# Enhanced 3D Mapping with an RGB-D Sensor via Integration of Depth Measurements and Image Sequences

**Bo Wu, Xuming Ge, Linfu Xie, and Wu Chen**

## Abstract

*State-of-the-art visual simultaneous localization and mapping (SLAM) techniques greatly facilitate three-dimensional (3D) mapping and modeling with the use of low-cost red-green-blue-depth (RGB-D) sensors. However, the effective range of such sensors is limited due to the working range of the infrared (IR) camera, which provides depth information, and thus the practicability of such sensors in 3D mapping and modeling is limited. To address this limitation, we present a novel solution for enhanced 3D mapping using a low-cost RGB-D sensor. We carry out state-of-the-art visual SLAM to obtain 3D point clouds within the mapping range of the RGB-D sensor and implement an improved structure-from-motion (SfM) on the collected RGB image sequences with additional constraints from the depth information to produce image-based 3D point clouds. We then develop a feature-based scale-adaptive registration to merge the gained point clouds to further generate enhanced and extended 3D mapping results. We use two challenging test sites to examine the proposed method. At these two sites, the coverage of both generated 3D models increases by more than 50% with the proposed solution. Moreover, the proposed solution achieves a geometric accuracy of about 1% in a measurement range of about 20 m. These positive experimental results not only demonstrate the feasibility and practicality of the proposed solution but also its potential.*

## Introduction

Red-green-blue-depth (RGB-D) sensors (such as the Kinect or the Structure sensors) have remarkable advantages, such as mobility and low cost, and have been used extensively in three-dimensional (3D) mapping and visual simultaneous localization and mapping (SLAM) (Dryanovski *et al.* 2013). An RGB-D sensor captures RGB images by a RGB camera and pixel-wise depth information by an infrared (IR) camera together with an IR projector, and it can thus produce textured 3D point clouds in the object coordinate system via transformation from the RGB camera coordinate system to the IR camera coordinate system (Tang *et al.* 2016). RGB-D sensors have thus been used for 3D mapping and modeling both in indoor and outdoor environments in recent years (Henry *et al.* 2010; Kerl *et al.* 2013; Tang *et al.* 2016). Although RGB-D sensors can be used to build 3D models of unprecedented richness, they have drawbacks that limit their application in practice: they measure depth only up to a limited distance (typically less than 5 m), and the gained depth values are much noisier than those provided by traditional laser scanners. Moreover, the field of view of a depth camera is far more constrained than that of traditional laser scanners that are typically used for 3D mapping and modeling.

The above-mentioned drawbacks of RGB-D sensors narrow the scope of their application in 3D mapping and modeling (Ye and Wu 2018). For example, when it is impossible to get close enough to scan targets in the working environment, an RGB-D sensor cannot provide desirable results. Other solutions are therefore needed to improve the feasibility and practicality of such sensors. This paper introduces the structure-from-motion (SfM) and multiview stereo (MVS) methods to generate extended 3D models in relatively distant ranges using RGB image sequences to supplement the 3D models obtained from the RGB-D sensors at short range. The proposed solution has three main advantages. (1) It requires no additional effort; although the SfM and MVS are offline productions, the RGB image sequences can be simultaneously collected with the online SLAM. (2) It provides improved SfM constrained by the additional depth information. (3) It uses scale-adaptive registration to fuse the multisensor point clouds.

The remainder of this paper is organized as follows. In the section "Related Work", the paper briefly introduces works related to the RGB-D SLAM and SfM. In the section "Enhanced 3D Mapping by Integrating Depth Measurements and Image Sequences", the proposed solution for enhanced 3D mapping is described in detail. In the "Experimental Evaluation" section, two challenging cases are tested to demonstrate the positive properties of the proposed system. In the final section, concluding remarks are presented.

## Related Work

The advent of RGB-D sensors has led to a great deal of progress in SLAM. Typically, a visual SLAM system consists of a camera-tracking frontend that uses visual odometry (Engel *et al.* 2014; Kerl *et al.* 2013) and a backend that generates and maintains a map of key-frames and reduce global drift via loop closure detection and map optimization (Mur-Artal and Tardós 2017; Gao *et al.* 2018). Most state-of-the-art methods estimate six degrees of freedom between adjacent frames based on the sparsely selected visual features to represent the camera motion (Liu *et al.* 2018). The two main classical approaches are to minimize the photometric error between consecutive stereo pairs (Comport *et al.* 2007) and to minimize the geometric error between 3D points (e.g., the iterative closest point [ICP] (Besl and McKay 1992) and the three-dimensional normal-distributions transform (Magnusson 2009). Tykkälä *et al.* (2011) and Whelan *et al.* (2013) used the minimization of photometric and geometric error to estimate camera motion. Newcombe *et al.* (2011) proposed an incremental strategy to register RGB-D frames. In recent years, visual-inertial SLAM has been increasingly used (Hesch *et al.* 2014) and these

The Hong Kong Polytechnic University, Department of Land Surveying & Geo-Informatics (xuming.ge@polyu.edu.hk).

Photogrammetric Engineering & Remote Sensing
Vol. 85, No. 9, September 2019, pp. 633–642.
0099-1112/19/633–642
© 2019 American Society for Photogrammetry and Remote Sensing
doi: 10.14358/PERS.85.9.633

approaches commonly rely on either pose-graph optimization (Kümmerle *et al.* 2011) or bundle adjustment (Lourakis and Argyros 2009) to minimize reprojection errors across frames. Dai *et al.* (2017) proposed a method called BundleFusion, in which a bundle adjustment (BA) from local to global was carried out and a sparse-to-dense alignment strategy was implemented to generate a dense 3D model from a coarse one. However, the previously mentioned SLAM developments have been most frequently considered in computer vision applications, and few studies have applied the RGB-D SLAM to 3D mapping applications, mainly because of their limited measuring scopes and lower data quality. Steinbrucker *et al.* (2013) proposed a large-scale multiresolution method to generate indoor mapping reconstruction from RGB-D sequences. Chow *et al.* (2014) introduced a mapping system that integrated a mobile 3D light detection and ranging system with two Kinect sensors and an inertial measurement unit to map indoor environments. Tang *et al.* (2016) carried out precise calibration of the RGB-D sensor to achieve enhanced 3D mapping. However, these improvements focused on data quality and were not related to measuring scope cases. Obviously, to improve the feasibility and practicality of low-cost RGB-D sensors, the measuring scope should be given greater attention. Our work fills this gap.

The SfM method is an image-based algorithm used to estimate the camera poses, scene geometry, and orientation from RGB sequences (Snavely *et al.* 2008; Westoby *et al.* 2012). In SfM, these estimations are solved simultaneously using a highly redundant iterative BA procedure based on a set of features that are automatically extracted from a set of multiple overlapping images (Snavely 2008). Because SfM does not rely on depth information from sensors, the valid range is typically greater than that of RGB-D sensors. A variety of SfM strategies have been proposed, including incremental (Frahm *et al.* 2010), hierarchical (Gherardi *et al.* 2010), and global approaches (Crandall *et al.* 2011). Moreover, many well-known open-source programs, such as VisualSfM (Wu 2011), Bundle (Snavely *et al.* 2006), and COLMAP (Schonberger and Frahm 2016), can be used to implement SfM and MVS. However, in addition to requiring the input images to have extreme overlapping regions in the stereo pairs, SfM must contain enriched textures in these regions to ensure the quality of the geometry structures. Moreover, a 3D visual model from SfM is scale-free and cannot be used directly on mapping issues.

Recent advances in deep learning have provided promising results for resolution of the related issues. Such innovations include the SE3-Nets (Byravan and Fox 2017), 3D image interpreter (Wu *et al.* 2016), depth convolutional neural network (CNN) (Garg *et al.* 2016), and SfM-Nets (Vijayanarasimhan *et al.* 2017). Although the robustness of SfM and the quality of the results have been improved by those state-of-the-art approaches, a gap remains to obtaining further mapping properties. To carry out a deep learning method, we require a large number of samples for training and a highly configured device (LeCun *et al.* 2015). To account for the properties of offline computing, we introduce additional depth information from the SLAM results into the BA to improve SfM. We then carry out a scale-adaptive registration to merge the point clouds from the depth sensor in short ranges and from the RGB image sequences in distant ranges to generate enhanced and extended 3D models.

## Enhanced 3D Mapping by Integrating Depth Measurements and Image Sequences

### Overview of the Proposed Approach

A calibration process must be carried out on the RGB-D sensor before starting the 3D mapping tasks. We implemented a precise calibration (Tang *et al.* 2016) on the RGB-D sensor to determine the precise spatial relationship between the RGB and IR (depth) cameras. We carried out the state-of-the-art online SLAM method—i.e., BundleFusion (Dai *et al.* 2017)—to generate dense 3D mapping results in short ranges using a commodity depth sensor. We then introduced the information from the depth sensor (i.e., point-to-pixel) into the SfM system (i.e., COLMAP) (Schonberger and Frahm 2016) to carry out constrained BA. Although the scale is imported in the improved BA within the additional depth constraints, distortions may occur during the SfM and MVS. Therefore, we further implemented scale-adaptive registration on those two-point clouds to generate enhanced and extended 3D mapping results. Figure 1 shows an overview of the proposed approach.

### 3D Mapping from the RGB-D Sensor in Short Ranges

The BundleFusion (Dai *et al.* 2017) was used for SLAM and 3D mapping in short ranges in our approach. The main difference between the BundleFusion and other approaches is that the former is a fully parallelizable sparse-to-dense global pose optimization framework. Sparse RGB features are detected by the scale-invariant feature transform (SIFT) detector (Lo and Siebert 2009; Wu *et al.* 2012), and then sparse correspondences are carefully established between the input frames, and then used to carry out a coarse global alignment. Mismatches are detected and removed to avoid false loop closures between all input frames. That is, detected keypoints are matched against all previous frames and carefully filtered to remove mismatches. After that, the coarse alignment is refined by optimizing for dense photometric and geometric consistency. We implement a hierarchical local-to-global strategy to optimize the camera tracking and orientations. On the first hierarchy level, $n$ continuous frames (e.g., $n = 10$) compose a chunk, and in each chunk a keyframe is defined and then other frames in the chunk are matched to the keyframe. Based on this, the program executes a local BA. On the second hierarchy level, all of the chunks are collected and the algorithm implements a global BA. The coarse global pose optimization ensures that the subsequent fine alignments can converge to a promising solution. In the fine global optimization step, the program also implements the same hierarchical strategy as that used in the coarse step. After executing the SLAM program, a dense 3D mapping can be obtained.

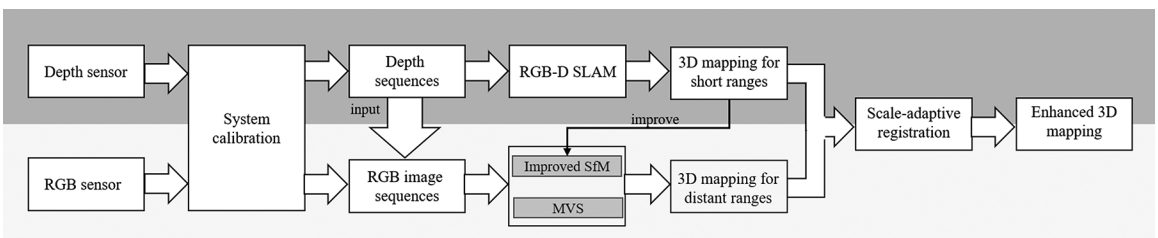


Figure 1. Workflow of the proposed approach.

As mentioned, the low-cost RGB-D sensor is inferior to the terrestrial laser scanner in terms of data quality. Thus, a noise and outlier removal strategy (Wolff *et al.* 2016) should be applied to the obtained point cloud model to allow the 3D mapping results (e.g., point clouds) to be generated and oriented simultaneously. With the above method, the mapping range is expected to reach 5–10 m in front of the sensor with typical off-the-shelf RGB-D sensors (Tang *et al.* 2017).

### 3D Mapping from RGB Images in Distant Ranges

RGB sequences are collected simultaneously when carrying out SLAM. Our work is based on an incremental SfM strategy and additional promising input constraints from the SLAM results. Features are extracted from the input images by applying a common feature detector, such as SIFT (Lo and Siebert 2009), and binary features (Heinly *et al.* 2012). Correspondences can then be set up among the input images in the overlaps based on the obtained features (Havlena and Schindler 2014; Johnson and Zhang 2014). A variety of approaches have been used to address this problem, such as Vocmatch. After obtaining the correspondences, the algorithm completes a geometric validation process to remove outliers using random sample consensus (RANSAC) (Bolles and Fischler 1981) to improve the robustness of the obtained correspondences. Moreover, as mentioned before, the corresponding depth information is introduced into this offline process. Thus, the point-to-pixel constraint (see Figure 2) can be applied as a filter to further purify the obtained correspondences. The algorithm provides a higher weight to correspondences with corrected depth information. Starting from a reconstruction, new images can be registered to the current model by solving the Perspective-n-Point problem (Bolles and Fischler 1981) using the gained correspondences. An efficient and robust multiview triangulation method in COLMAP is implemented in our algorithm to complete the transformation from two-dimensional to 3D. Image registration and triangulation are separate procedures even though their products are highly correlated. Thus, bundle adjustment is carried out to solve a nonlinear problem and to refine the pose-graph and orientations by minimizing the re-projection error

$$E=\sum_{\mathrm{j}}\mu_j\left(\left\|\pi\left(X_j,P_k\right)-p_j\right\|_2^2\right), \tag{1}$$

where $\pi$ is a function to transform a 3D point to image space, $\mu_j$ is a weight factor, $X_j$ is the camera pose, $p_j\in\mathbb{R}^2$ is a pixel coordinate in the image space, and $P_k\in\mathbb{R}^3$ is the corresponding 3D position in the object space. Based on the point-to-pixel information, we add the following additional constraint:

$$\alpha\left\|P_k-P_j\right\|_2^2=\left\|x_k-x_j\right\|_2^2, \tag{2}$$

where $\alpha$ is a scale factor, $x_k$ is a 3D coordinate from the point-to-pixel, and $P_k$ is its corresponding 3D coordinate from the triangulation. As Schonberger and Frahm (2016) mentioned, the additional steps of retriangulation and global BA can obviously improve the SfM results because BA is severely affected by outliers and can be subsequently filtered.

After obtaining a sparse 3D model by implementing the above process, we carry out an MVS process using COMLAP to obtain the dense 3D image-based model. It should be noted that although the RGB and IR cameras were calibrated before the mapping task and the depth information can be obtained in close ranges, SfM is a necessary step and cannot be skipped before MVS. This is because the SfM provides sparse point clouds to carry out MVS in distant ranges beyond the working ranges of the depth sensor, and the global optimization of SfM can further provide more accurate geometric solutions.

### Scale-Adaptive Registration of Point Clouds

Although the depth information from SLAM is introduced into SfM to generate a sparse 3D model, the scale may not be precise enough between the two-point clouds due to data noises and limited number of points used. Thus, we need to recover the scale of the image-based model, register it in the SLAM model, and then further produce enhanced 3D mapping results.

Point cloud registration is a key step in generating a complete 3D model with a straightforward solution (i.e., ICP) (Besl and McKay 1992). However, in our case, we cannot directly use ICP or even Scaled-ICP (Du *et al.* 2007) to register the point clouds because there are no initial values to start ICP or other fine registration strategies (Ge and Wunderlich 2016). Specifically, although some SLAM points are introduced into the SfM solution to provide the scale estimation and an initial alignment, due to the data noise and limited distribution of those points, there is no guarantee that the obtained alignment has favorable quality. Therefore, in order to ensure the reliability of the proposed approach, a scale-adaptive coarse registration strategy is employed to provide more accurate transformation parameters between the two-point clouds. After that, a Scaled-ICP is implemented for accurate coregistration of them. We carry out an extended feature-based four-point congruent sets (4PCS) method (Ge 2017; Aiger *et al.* 2008) to minimize the geometric errors

$$E_4=\sum_{i=1}^{4}\left(\left\|s_i-\mathrm{V}\cdot t_i\right\|^2\right), \tag{3}$$

where V is a transformation parameter set that includes the rotation, translation, and scale and ($s_i$, $t_i$) represents a correspondence from two-point clouds. To validate the transformation parameter set obtained by Equation 3, a validation

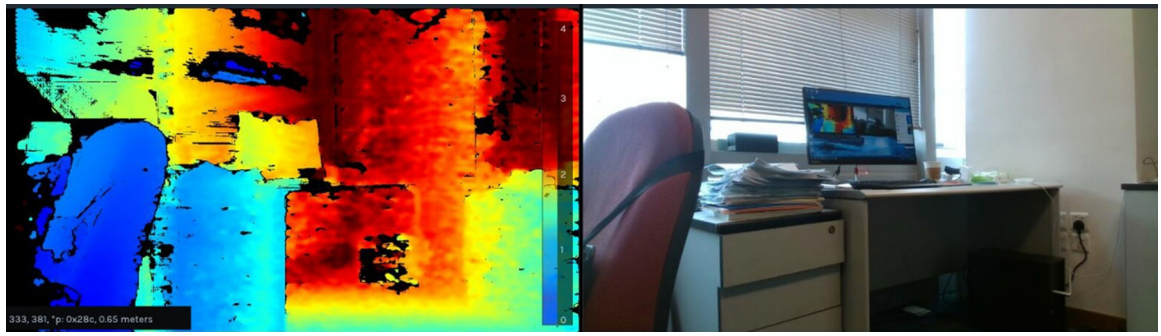


Figure 2. An example of the depth image (left) and the corresponding RGB image (right).

process as formulated by Equation 4 is carried out by using sample points from the two-point clouds. The sample rate depends on the overlapping rate between them.

$$E_n = \sum_{i=1}^{n} \left( \left\| S(s_i) - V \cdot T(t_i) \right\|^2 \right), \tag{4}$$

where S( ) and T( ) are overlapping subsets from two-point clouds.

The 4PCS strategy is an iterative solution and validation process. Based on the RANSAC platform, 4PCS randomly selects a four-point-base from a candidate set and then to set up correspondences in an inquiry set iteratively based on the geometric constraints and validation criteria. The geometric constraints in the original 4PCS are modified to satisfy our cases (see Figure 3). Letting $\mathcal{T} = \{\boldsymbol{a, b, c, d}\}$ be four coplanar points selected from a SLAM model, if the four points are not all collinear, the line ***ab*** intersects the line ***cd*** at an intermediate point ***e***. A corresponding base from a SfM + MVS model is $\mathcal{TT} = \{\boldsymbol{aa, bb, cc, dd}\}$ with an intermediate point ***ee***. The invariants and scale constrain the search for reasonable four-point bases (see Equations 5–9).

$$r_1 = \frac{\|a-e\|}{\|a-b\|} \approx \frac{\|aa-ee\|}{\|aa-bb\|}; \tag{5}$$

$$r_2 = \frac{\|c-e\|}{\|c-d\|} \approx \frac{\|cc-ee\|}{\|cc-dd\|}; \tag{6}$$

$$\left| \|a-b\| - \alpha \|aa-bb\| \right| \le \varepsilon; \tag{7}$$

$$\left| \|c-d\| - \alpha \|cc-dd\| \right| \le \varepsilon; \tag{8}$$

$$\left| cos(\langle ab, cd \rangle) - cos(\langle aabb, ccdd \rangle) \right| \le \omega, \tag{9}$$

where $r_1$ and $r_2$ are invariant ratios, $\alpha$ is the scale factor, $\varepsilon$ and $\omega$ represent the given thresholds in the distance and angle measurements, respectively, and $\langle\rangle$ is an operator to calculate the acute angle from crossed lines. Equations 5 and 6 represent the rule that intersection ratios of the diagonals in an arbitrary planar quadrangle are invariant under affine transformation (Huttenlocher 1991). Moreover, Equations 7–-9 strengthen geometric constraints for the correspondences in both the distance and angle measurements. Figure 4 shows an example of a setup of a four-point set in the SLAM dataset (left), and one of its corresponding sets in the SfM + MVS dataset (right).

## Experimental Evaluation

### Description of the Sensor and the Testing Sites

The performance of the proposed system is evaluated with two datasets captured from different scenes. We use a Tango system to capture data (see Figure 5). Smartphones such as Tango have built-in RGB and depth sensors, and the maximum working distance of the depth sensor is about 4 m. In our cases, the test scenarios are larger than the traditional working scenes of the RGB-D sensors (that is, indoor applications). The first dataset was captured along a corridor at a large public rest area (see Figure 6), and the subsequent test was carried out in a subway station (see Figure 7). In our experiments, we captured most of the data in the scenes from along the wall. Thus, the trajectory of the data collection does not cover the entire scene, and the middle parts of the scenes are obviously out of the working range of the depth sensor. Although these two experiments are specially designed to test and prove the proposed system, they can also be used to generate an enhanced 3D model for mapping issues when the trajectories cannot cover the whole scene or to build a large 3D model for a more economical approach.

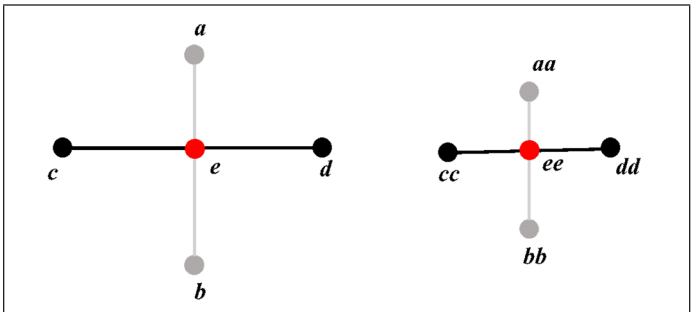


Figure 3. Illustration of the structure of the corresponding four-point set.

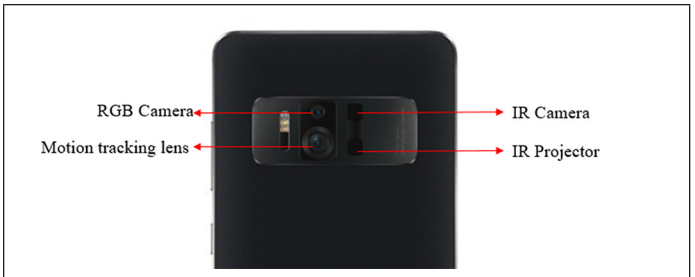


Figure 5. Hardware scheme of the RGB-D sensor.

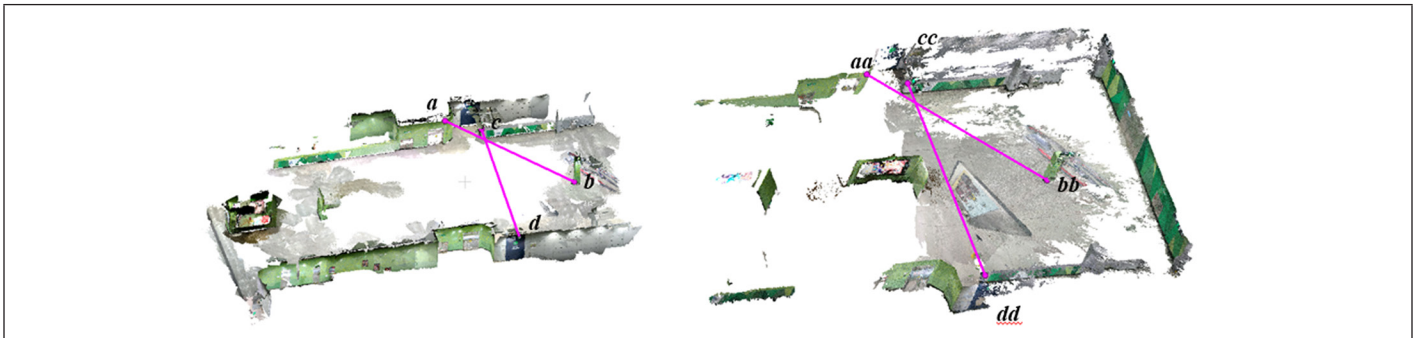


Figure 4. Selection of the four-point sets in the SLAM model (left) and the SfM + MVS model (right).

Figure 6. Corridor test site.

Figure 7. Subway station test site.

### Corridor Test Site

The important structural distances were measured by a laser range finder that was accurate to within 1 mm for distance measurements; these distances are used as the geometric structural ground truth in our tasks for mapping applications. Figure 8 shows the footprint of the corridor dataset within the geometric structural ground truth measurements. To assess 3D point cloud accuracy, we used a terrestrial laser scanner (Leica BLK 360) to capture the corresponding point clouds. We manually registered those point clouds to act as the ground truth and then implemented the 3D comparisons. For Leica BLK 360, the point measurement rate can reach 360 000 points per second with 3D point accuracy within 8 mm, which is sufficient to assess the gained point clouds from a low-cost RGB-D sensor.

As mentioned, we captured the data along the wall. The red curve in Figure 8 represents the trajectory of SLAM, and the yellow areas are the regions outside the working scope of the depth sensor. As Figure 8 shows, we designed some small loops in the trajectory to capture more RGB image sequences from the yellow areas. Figure 9a shows the SLAM point cloud model in this case, and Figure 9b presents the integrated model (i.e., the image-based model registered to the SLAM

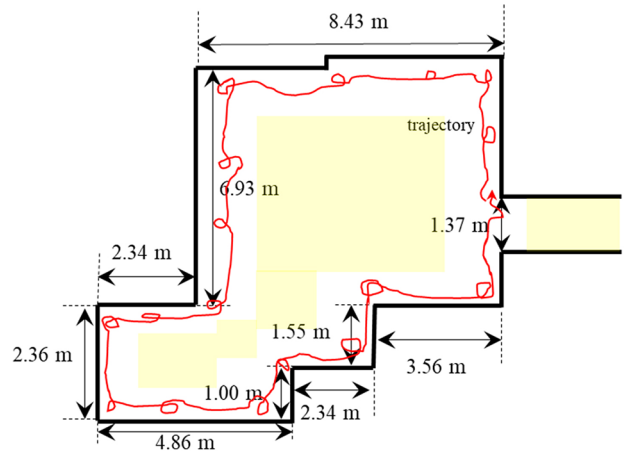


Figure 8. Floor plan of the corridor within the reference geometric information and the strategies of the designed experiment. The red line is the SLAM trajectory, and the yellow areas are outside the depth camera working range.

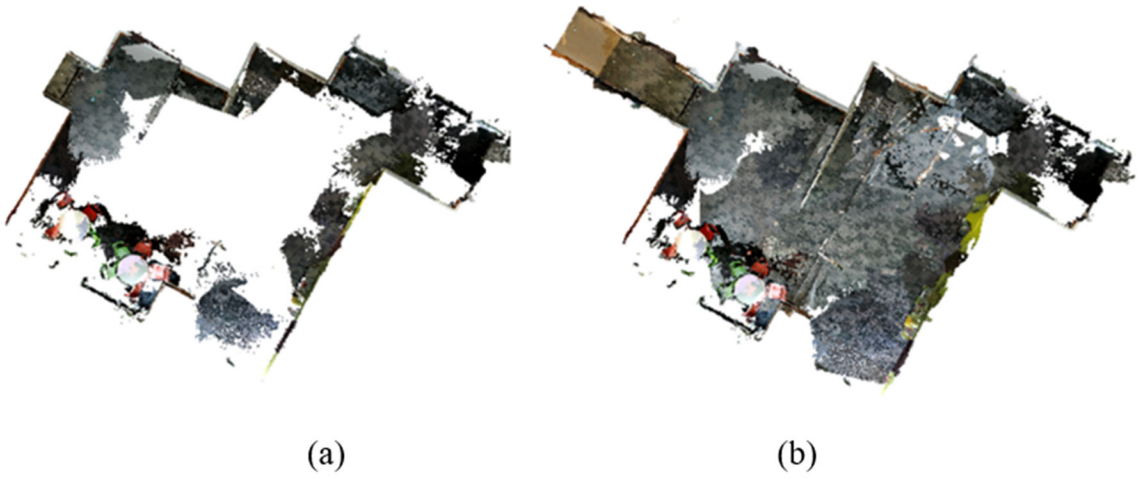


Figure 9. 3D mapping results. (a) SLAM point cloud model; (b) Registered point cloud model; that is, the SLAM model + the imaged-based model.

model). As Figure 9 shows, the image-based model can clearly compensate for the empty regions of the SLAM model, and the model coverage rate increases by approximately 56%.

Moreover, with the additional depth information to constrain the SfM process, the quality of the image-based point clouds is improved. Figure 10 shows the experimental results of the obtained point clouds from the SfM + MVS with/without the additional depth information. Comparing the point clouds in the two red squares it can be seen that the additional depth information can improve the geometric structure of the SfM + MVS solution. Although the improvement in the distant ranges (e.g., the blue square in Figure 10b) is not significant as that in the short ranges since the depth information from the IR camera is limited, the global geometric structure in the SfM + MVS solution benefits from the additional constraints.

Furthermore, favors from the designed scale-adaptive registration of the image-based point clouds can recover the scale information and ensure the correct geometric structure. Figure 11 displays the biases in the terms of footprint between the ground truth (black lines) of the geometric structure (i.e., the distances in Figure 8) and the generated enhanced and extended 3D mapping results (blue dashed lines). As Figure 11 shows, the obtained 3D mapping results have the correct geometric structure at both short and long ranges. The maximum difference is 6.9 cm at the whole loop closure point. One explanation for this situation is that there are insufficient feature points to support the SLAM to implement a reliable loop closure optimization. Thus, more overlaps in the trajectory and revisiting the starting location can benefit the SLAM solution. More information about the biases between the 3D mapping results and the geometric structural ground truth are presented in Table 1 and in Figures 11 and 12. Considering the accuracy requirements for the terrestrial laser scanner cases, 1 cm mapping accuracy is reasonable for the low-cost sensor (Ge and Wunderlich 2015). Figure 12 shows the 3D comparison in terms of cloud-to-cloud distances between the gained 3D model and the corresponding benchmark model. We can see that the gained 3D model has good quality, with point bias < 10 cm in the major areas. The maximum point bias is 0.2 m, mostly at boundary points (see Figure 12). Table 1 also displays the summary information of the 3D comparison, including the maximum bias, the minimum bias, the average bias, and the root -mean -square error (RMSE).

Table 1. Assessment of the corridor dataset.

| | Bias (cm) | | | |
|---|---|---|---|---|
| **Metric** | **Max.** | **Min.** | **Ave.** | **RMSE** |
| Geometric accuracy of structural edges | 6.9 | 0.7 | 4.1 | 2.2 |
| Comparison of point clouds | 19.3 | 0.0 | 8.4 | 11.2 |

## Subway Station Site

The subway station case (Figure 7) shows a working area of about 20 m × 40 m. We also introduced the laser range finder and the Leica BLK 360 scanner to produce the geometric structural ground truth and the point clouds ground truth to assess the gained enhanced and extended 3D mapping results. Figure 13 displays a footprint of this area with the geometric structural ground truth information. The red curve represents the trajectory of SLAM, and the yellow areas represent the regions outside the depth sensor's working range. The specially designed trajectory for this case also contains some loops (see Figure 13) for the same purposes as those in the first case. Figure 14a shows the 3D point cloud from the SLAM process and Figure 14b shows the image-based SfM model. Figure 14c displays the enhanced and extended 3D point clouds from the proposed approach. In the image-based model (i.e., Figure

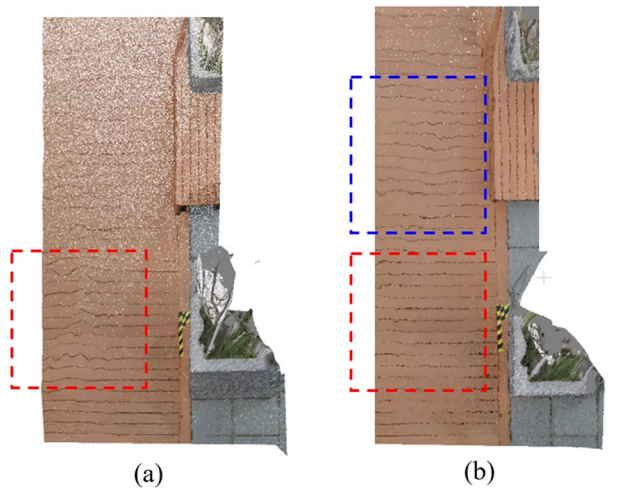


Figure 10. The textured point clouds generated from the SfM + MVS solution, (a) without the additional depth information, and (b) with the additional depth information.

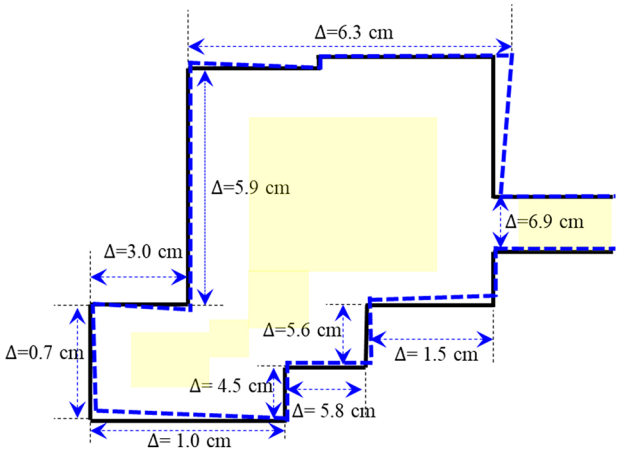


Figure 11. Biases of the enhanced 3D mapping results compared with the geometric ground truth in terms of the footprint. Black lines represent the ground truth, and blue dashed lines express the corresponding ground truth from the enhanced 3D mapping results.

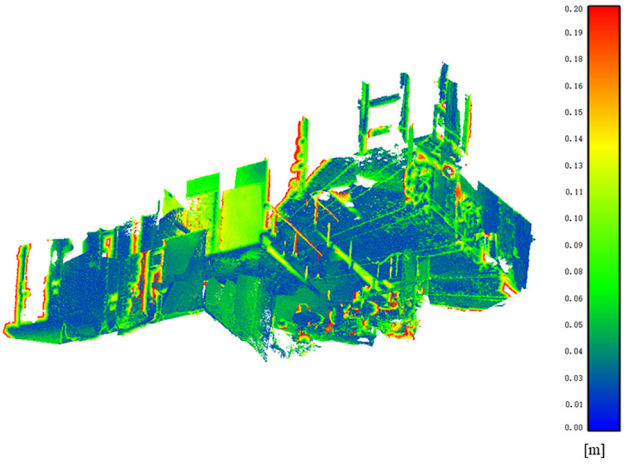


Figure 12. 3D comparison in the corridor dataset between the obtained 3D point clouds and the ground truth point clouds collected with a Leica BLK 360 scanner.

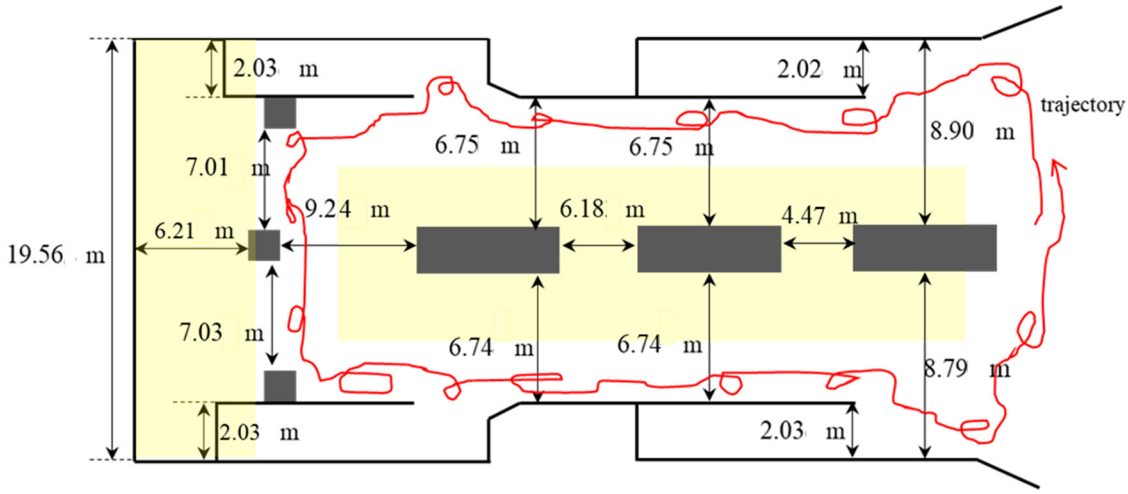


Figure 13. Floor plan of the subway station within the reference geometric information and the strategies of the designed experiment.

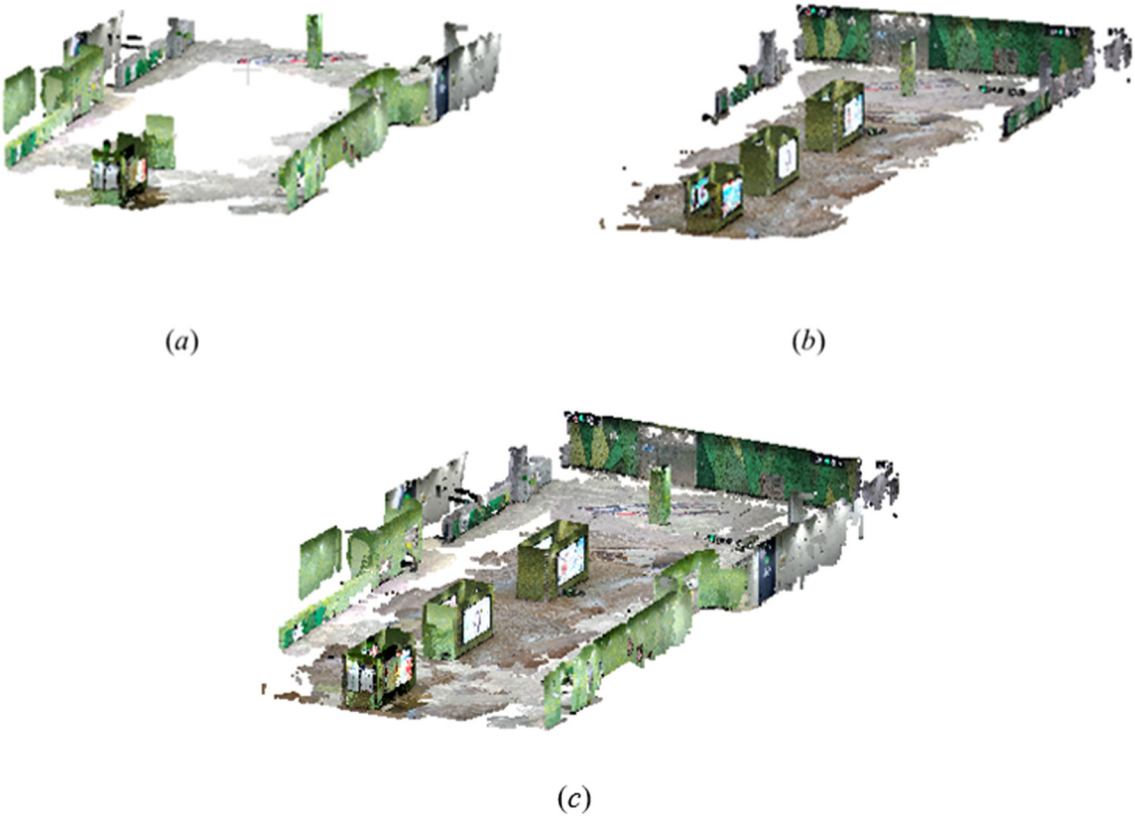


Figure 14. The textured 3D point clouds generated by (a) SLAM, (b) image-based SfM, and (c) the proposed integrated approach.

14b) the walls on the left and right were not detected by SfM because of the lack of texture information on those walls. It is obvious that the image-based point cloud model can be used to compensate the 3D SLAM point cloud model by the proposed solution so as to obtain an enhanced and extended 3D mapping results. In this case, the growth rate of model coverage is over 50%. Figure 15 shows the biases between the geometric structural ground truth (black lines) and the corresponding enhanced 3D mapping (blue dashed lines) results. The maximum bias increases to 22.8 cm at the longest tested distance (about 20 m) (see Figure 15 and Table 2). In this case, the accuracy of the gained enhanced 3D mapping results is lower than that in the first case (see Tables 1 and 2 and Figures 12 and 17). One interpretation is that the scanning scope of this area is much larger (about 800 $m^2$) than the first case, making it challenging for RGB-D SLAM to maintain high accuracy. We show an object's details from the enhanced 3D mapping results in Figure 16 to show that the proposed solution has good performance in recovering the geometric structures. We also assess the quality of the enhanced 3D point clouds by implementing the 3D comparison with the benchmark point clouds in terms of cloud-to-cloud distances (see Figure 17). First, the geometric structure of the main body is correct. The biases of the 3D point clouds are mostly less than 10 cm in

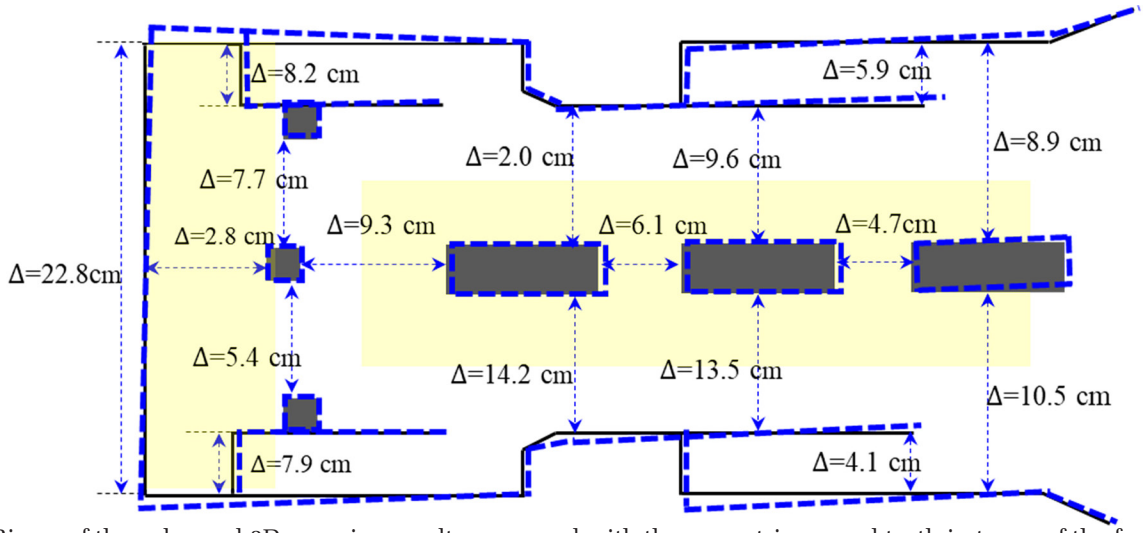


Figure 15. Biases of the enhanced 3D mapping results compared with the geometric ground truth in terms of the footprint. Markings have the same meanings as those in the corridor dataset.

the major surfaces. The average bias of the 3D point clouds is 9.3 cm (see Table 2). The poor results concentrated on the right part of the ground points because in that location, the terrestrial laser scanner was installed at ground level, so the point density and the quality of the laser points on the ground were poor because of the large incidence angles of the laser ray (Ge 2016).

Table 2. Assessment of the subway station dataset.

| | Bias (cm) | | | |
|---|---|---|---|---|
| **Metric** | **Max.** | **Min.** | **Ave.** | **RMSE** |
| Geometric accuracy of structural edges | 22.8 | 2.0 | 8.4 | 4.8 |
| Comparison of point clouds | 28.9 | 0.0 | 9.3 | 11.2 |

## Conclusions and Discussion

In this paper, we presented a novel solution, integrating a SLAM point cloud model, and a SfM + MVS point cloud model (i.e., image-based model) to generate enhanced and extended 3D mapping results using a low-cost RGB-D sensor. In the

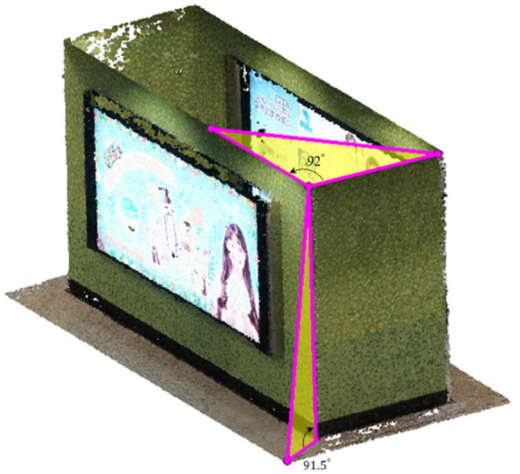


Figure 16. Object's details from the merged enhanced 3D mapping results.

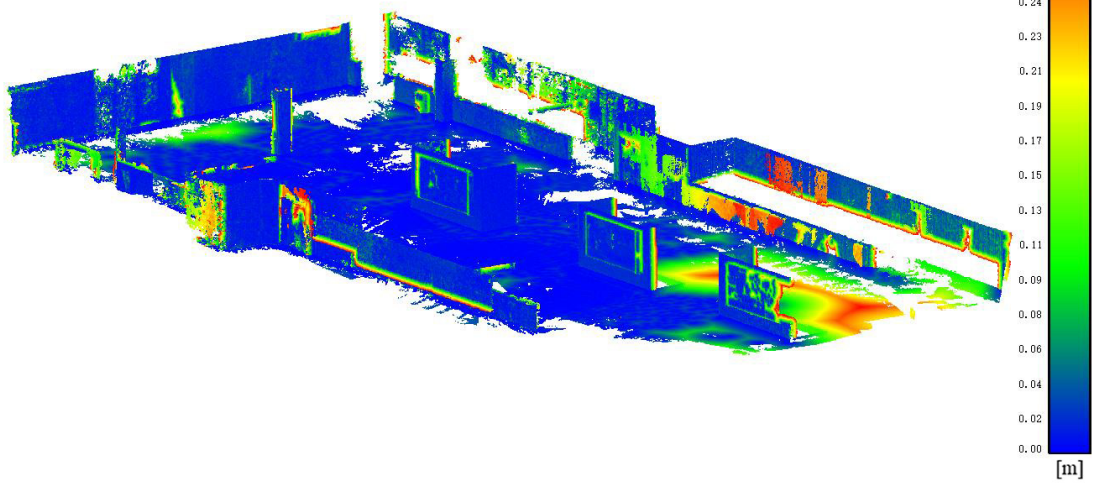


Figure 17. 3D comparison of the subway station dataset between the obtained 3D point clouds and the ground truth point clouds collected with a Leica BLK 360 scanner.

proposed approach, the two models can compensate for each other. Specifically, the image-based point cloud model can be used to extend the SLAM point clouds from short distances, whereas the SLAM model can provide a scale for the image-based model. Moreover, as discussed, the SLAM results can provide additional depth constraints for the SfM process to improve it. The designed scale-adaptive registration can then merge those kinds of point clouds into a common coordinate system to produce enhanced and extended 3D mapping results. Two challenging cases were used to evaluate the performance of the proposed solution. The theoretical analysis and experimental validation yield the following conclusions.

1. The incorporation of additional depth constraints from the SLAM results benefits the offline SfM; moreover, the data collection can be completed at one time measurement.
2. The fusion of the distant point cloud model from the RGB image sequences to the short-range point clouds from the depth sensors can significantly improve the coverage of 3D mapping results (more than 50% in our cases).
3. The designed scale-adaptive registration can ensure the geometric accuracy of the structural edges (i.e., accuracy in distant ranges is 1% at 20 m in our cases) and the 3D point quality (i.e., the bias is lower than 10 cm for the major surface of about 800 $m^2$).

Although RGB-D sensors are rarely used in real mapping cases, this paper shows the potential of such sensors to generate enhanced and extended 3D models with high mobility and low cost. Such low-cost equipment could be used to quickly build 3D models in large indoor spaces, such as shopping malls, hospitals, and airports, for a variety of indoor navigation applications. Thus, our future work will not only focus on the methods of related technical aspects but also consider the application of equipment in mapping and modeling projects.

## Acknowledgments

This work was supported by grants from the Hong Kong Polytechnic University (Project Nos. 1-ZEAB and 1-ZVN6) and grants from the National Natural Science Foundation of China (Project Nos. 41671426 and 41471345).

## References

Aiger, D., N. J. Mitra and D. Cohen-Or. 2008. 4-points congruent sets for robust pairwise surface registration. *ACM Transactions on Graphics* (*TOG*): 85.

Besl, P. J. and N. D. McKay. 1992. Method for registration of 3-D shapes, sensor fusion IV: Control paradigms and data structures. *International Society for Optics and Photonics*: 586–607.

Bolles, R. C. and M. A. Fischler. 1981. A RANSAC-based approach to model fitting and its application to finding cylinders in range data. *IJCAI*: 637–643.

Byravan, A. and D. Fox. 2017. Se3-nets: Learning rigid body motion using deep neural networks. Pages 173–180 in *2017 IEEE International Conference on Robotics and Automation* (*ICRA*).

Chow, J. C., D. D. Lichti, J. D. Hol, G. Bellusci and H. Luinge. 2014. IMU and multiple RGB-D camera fusion for assisting indoor stop-and-go 3D terrestrial laser scanning. *Robotics* 3 (3):247–280.

Comport, A. I., E. Malis and P. Rives. 2007. Accurate quadrifocal tracking for robust 3d visual odometry. *ICRA*, Citeseer: 40–45.

Crandall, D., A. Owens, N. Snavely and D. Huttenlocher. 2011. Discrete-continuous optimization for large-scale structure from motion. Pages 3001–3008 in *2011 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*).

Dai, A., M. Nießner, M. Zollhöfer, S. Izadi and C. Theobalt. 2017. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics* (*TOG*) 36 (4):76a.

Dryanovski, I., R. G. Valenti and J. Xiao. 2013. Fast visual odometry and mapping from RGB-D data. Pages 2305–2310 in *2013 IEEE International Conference on Robotics and Automation* (*ICRA*).

Du, S., N. Zheng, S. Ying, Q. You and Y. Wu. 2007. An extension of the ICP algorithm considering scale factor. Pages V-193-V-196 in *IEEE International Conference on Image Processing* (*ICIP*), 2007.

Engel, J., T. Schöps and D. Cremers. 2014. LSD-SLAM: Large-scale direct monocular SLAM. Pages 834–849 in *European Conference on Computer Vision*. Springer.

Frahm, J. M., P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu and M. Pollefeys. 2010. Building Rome on a cloudless day. Pages 368–381 in *European Conference on Computer Vision*. Springer.

Gao, X., R. Wang, N. Demmel and D. Cremers. 2018. LDSO: Direct sparse odometry with loop closure. *arXiv* preprint: arXiv:1808.01111.

Garg, R., B. G., V. K., G. Carneiro and I. Reid. 2016. Unsupervised CNN for single view depth estimation: Geometry to the rescue. Pages 740–756 in *European Conference on Computer Vision*. Springer.

Ge, X. 2017. Automatic markerless registration of point clouds with semantic-keypoint-based 4-points congruent sets. *ISPRS Journal of Photogrammetry and Remote Sensing* 130:344–357.

Ge, X. and T. Wunderlich. 2015. Target identification in terrestrial laser scanning. *Survey Review* 47 (341):129–140.

Ge, X. and T. Wunderlich. 2016. Surface-based matching of 3D point clouds with variable coordinates in source and target system. *ISPRS Journal of Photogrammetry and Remote Sensing* 111: 1–12.

Ge, X. 2016. Terrestrial laser scanning technology from calibration to registration with respect to deformation monitoring. Dissertation, Technische Universität München.

Gherardi, R., M. Farenzena and A. Fusiello. 2010. Improving the efficiency of hierarchical structure-and-motion. Pages 1594–1600 in *CVPR*.

Havlena, M. and K. Schindler. 2014. Vocmatch: Efficient multiview correspondence for structure from motion. Pages 46–60 in *European Conference on Computer Vision*. Springer.

Heinly, J., E. Dunn and J. Frahm. 2012. Comparative evaluation of binary features. Pages 759–773 in *Computer Vision–ECCV 2012*. Springer.

Henry, P., M. Krainin, E. Herbst, X. Ren and D. Fox. 2010. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. In the *12th International Symposium on Experimental Robotics* (*ISER*). Citeseer.

Hesch, J. A., D. G. Kottas, S. L. Bowman and S. I. Roumeliotis. 2014. Camera-IMU-based localization: Observability analysis and consistency improvement. *The International Journal of Robotics Research* 33 (1):182–201.

Huttenlocher, D. 1991. Fast affine point matching: An output-sensitive method. Pages 263–268 in *Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (*CVPR*), 1991.

Johnson, R. and T. Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv* preprint: arXiv:1412.1058.

Kerl, C., J. Sturm and D. Cremers. 2013. Dense visual SLAM for RGB-D cameras. Pages 2100–2106 in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*). Citeseer.

Kerl, C., J. Sturm and D. Cremers. 2013. Robust odometry estimation for RGB-D cameras. Pages 3748–3754 in *2013 IEEE International Conference on Robotics and Automation* (*ICRA*)..

Kümmerle, R., G. Grisetti, H. Strasdat, K. Konolige and W. Burgard. 2011. g 2 o: A general framework for graph optimization. Pages 3607–3613 in *2011 IEEE International Conference on Robotics and Automation* (*ICRA*). .

LeCun, Y., Y. Bengio and G. Hinton. 2015. Deep learning. *Nature* 521 (7553):436.

Liu, H., M. Chen, G. Zhang, H. Bao and Y. Bao. 2018. ICE-BA: Incremental, consistent and efficient bundle adjustment for visual-inertial SLAM. Pages 1974–1982 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Lo, T. R. and J. P. Siebert. 2009. Local feature extraction and matching on range images: 2.5 D SIFT. *Computer Vision and Image Understanding* 113 (12):1235–1250.

Lourakis, M. I. and A. A. Argyros. 2009. SBA: A software package for generic sparse bundle adjustment. *ACM Transactions on Mathematical Software* (*TOMS*) 36 (1):2.

Magnusson, M. 2009. The three-dimensional normal-distributions transform: An efficient representation for registration, surface analysis, and loop detection. Örebro Universitet.

Mur-Artal, R. and J. D. Tardós. 2017. Orb-slam2: An open-source slam system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics* 33 (5):1255–1262.

Newcombe, R. A., S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J., Davison and A. Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. Pages 127–136 in *2011 10th IEEE International Symposium on Mixed and Augmented Reality* (*ISMAR*).

Schonberger, J. L. and J. Frahm. 2016a. Structure-from-motion revisited. Pages 4104–4113 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Schonberger, J. L. and J. Frahm., J., 2016b. Structure-from-motion revisited. Pages 4104–4113 in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4104-4113.

Snavely, K. N. 2008. Scene reconstruction and visualization from internet photo collections. University of Washington, Wash..

Snavely, N., S. M. Seitz and R. Szeliski. 2006. Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics* (*TOG*): 835–846.

Snavely, N., S. M. Seitz and R. Szeliski. 2008. Modeling the world from internet photo collections. *International Journal of Computer Vision* 80 (2):189–210.

Steinbrucker, F., C. Kerl and D. Cremers. 2013. Large-scale multi-resolution surface reconstruction from RGB-D sequences. Pages 3264–3271 in *Proceedings of the IEEE International Conference on Computer Vision*.

Tang, S., Q. Zhu, W. Chen, W. Darwish, B. Wu, H. Hu and M. Chen. 2016. Enhanced RGB-D mapping method for detailed 3D indoor and outdoor modeling. *Sensors* 16 (10):1589.

Tykkälä, T., C. Audras and A. I. Comport. 2011. Direct iterative closest point for real-time visual odometry. Pages 2050–2056 in *2011 IEEE International Conference on Computer Vision Workshops* (*ICCV Workshops*). .

Vijayanarasimhan, S., S. Ricco, C. Schmid, R. Sukthankar and K. Fragkiadaki. 2017. Sfm-net: Learning of structure and motion from video. *arXiv* preprint: arXiv:1704.07804.

Westoby, M. J., J. Brasington, N. F. Glasser, M. J. Hambrey and J. M. Reynolds. 2012. "Structure-from-Motion" photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* 179:300–314.

Whelan, T., H. Johannsson, M. Kaess, J. J. Leonard and J. McDonald. 2013. Robust real-time visual odometry for dense RGB-D mapping. Pages 5724–5731 in *2013 IEEE International Conference on Robotics and Automation* (*ICRA*). .

Wolff, K., C. Kim, H. Zimmer, C. Schroers, M. Botsch, O. Sorkine-Hornung and A. Sorkine-Hornung. 2016. Point cloud noise and outlier removal for image-based 3D reconstruction. Pages 118–127 in *2016 Fourth IEEE International Conference on 3D Vision* (*3DV*).

Wu, B., Y. Zhang, and Q. Zhu. 2012. Integrated point and edge matching on poor textural images constrained by self-adaptive triangulations. *ISPRS Journal of Photogrammetry and Remote Sensing* 68 (2012):40–55.

Wu, C. 2011. VisualSFM: A visual structure from motion system. <http://ccwu.me/vsfm/>.

Wu, J., T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba and W. T. Freeman. 2016. Single image 3d interpreter network. Pages 365–382 in *European Conference on Computer Vision*. Springer.

Ye, L. and B. Wu. 2018. Integrated image matching and segmentation for 3d surface reconstruction in urban areas. *Photogrammetric Engineering & Remote Sensing* 84 (3):35–48.