M_{AUC} imposes a greater penalty on methods even when one of the metric score is lower compared to another.

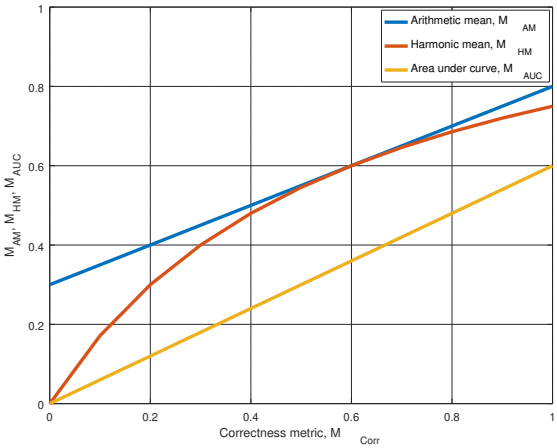


Figure 4. *Plot of different combined metrics with completeness metric as a constant 0.6 and correctness metric varying between 0-1.*

Normalized accuracy score

The accuracy score is normalized to be in range $[0,1]$ where 1 represents high accuracy and 0 represents low accuracy:

$$M_{Acc\_norm} = f(M_{Comp}, M_{Corr}) \times \frac{r - M_{Acc}}{r}, \quad (8)$$

where r is the *threshold* considered when calculating the accuracy score. The multiplication factor $f(M_{Comp}, M_{Corr})$ is used in the normalization to scale the accuracy metric as it is calculated only on the correctly reconstructed parts of the model.

Inference metric

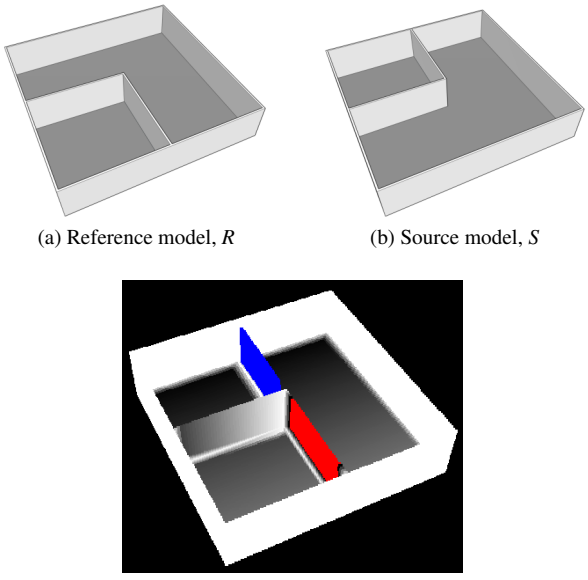
After substituting $f(M_{Comp}, M_{Corr})$ from (7) and M_{Acc_norm} from (8) in (4) we obtain the inference metric M_{Inf} as:

$$M_{Inf} = 90(M_{Comp}M_{Corr}) + 10(M_{Comp}M_{Corr}\frac{r - M_{Acc}}{r}) \quad (9)$$

It is a percentage value and is in a range $[0,100]$ where 100 represents a good model. The last column of Table 2 shows the final inference metric score M_{Inf} for the ISPRS benchmark results of the TUB1 model. It can be observed that giving a 10% weight to the normalized accuracy metric helps to penalize the reconstruction algorithm for a low accuracy score but at the same time it does not affect the main formulation.

Visualization

Though a single metric value is enough to choose between different reconstruction methods, it cannot help to pinpoint the source of error. Visualization is an important tool to localize the error regions and help to improve the reconstruction algorithm in order to get a good metric score.



(a) Reference model, R (b) Source model, S

(c) Visual representation, Q

Figure 5. *Example of error localization using visual representation.*

The inference metric cannot be directly visualized because it is just a single value, so we try to visualize the different components of the metric by generating an RGB point cloud. Figure 5 shows the point cloud representation of the model with the following color coding for each point:

- Red (incompleteness) - Points belonging to regions that are present in the ground truth model but missing in the source model.

- Blue (incorrectness) - Points belonging to regions that are not supposed to be present in the source model when compared with ground truth model.
- Gray (inaccuracy) - Points belonging to regions reconstructed within the buffer range. Gray level values close to 255 (white) are more accurate compared to values close to 0 (black) which are less accurate.

It is very evident that with such a representation it is not only easy to identify the missing or wrongly reconstructed structures but also to know the accuracy with which they are reconstructed.

Experiments

This section provides various experiments to verify the robustness of the implemented metric on both synthetic and real world data.

Synthetic data

A step-by-step hierarchy test was done with a synthetic model of a room shown in Figure 6 used as the ground truth reference model. This room was modelled with multiple components observed in real world data like doors, windows, etc.

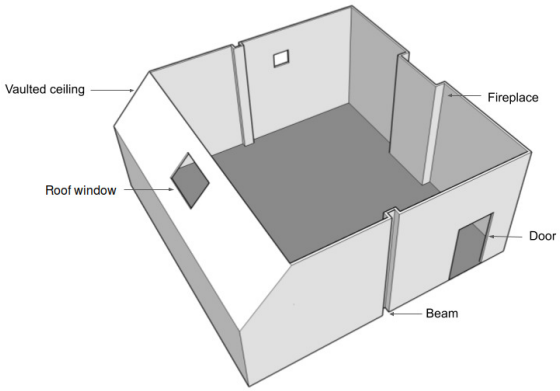


Figure 6. *Synthetic ground truth reference model used for experiments.*

Table 3: Results obtained on synthetic data (@25cm buffer size and accuracy threshold).

Source model	M_{Comp}	M_{Corr}	M_{Acc}	M_{Inf} %
1. Simple cube	0.91	0.85	0.33	77.51
2. Vaulted ceiling	0.97	0.94	0.72	92.30
3. Fireplace	0.98	0.95	0.25	93.53
4. Door	0.98	0.98	0.38	96.92
5. Roof window	0.98	0.99	0.37	98.41
6. Beam	1.00	0.99	0.37	99.89

We used six source models starting with a simple cube and increasing the complexity of the model at each step by adding the different components of the room and thereby getting closer to the reference model. The inference metric M_{Inf} was then calculated to see if they could identify the improvement of the model at every step. Table 3 lists the different components added to the source model at each step and their corresponding metric value when compared to the reference model.

Real world data

The data for these experiments were obtained by performing scans of real houses, room by room, using various 3D scanners such as Matterport, DotProduct DPI-8 and Geoslam ZEB1. Three houses were scanned with a total of 46 rooms. Figure 7 shows the input point cloud for one of the houses scanned using a DotProduct DPI-8 scanner.

Figure 7. *Input scanned point cloud obtained for House 1 ground floor.*

Table 4: Inference metric average for each house for different versions of reconstruction algorithm (@25cm buffer size and accuracy threshold).

Version	House 1	House 2	House 3
1.0	65.47	69.14	68.20
2.0	69.33	77.10	79.68
2.1	76.41	79.81	80.10
2.2	78.87	81.94	80.18
3.0	81.95	82.90	80.06
3.1	84.55	83.27	83.63

The initial reconstruction pipeline and the improvements done to the pipeline which were tested using inference metrics are explained below:

1. *Verify results of initial reconstruction pipeline:* The pipeline tested was an initial setup without any optimizations. Table 4 row v1.0 shows the average inference metrics of each house.
2. *Verify results after automated clutter removal:* The initial scan of the house contains a lot of "clutter" such as tables, sofas, beds etc. which strongly influence the reconstruction and create planes which are not part of the room structure. Table 4 rows v2.0, v2.1, v2.2 shows the average inference metrics of each house obtained after different improvements implemented to automate the removal of clutter.
3. *Improve plane detection technique:* Random sample consensus (RANSAC) based and region growing based plane detection are the two important plane detection techniques in CGAL library. The initial reconstruction pipeline used the RANSAC based technique. The region growing based technique was later used after observing that the technique generated more stable results when compared to results from

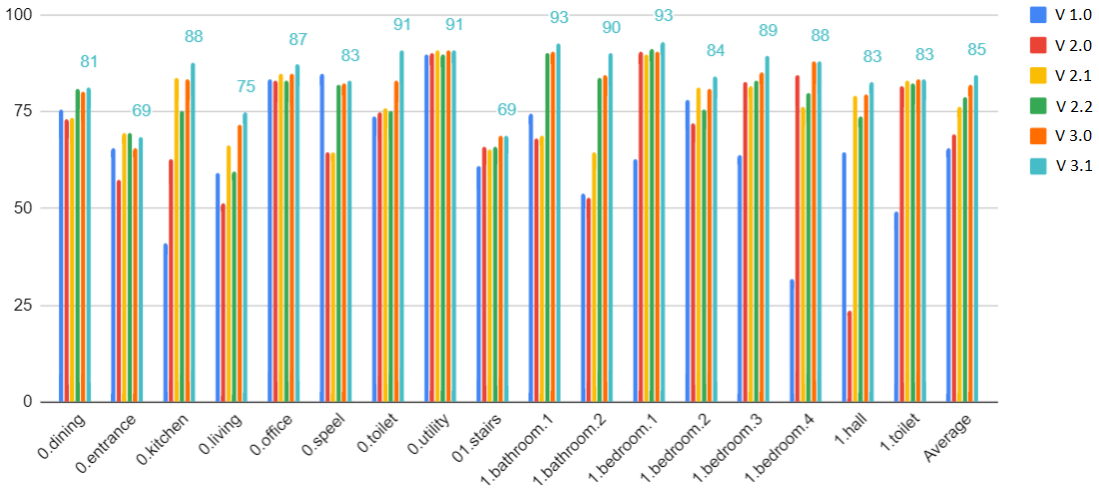
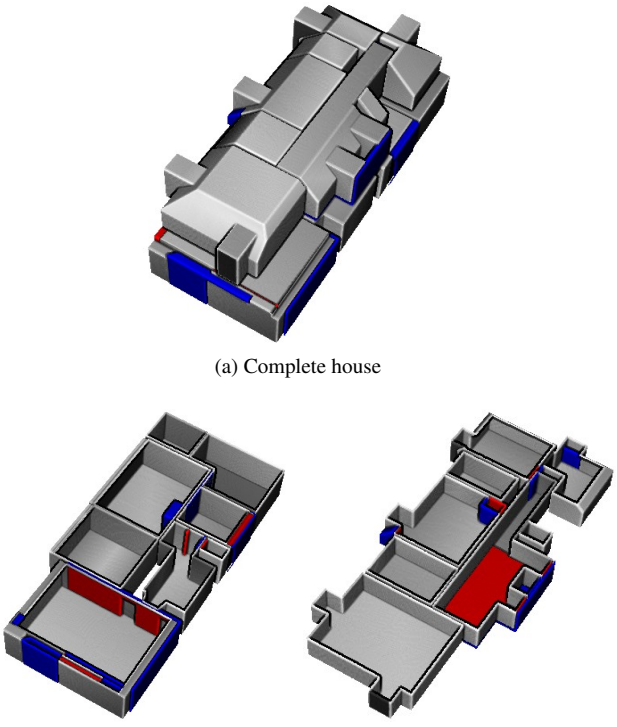


Figure 8. *Graph showing the reconstruction results of House 1 for every version release with the help of inference metric (@25cm buffer size and accuracy threshold).*

the RANSAC based technique, which were verified with the inference metric scores. Table 4 row v3.0 shows the average inference metrics of each house obtained after replacing the initial RANSAC based plane detection with the region growing based technique.

4. *Verify results after automated door detection:* Table 4 row v3.1 shows the average inference metrics of each house obtained after implementing a door detection algorithm.



(a) Complete house

(b) Ground floor cross section

(c) First floor cross section

Figure 9. *Visual representation of House 1 with 17 rooms for version 3.1.*

Figure 8 shows the inference metrics for all the rooms in *House 1*. Each color represents a different version of the reconstruction pipeline implementation. Figure 9 shows the visual representation of *House 1* for version 3.1 with the errors localized.

Conclusion

In this paper, a single metric for comparing the performance of 3D reconstruction of building interiors was presented. A corresponding visualization technique was also introduced for localizing the reconstruction error, thereby improving the algorithm by visualising the different components of the metric like incompleteness, incorrectness and inaccuracy.

The current implementation works well for geometric 3D models as it takes into account only their geometric characteristics like shape, size and angle. It was observed that structures like the ceilings were reconstructed correctly most of the time as they do not have occlusions. And since this structure is huge compared to smaller structures such as doors and windows, possible future work could examine assigning different weights to different semantic structures while calculating the intersections with the ground truth. We believe that this would make better sense when comparing semantically rich 3D models.

Acknowledgments

This work is supported by Flanders Innovation & Entrepreneurship (VLAIO) through an O&O company project (HBC.2017.0467). We thank 3Frog for providing the scanned indoor scenes with their corresponding 3D Blender models and their assistance with generating the synthetic training data.

References

[1] Pingbo Tang, Daniel Huber, Burcu Akinci, Robert Lipman, and Alan Lytle. Automatic reconstruction of as-built building information models from laser-scanned point clouds: A review of related techniques. *Automation in construction*, 19(7):829–843, 2010.

[2] Liangliang Nan and Peter Wonka. Polyfit: Polygonal surface reconstruction from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2353–2361, 2017.

[3] Inge Coudron, Steven Puttemans, and Toon Goedemé. Polygonal reconstruction of building interiors from cluttered pointclouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[4] Sebastian Ochmann, Richard Vock, and Reinhard Klein. Automatic reconstruction of fully volumetric 3d building models from oriented point clouds. *ISPRS journal of photogrammetry and remote sensing*, 151:251–262, 2019.

[5] H Tran and K Khoshelhama. A stochastic approach to automated reconstruction of 3d models of interior spaces from point clouds. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pages 299–306, 2019.

[6] H Tran, K Khoshelham, A Kealy, and L Díaz-Vilariño. Shape grammar approach to 3d modeling of indoor environments using point clouds. *Journal of Computing in Civil Engineering*, 33(1):04018055, 2018.

[7] E Maset, L Magri, and A Fusiello. Improving automatic reconstruction of interior walls from point cloud data. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2019.

[8] Mattia Previtali, Lucía Díaz-Vilariño, and Marco Scaioni. Indoor building reconstruction from occluded point clouds using graph-cut and ray-tracing. *Applied Sciences*, 8(9):1529, 2018.

[9] Martin Rutzinger, Franz Rottensteiner, and Norbert Pfeifer. A comparison of evaluation techniques for building extraction from airborne laser scanning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2(1):11–20, 2009.

[10] Mohammad Awrangjeb and Clive S Fraser. An automatic and threshold-free performance evaluation system for building extraction techniques from airborne lidar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(10):4184–4198, 2014.

[11] K Khoshelham, H Tran, L Díaz-Vilariño, M Peter, Z Kang, and D Acharya. An evaluation framework for benchmarking indoor modelling methods. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42(4), 2018.

[12] H Tran, K Khoshelham, and A Kealy. Geometric comparison and quality evaluation of 3d models of indoor environments. *ISPRS journal of photogrammetry and remote sensing*, 149:29–39, 2019.

[13] Kourosh Khoshelham, L Díaz Vilariño, Michael Peter, Zhizhong Kang, and Debaditya Acharya. The isprs benchmark on indoor modelling. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, 42, 2017.

Author Biography

Umamaheswaran Raman Kumar is currently a Ph.D. student at the EAVISE research group in KU Leuven, Belgium. He obtained his Erasmus Mundus Joint Master Degree in Computer Vision and Medical Imaging from University of Girona, Spain in 2018. His research interest includes 3D vision and modeling, augmented reality, machine learning and software engineering.

Inge Coudron received her M.Sc. degree in Electrical Engineering from KU Leuven, Belgium in 2013. After obtaining her engineering degree, she studied a Master after Master in Artificial Intelligence. Wanting to put this knowledge into practice, she pursued a PhD at the faculty of Engineering Technology from KU Leuven, Belgium. There she started working on several 3D related research projects including 3D object detection and semantic 3D modelling of building interiors.

Steven Puttemans is currently a scientific advisor on innovation support for Flanders Innovation Entrepreneurship (VLAIO), Belgium. He obtained his doctoral degree in Engineering Technology from KU Leuven, Belgium in 2017. His research focused on industrially relevant applications of 2D and 3D object detection, with a main focus on integrating application specific knowledge into the solution.

Patrick Vandewalle received a M.Sc. degree in electrical engineering from KU Leuven, Belgium, in 2001, and a Ph.D. degree from EPFL, Switzerland, in 2006. From 2007 to 2018, he worked at Philips Research, The Netherlands, as a senior research scientist. He is now an associate professor at KU Leuven, Belgium. His current research in the EAVISE research group focuses on 3D processing, reconstruction, computer vision and AR/VR.

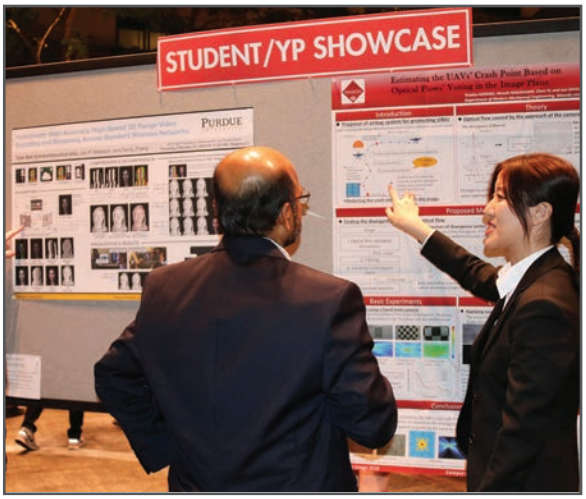
JOIN US AT THE NEXT EI!

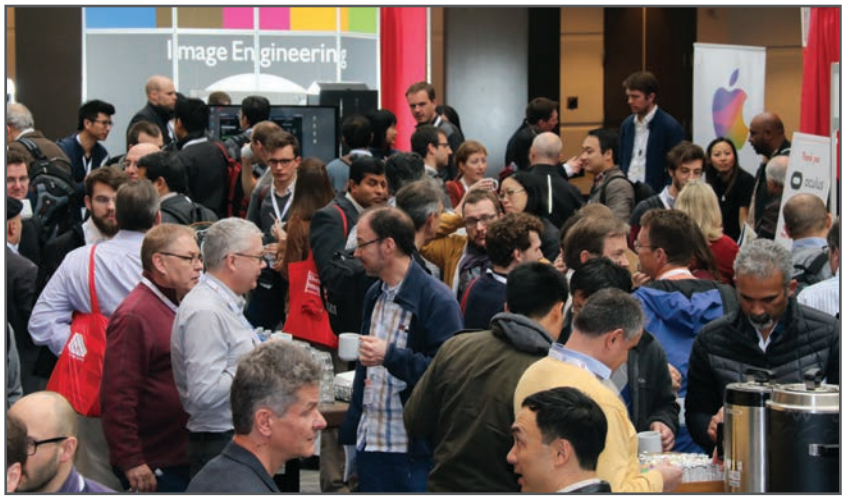
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org