International Journal of Automation and Computing 15(4), August 2018, 443-453
DOI: 10.1007/s11633-018-1114-2

Elaborate Scene Reconstruction with a Consumer Depth Camera

Jian-Wei Li[1,2] Wei Gao[1,2] Yi-Hong Wu[1,2]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100080, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: A robust approach to elaborately reconstruct the indoor scene with a consumer depth camera is proposed in this paper. In order to ensure the accuracy and completeness of 3D scene model reconstructed from a freely moving camera, this paper proposes new 3D reconstruction methods, as follows: 1) Depth images are processed with a depth adaptive bilateral filter to effectively improve the image quality; 2) A local-to-global registration with the content-based segmentation is performed, which is more reliable and robust to reduce the visual odometry drifts and registration errors; 3) An adaptive weighted volumetric method is used to fuse the registered data into a global model with sufficient geometrical details. Experimental results demonstrate that our approach increases the robustness and accuracy of the geometric models which were reconstructed from a consumer-grade depth camera.

Keywords: 3D reconstruction, image processing, geometry registration, simultaneous localization and mapping (SLAM), volumetric integration.

1 Introduction

Reconstructing the real world scenes is known as a particularly challenging problem in computer vision field. Many tools have been applied to perceive accurate 3D world, including stereo cameras, laser range finders, monocular cameras, and depth cameras.

The emergence of consumer depth cameras, in particular the Microsoft Kinect, provides an opportunity to develop reconstruction systems conveniently. Izadi et al.[1, 2] introduced the Kinect-fusion algorithm which used a volumetric representation of the scene, known as the truncated signed distance function (TSDF)[3], in conjunction with fast iterative closest point (ICP)[4] pose estimation to provide a real-time fused dense model of the scene. Kinect-fusion works according to fixed grid spaces and the algorithm has no loop closure detection or global optimization. Therefore, it has good effectiveness only for local small scenes.

When we reconstruct complete and high-quality real world scenes with consumer-grade depth cameras, the principal problems are serious sensor noise and accumulated visual odometry errors which may result in distortions in the reconstructed 3D models. For the past few years, researchers have explored a number of approaches to address these issues.

Some systems achieved high accuracy localization by combining the depth data with red green blue (RGB) images[5−7] or an inertial measurement unit (IMU)[8−10]. But most depth cameras are not accompanied by color cameras. Even if the color camera is present, their view points are different and their shutter may not be perfectly synchronized. Besides, a consumer-grade IMU also suffers from sensor noise and is subject to large drifts over time.

Other systems tried to detect loop closures more explicitly and distributed the accumulated error across the pose graphs[11−13]. Choi et al.[11] have demonstrated impressive globally optimized 3D surface models, which extended the frame-to-model incremental tracking and reconstruction technique utilized in Kinect-fusion. The key idea of Choi′s algorithm is to combine geometric registration of scene fragments with a robust global optimization based on line processes. However, this algorithm also suffers from failure in geometric registration in part derived from a uniform segmentation strategy.

In this paper, we present an elaborate and robust scene reconstruction method, which can be applied to real-world scenes and has high reconstruction quality. The main contributions of our work contain three aspects: First, in order to increase the accuracy of 3D model, we smooth the depth images by a depth adaptive bilateral filter according to the depth camera′s noise model. Second, to reduce the visual odometry drift and improve the geometric registration accuracy, we propose a content-based segmentation to partition the depth image sequence into fragments, and perform geometric registration from local to global. Third, we fuse the data with an adaptive weighting TSDF by which the details of areas with high accuracy and regions of interest (ROI) can be preserved.

This paper is structured as follows. Section 2 discusses the related work of indoor scene 3D reconstruction while

Research Article
Manuscript received September 6, 2017; accepted December 27, 2017; published online April 17, 2018
Recommended by Associate Editor Jangmyung Lee

Section 3 describes the pipeline of our 3D reconstruction system. The details of the proposed method are presented in Section 4. Section 5 presents experiment results and discussions while Section 6 presents some concluding remarks.

2 Related work

Many algorithms are designed for depth image augmentation, complete scene processing, and volumetric integration. Now we briefly discuss the related work and further state the detailed motivations of our methods.

The raw depth images obtained from commercially available depth cameras are easy to use, but are affected by significant amounts of noise. Researchers have made lots of analysis for the accuracy and resolution of depth data[14−18]. A commonly used modification is the bilateral filter[19] which modifies the weighting to account for variation of intensity thereby effectively carrying out a robust smoothing operation. But the bilateral filter applied to depth images implicitly assumes that depth values have uniform uncertainty. Xiao et al.[20] improved the depth map by using TSDF to voxelize the space, accumulating the depth map from nearby frames using the camera poses, and then using ray casting to get a reliable depth map for each frame. Chen and Koltun[21] developed a global high-resolution media resource function (MRF) optimization approach to improve the accuracy of depth images. The algorithm performed block coordinate descent by optimally updating a horizontal or vertical line in each step. The idea of joint bilateral upsampling[22] is to apply a spatial filter to the low resolution image, while a range filter is jointly applied on another full resolution image. It is used to augment the quality of image with the help of a high resolution color image. In contrast to these, we smooth the depth image by a depth adaptive bilateral filter which is derived from the noise model of a structured-light stereo based depth camera, and can be used easily.

A complete scene is reconstructed from views acquired along the camera trajectory, and each view exposes only a small part of the environment. Whelan et al.[12, 23] permitted the area mapped by the TSDF to move over time, which allows to continuously augment the reconstructed surface in an incremental fashion as the camera translates and rotates in the real world. An inherent problem is dealing with the tracking drift due to accumulated pose estimation errors. Zeng et al.[24] introduced 3DMatch to robustly match local geometry, which is a data-driven local feature learner that jointly learns a geometric feature representation and an associated metric function from a large collection of real world scanning data. Halber and Funkhouser[25] introduced a fine-to-coarse algorithm that detects planar structures spanning multiple RGB-D frames and establishes geometric constraints between them as they become aligned. Detection and enforcement of these structural constraints in the inner loop of a global registration algorithm guides the solution towards more accurate global registrations, even without detecting loop closures. Choi et al.[11, 13] dealt with the accumulated pose estimation errors by reconstructing locally smooth scene fragments and deforming these fragments in order to align to each other. However, it is not very effective for the reconstruction of real world scenes with a hand-held camera. Therefore, we extend this method and design a content-based segmenting strategy to increase the accuracy of local fusion and global registration.

In volumetric integration, TSDF is discretized into a voxel grid to represent a physical volume of space. Each voxel $\boldsymbol{v}$ contains a signed distance d indicating the distance from the cell to a surface and a weight w representing confidence in the accuracy of the distance. The actual world surfaces are encoded as the zero crossings of the distance field and can be extracted by ray casting[26] or marching cubes[27]. The weight w trivially assumes a constant for all voxels, i.e., $w = 1$. It is suitable for distance sensors that can deeply penetrate objects, such as radar. Curless and Levoy[3] assigned a constant weight to all voxels up to a certain penetration depth, after which the weight linearly decreases to zero at a penetration depth. Newcombe et al.[2] proposed an exponential weighting function motivated by a Gaussian noise model of depth measurements. Zollhöfer et al.[28] obtained fine-scale detail through volumetric shading-based refinement (VSBR) of a distance field to solve the problem of over-smoothing. However, this algorithm is effective only in the controlled light source reconstruction. Zhou and Koltun[29] detected points of interest in the scene based on their distance from the principal axis and preserved detailed geometry around them with a global optimization. Inspired by these methods, we propose an adaptive weighting function whose value varies with the position of the points, and give higher weights to the points with high accuracies and interests.

3 Pipeline for scene reconstruction

An overview of our scene reconstruction framework is shown in Fig. 1. The scene reconstruction pipeline consists of three main stages and each stage is briefly described as follows.

Image capture and processing. The raw depth images are captured with a depth camera based on structured-light stereo, such as Microsoft Kinect for Windows and Asus Xtion Pro Live. Before the scene reconstruction, we improve the quality of depth images by the proposed depth adaptive bilateral filter algorithm, which can effectively remove the noises from these depth cameras.

Local-to-global registration. We introduce a local-to-global registration strategy to reduce visual odometry drift errors and achieve complete scene reconstruction. The large scene is partitioned into fragments of various sizes with the proposed content-based segmentation method. All fragments are locally fused with ICP registration algorithm,

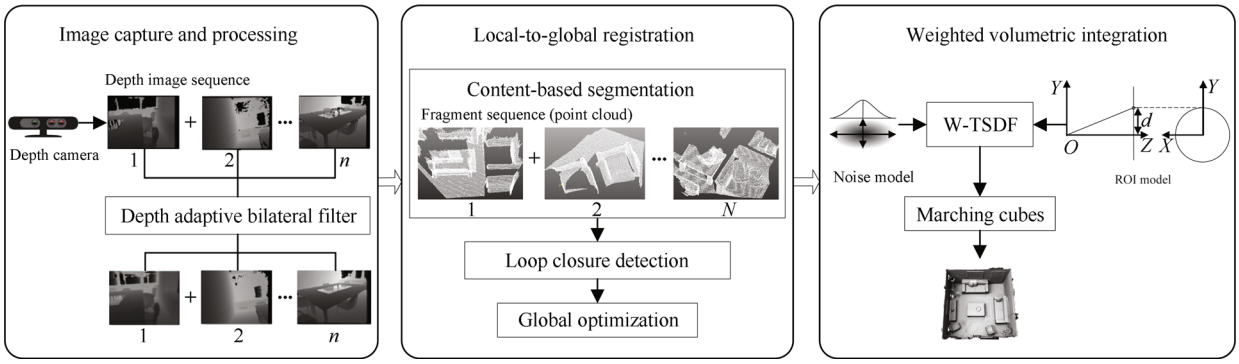


Fig. 1 Pipeline of the proposed scene reconstruction system

and a global loop closure is detected for each pairs of fragments with a geometric registration algorithm[11, 30]. The benefit of this registration is that we can get more reliable geometry information because information extracted from content-based fragment is more complete than an individual depth image. The pose of fragment i and the rigid transformation aligning fragment i to fragment $i+1$ are computed by Kintinuous framework[23]. The false positive loop closures are pruned by the pose graph optimization[11] with the line process constraint using the g2o framework[31].

Weighted volumetric integration. The registered fragments are fused into a global model through volumetric integration. The proposed weighting function of TSDF is based on the camera′s noise characteristics and the proposed ROI model. Therefore, the details of areas with high accuracies and regions of interest can be preserved. The final mesh model is extracted with the marching cubes algorithm[27].

4 The proposed method

The proposed new techniques specifically include three aspects: depth adaptive bilateral filter, content-based segmentation, and adaptive weighted TSDF (W-TSDF). The following subsections describe the core methods in our system.

4.1 Depth adaptive bilateral filter

The consumer depth camera based on structured-light stereo can be treated as a pair of stereo cameras in a canonical position[16]. The depth z of a point is proportional to the disparity D i.e., $z = \frac{fB}{D}$, where B is the baseline and f is the focal length of the camera. The most significant source of disparity errors is quantization noise which arises when the disparity is estimated with a given finite precision. We differentiate the depth z with respect to the disparity D, and get a relationship as follows:

$$\frac{\partial z}{\partial D} = -\frac{z^2}{fB}. \tag{1}$$

The standard deviation (STD) of noise in depth measurement is proportional to the square of the depth. Thus, we propose a depth adaptive bilateral filtering method which is more effective to smooth depth images than the bilateral filtering[19].

Consider an observed depth image $\boldsymbol{Z}(\boldsymbol{u})$ where $\boldsymbol{u}$ denotes the location of a pixel. The depth estimation smoothed by the depth adaptive bilateral filtering is

$$\hat{\boldsymbol{Z}}(\boldsymbol{u}) = \frac{1}{W}\sum_{N(\boldsymbol{u}_k)} w_s(\boldsymbol{u}-\boldsymbol{u}_k) w_c(\boldsymbol{Z}(\boldsymbol{u})-\boldsymbol{Z}(\boldsymbol{u}_k))\boldsymbol{Z}(\boldsymbol{u}) \tag{2}$$

where w_s and w_c are Gaussian functions for spatial and range weighting with standard deviations of δ_s and δ_c, respectively, $N(\boldsymbol{u}_k)$ is the neighborhood of $\boldsymbol{u}$, and W is an overall normalizing factor to have a total sum of 1 over $N(\boldsymbol{u}_k)$.

$$\begin{cases} w_s = \exp\left(-\dfrac{(\boldsymbol{u}-\boldsymbol{u}_k)^2}{2(\delta_s)^2}\right) \\ w_c = \exp\left(-\dfrac{(\boldsymbol{Z}(\boldsymbol{u})-\boldsymbol{Z}(\boldsymbol{u}_k))^2}{2(\delta_c)^2}\right). \end{cases} \tag{3}$$

Unlike the bilateral filter, here the values of δ_c for the depth image are not fixed but vary with the depths. It can be approximated as

$$\delta_c = K\boldsymbol{Z}(\boldsymbol{u})^2 \tag{4}$$

where K is constant and its value depends on the camera parameters. In our experiments, K is set to be 16 and δ_s is 4.5 (in pixels).

Fig. 2 gives an example of the results of a depth image by the standard bilateral filter and the proposed depth adaptive bilateral filter. The color is for visualization only. We can see from the point cloud and mesh models that depth adaptive bilateral filter for the depth image is more effective to remove the noise and protect the edges. Both the foreground and background regions are appropriately smoothed while preserving depth discontinuity features since the proposed filter is adaptive to the variation of depth.

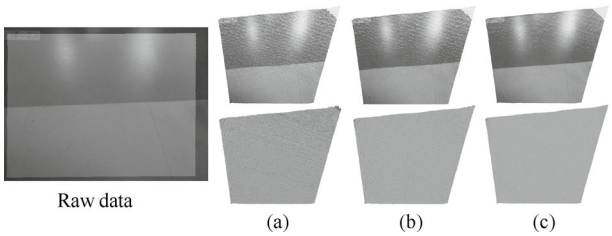


Fig. 2 Results of filtering on the depth image. Raw data shows the fusion image of a floor with the raw depth and color information. (a)–(c): Results with the raw depth image, depth image with bilateral filter, and depth image with the proposed adaptive bilateral filter, respectively. Top shows point clouds; Bottom shows mesh models.

4.2 Content-based segmentation

The segmentation of a depth image sequence is the key of the local-to-global registration. Segmentation based on visual content can effectively reduce the odometry drift and make the global loop closure more reliable.

The data obtained from a hand-held depth camera is usually related to the camera′s movement state and the complexity of the objects. Important objects or objects with lots of details are usually scanned fully and slowly, while other objects with less information, such as floors, walls, and some simple objects are scanned quickly. Besides, different people have different scanning habits. We assume that the regions with closer content have good consistency. By measuring the con-visibility information of depth images between the i-th depth frame and the first depth frame during the visual odometry estimation, we estimate the dissimilarity of visual contents between them, and decide the segmentation time. The relationship of con-visibility is shown in Fig. 3.

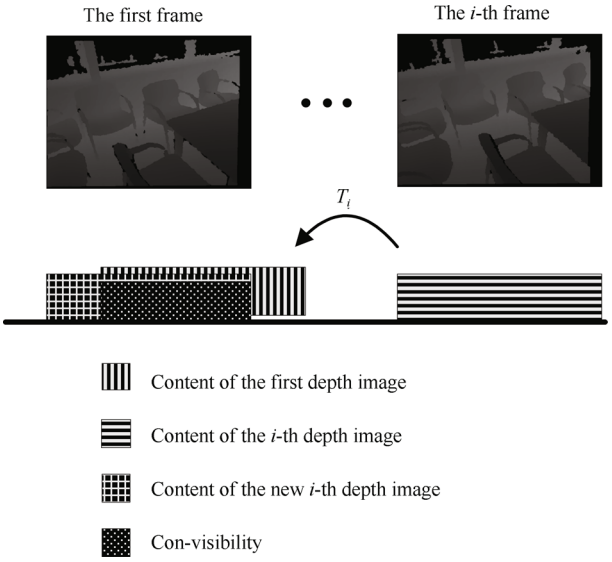


Fig. 3 Con-visibility between two depth images

Consider the estimated pose $\boldsymbol{P}_i$ for the camera of each frame, the rigid transformation $\boldsymbol{T}_{ij}$ aligns $\boldsymbol{P}_i$ to $\boldsymbol{P}_j$. We initialize the pose of the first depth frame for each fragment to be $\boldsymbol{P}_0$, and define the rigid transformation $\boldsymbol{T}_i$ aligns $\boldsymbol{P}_i$ to $\boldsymbol{P}_0$. In other words, the pose of the first depth frame for each fragment has the same value of the world coordinate system. We construct the content-based segmentation as follows:

First, we reconstruct the 3D point $\boldsymbol{p}(x_p, y_p, z_p)$ of the i-th depth frame corresponding to the pixel $\boldsymbol{u}_p$ using the inverse of the projection function π as: $\boldsymbol{p} = \pi^{-1}(\boldsymbol{u}_p, z(\boldsymbol{u}_p))$, where $z(\boldsymbol{u}_p)$ is the depth value of pixel $\boldsymbol{u}_p$ in the i-th depth frame.

Second, we transform the 3D point $\boldsymbol{p}$ from the i-th depth frame coordinate system to the world coordinate system, and obtain a 3D point $\boldsymbol{q}$, i.e., $\boldsymbol{q} = \boldsymbol{T}_i\boldsymbol{p}$. And then we reconstruct a new depth image with the pixel $\boldsymbol{u}_q$ by reprojecting the 3D point $\boldsymbol{q}(x_q, y_q, z_q)$ to 2D image plane:

$$\boldsymbol{u}_q = \left(\frac{f_x x_q}{z_q} + c_x,\ \frac{f_y y_q}{z_q} + c_y\right)^{\mathrm{T}} \tag{5}$$

where f_x and f_y are the focal lengths of depth camera.

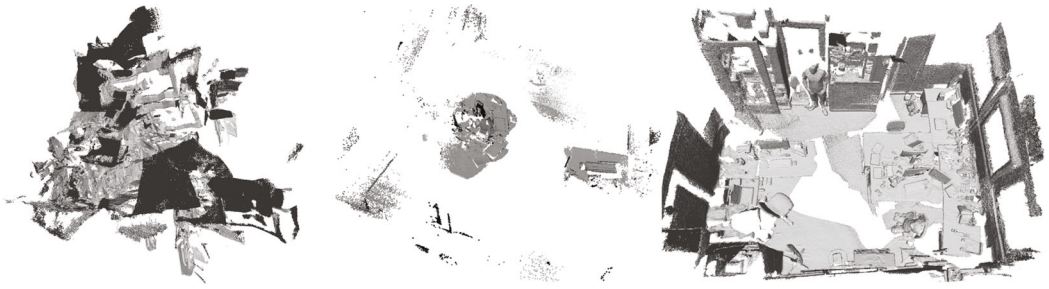
Third, we compute the number of available pixels of the new i-th depth frame and the first depth frame respectively, and then obtain the ratio ρ of the con-visibility as: $\rho = \frac{n^i}{n^0}$, here n^i and n^0 are the number of available pixels in the new i-th depth frame and the first depth frame respectively. To remove the invisible part, we perform a depth test for the new i-th depth frame before computing its number. Once the ratio ρ is less than the threshold, the pose of the camera will be initialized, and a new scene fragment will begin. In our experiments, the range of the ratio threshold is $0.7-0.9$. We also set the upper and lower thresholds for the number of frames in each fragment, so that the fragment will not be too small or too large.

Segmenting the input depth image sequence into fragments with the same size is a simple method, but it is difficult to select an appropriate number of frames for each fragment to reconstruct a good 3D scene model. We made some reconstruction experiments on augmented ICL-NUIM dataset by no segmentation, uniform segmentation (50-frame) and content-based segmentation for depth image sequence. The odometry drifts estimated with Median error and root mean square error (RMSE) are shown in Table 1. It comes out that the proposed segmentation can effectively reduce the odometry drift on average.

We also made some reconstruction experiments on real world scenes by the methods of uniform segmentation and the content-based segmentation for depth image sequence. The comparison results can be seen from Figs.4 (a) and 5 (a) to Figs. 4 (c) and 5 (c). The depth image sequences with the method of Choi et al.[11] are partitioned into fragments of 50 frames. Compared with the uniform segmentation, the content-based segmentation can automatically adjust the size of the fragments according to different datasets and data scanned by different operators. It can provide a good initial value for pose-graph optimization to increase the robustness, since the number of frames in each fragment is adaptive to the scanning process.

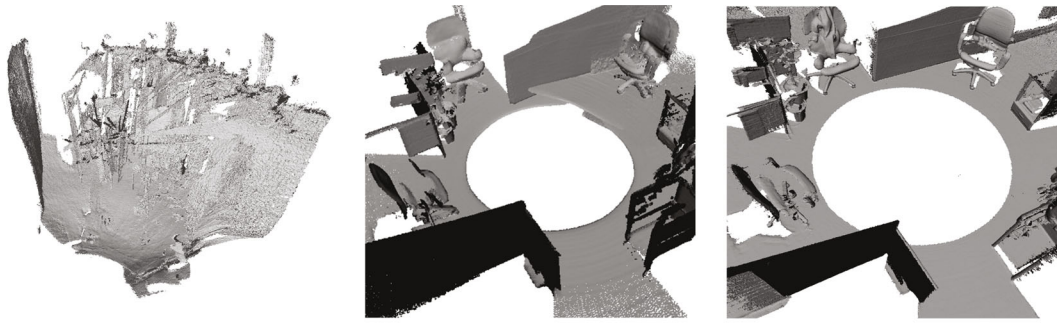
Table 1 Comparison of odometry drift (Median and RMSE) on augmented ICL-NUIM sequences. Note that the camera trajectories are estimated by Kinfu of point cloud library (PCL)

Sequence	No segmentation		Uniform segmentation		Content-based segmentation	
	Median	RMSE	Median	RMSE	Median	RMSE
Living room 1	0.309	0.424	**0.257**	0.406	0.272	**0.339**
Living room 2	0.566	0.617	0.399	0.396	**0.225**	**0.304**
Office 1	**0.124**	0.280	0.212	0.251	0.204	**0.229**
Office 2	**0.172**	**0.203**	0.273	0.308	0.276	0.300
Average	0.293	0.381	0.285	0.340	**0.244**	**0.293**



(a) (b) (c)

Fig. 4 Reconstruction results of fr1/room scene from the RGB-D simultaneous localization and mapping (SLAM) Dataset. (a) Results with the method of Choi et al.[11]. (b) Surfel model with Elasticfusion[32]. (c) Results with the proposed method.



(a) (b) (c)

Fig. 5 Reconstruction results of indoor scene which is scanned through a robot equipped with Microsoft Kinect for Windows. (a) Results with the method of Choi et al.[11] (b) Surfel model with Elasticfusion[32]. (c) Results with the proposed method.

4.3 Adaptive weighted TSDF

In this subsection, a TSDF with new adaptive weights is proposed to merge the registered data into a complete scene model, where different positions of points are considered. This can give sufficient details to the regions with high accuracies and interests.

For a given voxel $\boldsymbol{v}$ in the fused scene model F, the corresponding signed distance value $F(\boldsymbol{v})$ can be computed with respect to n input frames of a given depth image sequence:

$$\begin{cases} F(\boldsymbol{v}) = \dfrac{\sum\limits_{i=1}^{n} f_i(\boldsymbol{v}) w_i(\boldsymbol{v})}{W(\boldsymbol{v})} \\ W(\boldsymbol{v}) = \sum\limits_{i=1}^{n} w_i(\boldsymbol{v}) \end{cases} \tag{6}$$

where the signed distance function $f_i(\boldsymbol{v})$ is the projective distance (along the Z axis) between a voxel and the i-th depth frame $\boldsymbol{Z}_i$, and is defined as

$$f_i(\boldsymbol{v}) = [\boldsymbol{K}^{-1}\boldsymbol{Z}_i(\boldsymbol{u})[\boldsymbol{u}^{\mathrm{T}}, 1]^{\mathrm{T}}]_z - [\boldsymbol{v}]_z \tag{7}$$

where $\boldsymbol{u} = \pi(\boldsymbol{K}\boldsymbol{v})$ is the pixel into which the voxel center projects. We compute distance along the principal (Z) axis of the camera frame using the z component denoted as $[\cdot]z$. $\boldsymbol{K}$ is the known 3×3 camera intrinsic matrix, and π performs perspective projection.

The weighting function $w_i(\boldsymbol{v})$ represents the confidence in the accuracy of the distance. We can see from (7) that the value of signed distance function varies with the position of points. Thus, to get an accurate volumetric integration, the weighting function should take the position of points into account. In the following, we derive a more accurate weight function.

On one hand, as discussed in Section 4.1, the main noise in depth measurements is quantization noise, and the depth estimate z_i has standard deviation proportional to z_i^2. Con-

sider the observation $z_i = \mu + n_i$, where noise term n_i is Gaussian independently distributed, i.e., $n_i \sim N(0, \sigma_i^2)$. The maximum likelihood estimate (MLE) for μ is given as $\hat{\mu} \propto \sum_i \frac{z_i}{\sigma_i^2}$. This means that each individual measurement of z_i is weighted by a factor inversely proportional to the variance of the observation. Therefore, the weights should be inversely proportional to the fourth power of depth, i.e., $w_i \propto \frac{1}{z_i^4}$.

On the other hand, since the consumer depth camera has a narrow field of view, when we scan around the objects in a scene, we usually make the principal axis of the camera directly aligned the regions in which we are interested because the errors increase with the distances from points to the principal axis increasing. In order to emphasize the regions of interest, we give high weights to the points based on their distances from the principal axis.

Let $\boldsymbol{p}_i(x_i, y_i, z_i)$ be a point in the three dimension spaces, its squared-distance to Z-axis is $d_i^2 = x_i^2 + y_i^2$. The circle with a radius of d_i denotes the regions of interest (ROI), which is shown in Fig. 6.

Thus, we propose an adaptive weighting function motivated by the depth noise and ROI model, and the weighting function is assigned as follows:

$$w_i(\boldsymbol{v}) = \begin{cases} \dfrac{\exp(-\frac{d_i^2}{2\delta_r^2})}{z_i^4}, & 0 < z_i < d \\ 0, & z_i \geq d \end{cases} \tag{8}$$

where we use an Gaussian exponential model which uses Gaussian lateral noise[16] as the exponent to indicate the ROI model. δ_r is the STD of lateral noise in the depth image, and its value is $815 \times \frac{\boldsymbol{Z}(\boldsymbol{u})}{f}$. d is the radius threshold of ROI. We set an accurate weight to the signed distance function if the depth value is less than d, and set $w_i(\boldsymbol{v}) = 0$ if the depth value is greater than d. The choice of d mainly depends on the size of the scene. In most of our experiments, the value of d is set to be 2.8 m.

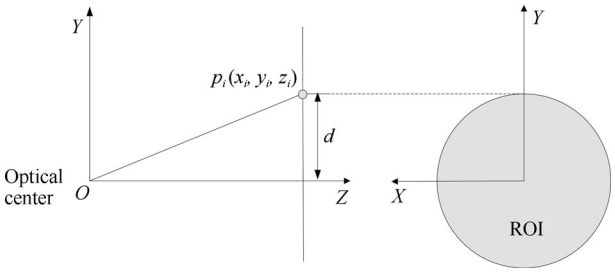
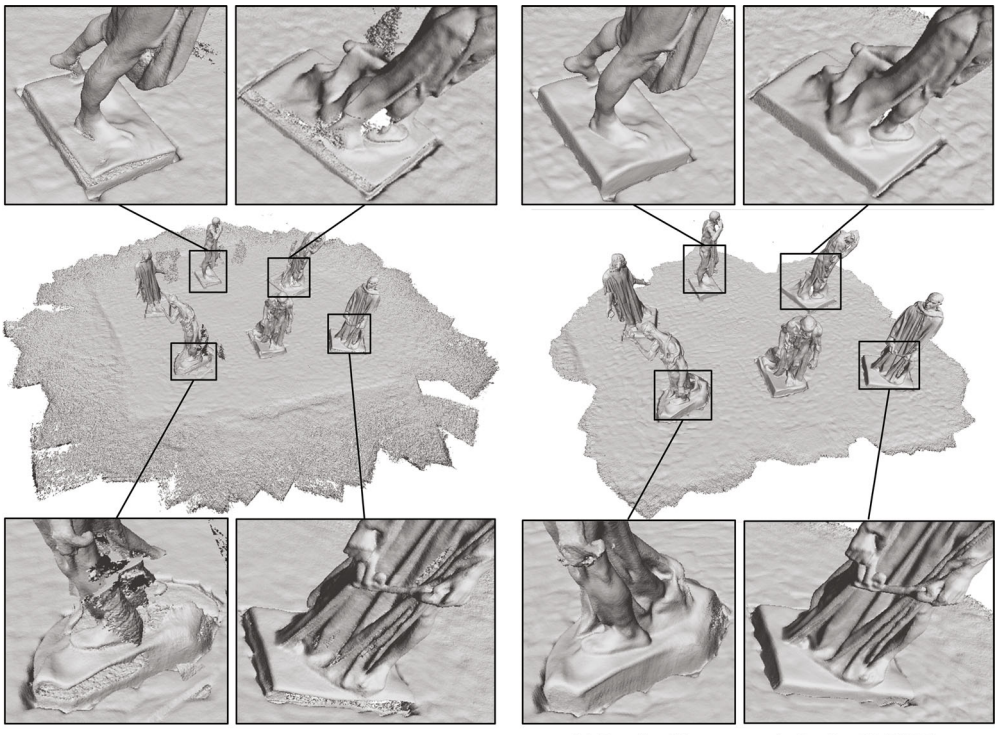


Fig. 6 Regions of interest model

Fig. 7 shows a comparison of volumetric integration by the standard TSDF with $w = 1$ and the proposed adaptive W-TSDF. It indicates that the adaptive W-TSDF not only improves the details of reconstruction but also removes the noise regions effectively.

5 Experiments

To illustrate the effectiveness of the proposed reconstruction method, we have carried out some experiments to evaluate the qualitative performance of the system.



(a) Result of the standard TSDF with $w = 1$ (b) Result of the proposed adaptive W-TSDF

Fig. 7 Comparison of volumetric integration with the standard TSDF and the proposed adaptive W-TSDF

5.1 Hardware

For all experiments, we ran the proposed system on a standard desktop PC with an Intel Core i7-4790 3.6 GHz CPU, and an Nvidia GeForce GTX 750 Ti 2GB GPU.

5.2 Data

One part of the data used in our experiments is captured by us with Microsoft Kinect for Windows. It streams VGA resolution (640×480) range and color images at 30 Hz. The operator who scans the scenes had no special training and no preview of the reconstruction. The consumer depth camera moves freely in the real world scenes without interference. Thus, the depth image sequences captured by us have more noise than the public datasets. We also use three public datasets:

RGB-D SLAM dataset. This dataset is provided by Handa et al.[33] for the evaluation of visual odometry and visual SLAM systems. Our experiments are conducted on the fr1/room sequence which is a complete indoor scene captured by the robot with Microsoft Kinect for Windows. The data is recorded at full frame rate (30 Hz) and sensor resolution (640×480).

3D scene Dataset. This dataset is provided by [30]. They used an Asus Xtion Pro Live camera, which streams VGA resolution (640×480) range and color images at 30 Hz. This camera uses the same prime sense range sensor as the Microsoft Kinect, but is somewhat smaller and lighter. In our experiments, we use two data sequences: Burghers and copy room.

Augmented ICL-NUIM dataset. The original ICL-NUIM dataset is based on the synthetic environments provided by [33]. Choi et al.[11] have augmented it in a number of ways to adapt it for evaluation of complete scene reconstruction pipeline. The average trajectory length is 36 m and the average surface area coverage is 88 %. Our experiments are conducted on four input sequences that model through hand-held imaging for the purpose of reconstruction: Living room 1, Living room 2, Office 1 and Office 2.

5.3 Real-world scenes

From the overall analysis of our experimental results, our reconstruction system is more robust than the state-of-art system proposed by Choi et al.[11] and Elasticfusion[32], especially in real-world scenes scanned through the robot. The precision of the reconstructed models with the data captured by us and public datasets are both better than the state-of-art offline reconstruction method.

Fig. 4 shows reconstruction results of fr1/room scene from the RGB-D SLAM dataset. Fig. 5 shows the reconstruction results of real-world indoor scene scanned by the robot equipped with Microsoft Kinect for Windows. The reconstruction results of the scene with Choi et al.[11] and Elasticfusion[32] are shown in Figs. 4(a), 5(a), 4(b) and 5(b). Both of the reconstructions are not good due to erroneous alignments. Figs. 4(c) and 5(c) show the reconstruction results with the proposed method. The numbers of depth frames used in Figs. 4 and 5 are 1 352 and 2 082, respectively. Fig. 8 shows the reconstruction result of a dynamic working area scanned by the robot equipped with Microsoft Kinect for Windows. Although the data captured by the robot have less surface information since the movements of robot are less flexible, we still get the complete scene models with good geometric structures since the proposed method is more robust.

Fig. 9 shows the reconstruction result of a real-world indoor scene manually scanned with Microsoft Kinect for Windows. The left of Fig. 9 illustrates the complete scene model reconstructed by the proposed reconstruction system. The room is 4 m wide and 5 m long. The number of depth frames is 6 000, and the total camera trajectory length is about 68 m. The center region of the room is a desk with a laptop on it. There are some dynamic disturbances by the wire to have more noises. During the scanning process, the data link of our Kinect is connected

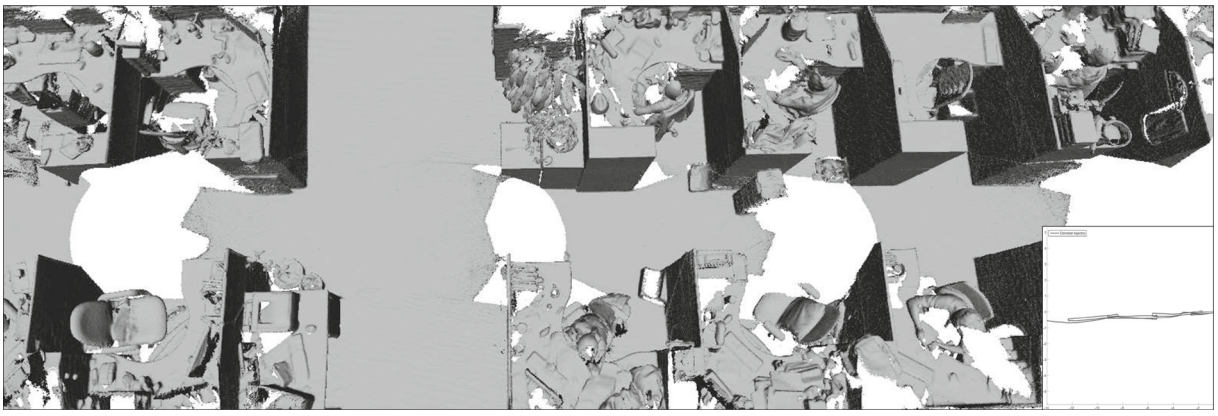


Fig. 8 Reconstruction results of the dynamic working area (with 10 000 frames) scanned through a robot equipped with Microsoft Kinect for Windows

to the laptop and the power cord of it is plugged into a power strip with long stumbled tail. However, we still get a complete model of the scene. As shown in top right of Fig. 9, the Venus and sofa in which we are interested are reconstructed with high-fidelity. The bottom right of Fig. 9 shows the reconstruction results of the corresponding objects in the room with the state-of-art method. Both the models of the desk and laptop have weak geometric details due to the reflective laptop surface and disturbed wires. We can obviously see from Fig. 9 that the 3D models reconstructed with the proposed method are better since our system is more robust to reconstruct real-world scenes with a consumer depth camera.

Fig. 10 shows the reconstruction results of real-world scene from 3D scene dataset, which is manually scanned with Asus Xtion Pro Live camera. The reconstruction results of the scene with Choi et al.[11] and Elasticfusion[32] are shown in Figs. 9(A) and 9(B). The corresponding details of 3D models for the burghers and copy room are shown in Figs. 7 and 11, respectively. The depth image sequences of the burghers and copy room are reconstructed with the method of Choi et al.[11] by segmenting the depth image sequence into fragments of 50 frames each and the proposed method respectively. The statues of the burghers are 2 m tall and the total camera trajectory length is about 184 m. The size of the copy room is 14 m^2 and the trajectory length is about 69 m. As can be seen from Figs. 7 and 11, the proposed method has an obvious advantage in retaining the geometric details. The models reconstructed with the proposed method are more accurate and perfect, and the details of regions marked with rectangles are preserved better than the state-of-art method.

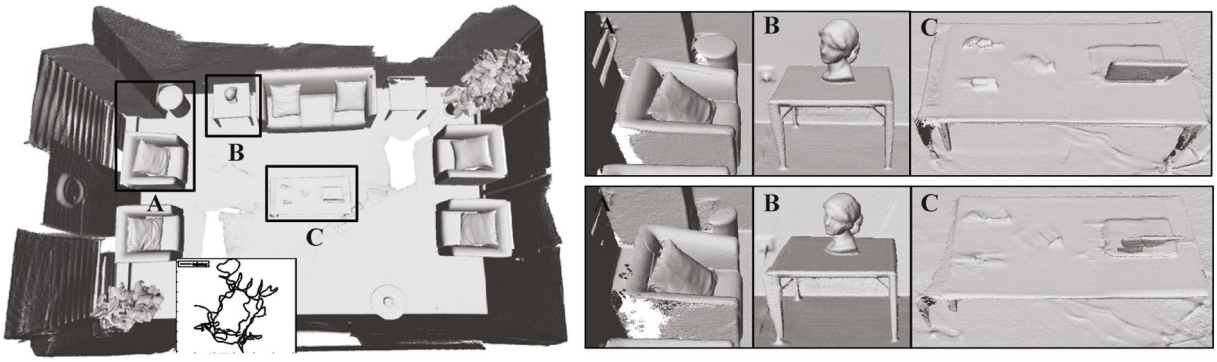
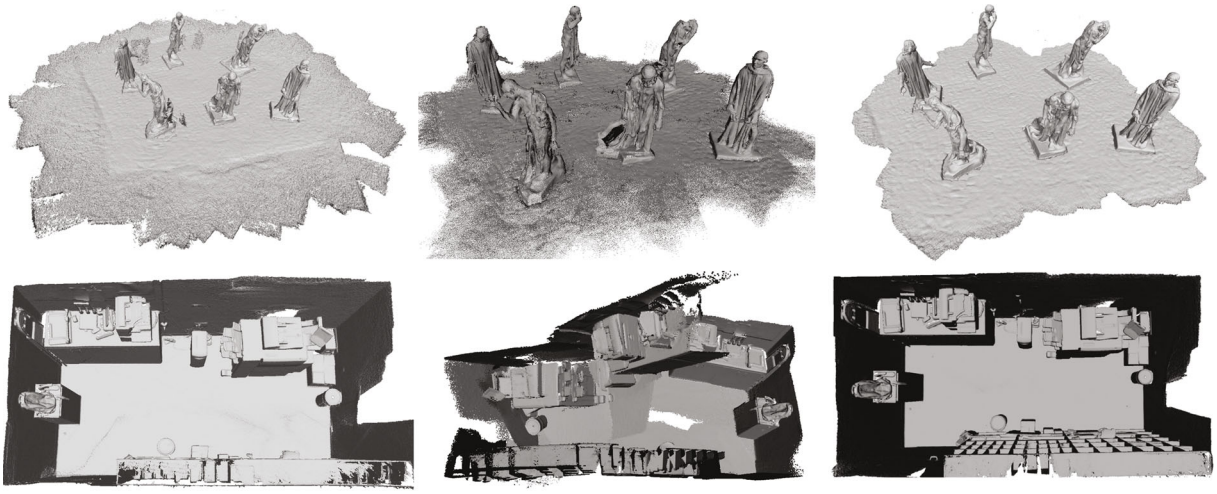


Fig. 9 Reconstruction results of indoor scene, which is manually scanned with Microsoft Kinect for Windows. The left shows the complete scene model and the camera trajectory information with the proposed method. The right shows the enlarged views of A, B and C in the room. Top of right shows the results with the proposed method. The bottom of right shows the results with the method of Choi et al.[11]



(a) (b) (c)

Fig. 10 Reconstruction results of real-world scene from 3D scene dataset, which is manually scanned with Asus Xtion Pro Live. The top shows the burghers. The bottom shows the copy room. (a) Results with the method of Choi et al.[11]. (b) Surfel model with Elasticfusion[32]. (c) Results with the proposed method.

5.4 Synthetic scenes

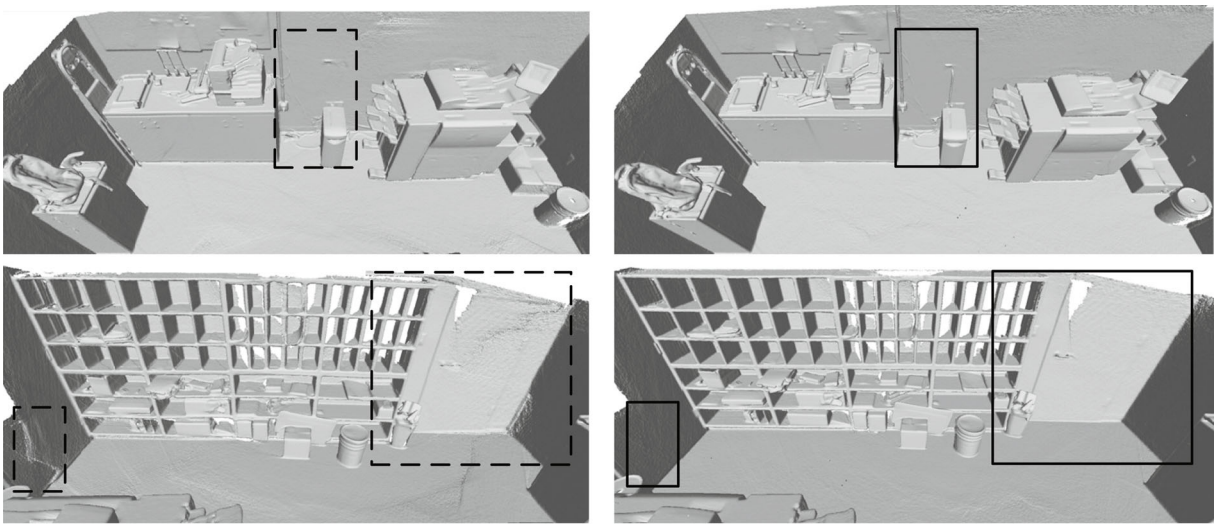
To evaluate the accuracy of camera trajectory and 3D model surfaces, we use four depth image sequences of augmented ICL-NUIM scenes. The accuracies are estimated by Kintinuous[23], DVO SLAM[34], SUN3D SfM[20], Choi et al.[11], Elasticfusion[32] and the proposed method. Note that the results of Choi et al.[11], Elasticfusion[32] were run by ourselves. And the results of DVO SLAM[34] and SUN3D SfM[20] are included in the paper of Choi et al.[11]

Camera trajectory evaluation. Table 2 reports the accuracy of the camera trajectories using the root mean square error (RMSE) metric described by Handa et al.[33] The RMSE of trajectory errors are in meters. As can be seen from Table 2, the average accuracy of trajectories with the proposed method is higher since it can reduce the accumulated odometry errors.

Assessing quality of 3D reconstruction. We use the open-source tool called CloudCompare to evaluate the surface reconstruction quality. The reconstruction surfaces of Augmented ICL-NUIM scenes can be compared against the ground-truth 3D model surfaces. The median distance of each reconstructed model to the ground-truth surface are reported in Table 3. It indicates that our method can effectively reduce the average error.

6 Conclusions

We presented a robust approach to elaborate scene reconstruction from a consumer depth camera. The main contribution of our research is using the local-to-global registration to obtain complete scene reconstruction and then the accuracy of 3D scene models is improved in the process of depth images filtering and weighted volumetric integration. The experimental results demonstrated that the proposed approach improves the robustness of reconstruction and enhances the fidelity of the 3D models produced from a consumer depth camera.



(a) Reconstruction results with the method of Choi et al. [11] (b) Reconstruction results with the proposed method

Fig. 11 Reconstruction details of the copy room from 3D scene dataset

Table 2 Accuracy of estimated camera trajectories (RMSE in meters) on augmented ICL-NUIM sequences

Sequence	Kintinuous[23]	DVO SLAM[34]	SUN3D SfM[20]	Choi et al.[11]	Elasticfusion[32]	The proposed
Living room 1	0.27	1.02	0.21	0.10	0.62	**0.09**
Living room 2	0.28	0.14	0.23	0.13	0.37	**0.11**
Office 1	0.19	0.11	0.24	0.13	0.13	**0.08**
Office 2	0.26	0.11	0.12	**0.09**	0.13	0.10
Average	0.250	0.345	0.200	0.113	0.313	**0.095**

Table 3 Surface reconstruction accuracy (median distance in meters) on augmented ICL-NUIM sequences

Sequence	Kintinuous[23]	DVO SLAM[34]	SUN3D SfM[20]	Choi et al.[11]	Elasticfusion[32]	The proposed
Living room 1	0.17	0.16	0.08	**0.03**	0.39	**0.03**
Living room 2	0.10	0.05	0.06	0.05	0.28	**0.03**
Office 1	0.09	0.08	0.11	**0.02**	0.03	**0.02**
Office 2	0.09	0.07	0.06	0.03	0.05	**0.02**
Average	0.113	0.090	0.078	0.033	0.188	**0.025**

Acknowledgements

This work was supported by the National Key Technologies R & D Program (No. 2016YFB0502002) and in part by National Natural Science Foundation of China (Nos. 61472419, 61421004 and 61572499).

References

[1] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, A. Fitzgibbon. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ACM, Santa Barbara, USA, pp. 559–568, 2011.

[2] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, A. Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In *Proceedings of the 10th IEEE International Symposium on Mixed and Augmented Reality*, IEEE, Basel, Switzerland, pp. 127–136, 2011.

[3] B. Curless, M. Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, ACM, New York, USA, pp. 303–312, 1996.

[4] S. Rusinkiewicz, M. Levoy. Efficient variants of the ICP algorithm. In *Proceedings of the 3rd International Conference on 3D Digital Imaging and Modeling*, IEEE, Quebec, Canada, pp. 145–152, 2001.

[5] C. Kerl, J. Sturm, D. Cremers. Robust odometry estimation for RGB-D cameras. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Karlsruhe, Germany, pp. 3748–3754, 2013.

[6] F. Steinbrücker, J. Sturm, D. Cremers. Real-time visual odometry from dense RGB-D images. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, IEEE, Barcelona, Spain, pp. 719–722, 2011.

[7] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, J. McDonald. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *International Journal of Robotics Research*, vol. 34, no. 4–5, pp. 598–626, 2015.

[8] J. Huai, Y. Zhang, A. Yilmaz. Real-time large scale 3D reconstruction by fusing Kinect and IMU data. In *Proceedings of ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, ISPRS, La Grande Motte, France, vol. II-3/W5, pp. 491–496, 2015.

[9] M. Niessner, A. Dai, M. Fisher. Combining inertial navigation and ICP for real-time 3D surface reconstruction. *Eurographics*, E. Galin, M. Wand, Eds., Strasbourg, France: The Eurographics Association, pp. 13–16, 2014.

[10] K. H. Yang, W. S. Yu, X. Q. Ji. Rotation estimation for mobile robot based on Single-axis gyroscope and monocular camera. *International Journal of Automation and Computing*, vol. 9, no. 3, pp. 292–298, 2012.

[11] S. Choi, Q. Y. Zhou, V. Koltun. Robust reconstruction of indoor scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Boston, USA, pp. 5556–5565, 2015.

[12] T. Whelan, M. Kaess, J. J. Leonard, J. McDonald. Deformation-based loop closure for large scale dense RGB-D SLAM. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Tokyo, Japan, pp. 548–555, 2013.

[13] Q. Y. Zhou, S. Miller, V. Koltun. Elastic fragments for dense scene reconstruction. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 473–480, 2013.

[14] A. Chatterjee, V. M. Govindu. Noise in structured-light stereo depth cameras: Modeling and its applications. arXiv: 1505.01936, 2015.

[15] K. Khoshelham. Accuracy analysis of Kinect depth data. In *Proceedings of International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, ISPRS, Calgary, Canada, vol. XXXVIII-5/W12, pp. 133–138, 2011.

[16] K. Khoshelham, S. O. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[17] C. V. Nguyen, S. Izadi, D. Lovell. Modeling Kinect sensor noise for improved 3D reconstruction and tracking. In *Proceedings of the 2nd International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, IEEE, Zurich, Switzerland, pp. 524–530, 2012.

[18] J. Smisek, M. Jancosek, T. Pajdla. 3D with Kinect. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, IEEE, Barcelona, Spain, pp. 1154–1160, 2011.

[19] C. Tomasi, R. Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the 6th International Conference on Computer Vision*, IEEE, Bombay, India, pp. 839–846, 1998.

[20] J. X. Xiao, A. Owens, A. Torralba. SUN3D: A database of big spaces reconstructed using SFM and object labels. In *Proceedings of IEEE International Conference on Computer Vision*, IEEE, Sydney, Australia, pp. 1625–1632, 2013.

[21] Q. F. Chen, V. Koltun. Fast MRF optimization with application to depth reconstruction. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Columbus, USA, pp. 3914–3921, 2014.

[22] J. Kopf, M. F. Cohen, D. Lischinski, M. Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics*, vol. 26, no. 3, Article number 96, 2007.

[23] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, J. McDonald. Kintinuous: Spatially extended KinectFusion. In *Proceedings of RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, Sydney, Australia, pp. 3–14, 2012.

[24] A. Zeng, S. Song, M. Niessner, M. Fisher, J. X. Xiao, T. Funkhouser. 3DMatch: Learning the matching of local 3D geometry in range scans. In *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Puerto Rico, USA, 2017.

[25] M. Halber, T. Funkhouser. Structured global registration of RGB-D scans in indoor environments. arXiv:1607.08539, 2016.

[26] S. Parker, P. Shirley, Y. Livnat, C. Hansen, P. P. Sloan. Interactive ray tracing for isosurface rendering. In *Proceedings of the Conference on Visualization*, IEEE, North Carolina, USA, pp. 233–238, 1998.

[27] W. E. Lorensen, H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *ACM SIGGRAPH Computer Graphics*, vol. 21, no. 4, pp. 163–169, 1987.

[28] M. Zollhöfer, A. Dai, M. Innmann, C. L. Wu, M. Stamminger, C. Theobalt, M. Niessner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics*, vol. 34, no. 4, Article number 96, 2015.

[29] Q. Y. Zhou, V. Koltun. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics*, vol. 32, no. 4, Article number 112, 2013.

[30] Q. Y. Zhou, J. Park, V. Koltun. Fast global registration. In *Proceedings of 14th European Conference on Computer Vision*, Amsterdam, The Netherlands, pp. 766–782, 2016.

[31] R Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, W. Burgard. g2o: A general framework for graph optimization. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Shanghai, China, pp. 3607–3613, 2011.

[32] T. Whelan, S. Leutenegger, R. F. Salas-Moreno, B. Glocker, A. J. Davison. ElasticFusion: Dense SLAM without a pose graph. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, 2015.

[33] A. Handa, T. Whelan, J. McDonald, A. J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *Proceedings of IEEE International Conference on Robotics and Automation*, IEEE, Hong Kong, China, pp. 1524–1531, 2014.

[34] C. Kerl, J. Sturm, D. Cremers. Dense visual SLAM for RGB-D cameras. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, Tokyo, Japan, pp. 2100–2106, 2013.

Jian-Wei Li received the B. Sc. degree in measurement, control technology and instrument, the M. Sc. degree in detection technology and automatic equipment from Beijing Jiaotong University, China in 2008 and 2011, respectively. She is currently a Ph. D. degree candidate in pattern recognition and intelligent system from National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China.

Her research interests include 3D reconstruction from images and SLAM technology.

E-mail: jianwei.li@nlpr.ia.ac.cn

ORCID iD: 0000-0002-4523-1692

Wei Gao received the B. Sc. degree in computational mathematics, the M. Sc. degree in pattern recognition and intelligent system from Shanxi University and the Ph. D. degree in pattern recognition and intelligent system from Institute of Automation, Chinese Academy of Sciences, China in 2002, 2005 and 2008, respectively. Since July 2008, he has joined Robot Vision Group of National Laboratory of Pattern Recognition, where he is currently an associate professor.

His research interests include 3D reconstruction from images and SLAM technology.

E-mail: wgao@nlpr.ia.ac.cn (Corresponding author)

ORCID iD: 0000-0003-2257-5684

Yi-Hong Wu received the Ph. D. degree in geometric invariants and applications from the Institute of Systems Science, Chinese Academy of Sciences, China in 2001. She is currently a professor at Institute of Automation, Chinese Academy of Sciences, China.

Her research interests include vision geometry, image matching, camera calibration, camera pose determination, SLAM, and their applications.

E-mail: yhwu@nlpr.ia.ac.cn