Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Review:

A survey on indoor 3D modeling and applications via RGB-D devices*

Zhilu YUAN[1], You LI[1], Shengjun TANG[†‡1], Ming LI[2], Renzhong GUO[1], Weixi WANG[1]

[1]*School of Architecture and Urban Planning, Research Institute for Smart Cities, Shenzhen University & China Guangdong–Hong Kong–Macau Joint Laboratory for Smart Cities & Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518060, China*

[2]*State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China*

[†]E-mail: shengjuntang@szu.edu.cn

Received Mar. 7, 2020; Revision accepted July 2, 2020; Crosschecked June 1, 2021

Abstract: With the fast development of consumer-level RGB-D cameras, real-world indoor three-dimensional (3D) scene modeling and robotic applications are gaining more attention. However, indoor 3D scene modeling is still challenging because the structure of interior objects may be complex and the RGB-D data acquired by consumer-level sensors may have poor quality. There is a lot of research in this area. In this survey, we provide an overview of recent advances in indoor scene modeling methods, public indoor datasets and libraries which can facilitate experiments and evaluations, and some typical applications using RGB-D devices including indoor localization and emergency evacuation.

Key words: 3D indoor mapping; RGB-D; Indoor localization; Construction monitoring; Emergency evacuation
https://doi.org/10.1631/FITEE.2000097 **CLC number:** P232

1 Introduction

In the past few decades, three-dimensional (3D) modeling of indoor environments has been a very popular research area. Generation of detailed 3D maps for indoor environments is essential for many mobile robot applications, including indoor navigation, facility management, virtual reality, and building information models (BIMs) (He and Habib, 2018). Three-dimensional indoor models can be obtained by either active or passive remote sensing systems. Passive sensors (such as imaging sensors) whose representative product is the ZED camera of Stereolabs, are able to acquire 3D information by applying the structure from the motion method (Konolige and Agrawal, 2008; Westoby et al., 2012; Ortiz et al., 2018), and are usually at a low cost. However, to extract 3D information from two-dimensional (2D) images, these sensors need extensive post-processing, including image matching and pose estimation, which are time-consuming, and especially suffer from dark environments, poorly textured areas, and motion blurs. Active sensors provide alternative ways to obtain accurate 3D models for indoor scenes. Laser scanners are typical devices to capture precise and reliable 3D information from indoor scenes. However, most existing laser scanning systems are

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 71901147, 41801392, 41901329, 41971354, and 41971341), the Research Program of Shenzhen S&T Innovation Committee, China (No. JCYJ20180305125131482), the Open Fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, MNR, China (Nos. KF-2019-04-010, KF-2019-04-014, and KF-2018-03-066), the Natural Science Foundation of Guangdong Province, China (Nos. 2019A1515010748 and 2019A1515011872), the Foundation of High-Level University Phase II, China (No. 000002110335), the Foundation of Shenzhen University for New Researchers, China (No. 2019056), the Innovation Team Program of Department Education of Guangdong Province, China (No. 2017KCXTD028), and the Guangdong Science and Technology Strategic Innovation Fund (the Guangdong–Hong Kong–Macau Joint Laboratory Program) (No. 2020B1212030009)

ORCID: Zhilu YUAN, https://orcid.org/0000-0002-7431-6599; Shengjun TANG, https://orcid.org/0000-0002-8262-7397

expensive and are short of RGB information (Chen C et al., 2018; Tang et al., 2019a). Recently, RGB-D sensors, such as Kinect (https://developer.microsoft.com/en-us/windows/kinect), Structure Sensor (https://structure.io/), and Intel RealSense (https://www.intelrealsense.com/lidar-camera-l515), have gained their popularity. They are inexpensive, lightweight, and accurate in acquiring 3D information and have promoted a rapid progress in indoor mapping. This technology combines laser scanning and visual systems. It collects depth and color information synchronously at high data rates. Considering their low cost and acceptable high accuracy, RGB-D sensors are the optimum option for indoor 3D modeling and related applications. This survey focuses on 3D modeling of indoor scenes and related applications using RGB-D devices.

1. For the modeling part, we have witnessed the development of 3D dense mapping and simultaneous localization and mapping (SLAM) pipelines, as shown in Fig. 1. Those that depend on only the RGB-D devices are now an important part. Visual SLAM systems are too diverse and complex to discuss. Here we focus only on RGB-D SLAM. The accuracy of the 3D maps rises as the accuracy of frame registration increases. We can classify them based on the differences of the registration method. Newcombe et al. (2011) first proposed a dense mapping system in which a global volumetric model is built to integrate all the depth data flow from the Kinect sensor. As the iterative closest point (ICP) algorithm is implemented, the camera pose can be obtained by capturing the trajectory of the live depth frame through the global surface model (Zeng et al., 2012; Chen JW et al., 2013; Whelan et al., 2013). On the other hand, a sparse style RGB-D system was proposed first by dos Santos et al. (2016). The feature-based SLAM system takes advantages of using fewer meaningful points to estimate the camera pose. As a consequence, this model requires less computational cost. Take an early feature-based SLAM system as an example. It was first introduced by Engelhard et al. (2011). This model first extracts the speeded up robust features (SURFs) from color images, and then maps it into the depth image. Meanwhile, the corresponding 3D points can be obtained to estimate the camera pose. Based on the proposed RBG-D SLAM method, the sensors can be used to model different kinds of objects, such as indoor space, tunnel, small objects, and human body, providing millimeter-level accuracy for small object reconstruction and centimeter-level accuracy for enclosure or semi-enclosed space modeling.

2. Since commodity RGB-D sensors were introduced, the 3D geometry capture technology has blossomed rapidly. Applications based on these technologies are expanding over a broad field, such as semantic understanding of scenes, indoor localization, and BIM reconstruction. For semantic understanding of scenes, Song et al. (2015) introduced a dataset called “SUN

Fig. 1 Samples of 3D modeling using RGB-D devices

RGB-D." This dataset contains 10 335 RGB-D images, which are either 2D or 3D. Dense annotations are embedded in these images, for both objects and rooms. It has been used for an object recognition task moving towards total scene understanding. Similarly, Dai et al. (2017) presented a ScanNet framework, which contains 2.5 million views in 1513 scenes annotated with 3D camera poses, surface reconstructions, and semantic segmentations. For indoor localization, Li et al. (2018) presented a visual localization and navigation method based on the Tango sensor, in which the direct depth information was employed to recover the real scale of the scenes and used to estimate the motion of the sensor. For semantically rich indoor reconstruction, Chen K et al. (2014) suggested working out the links between objects and their knowledge from the databases. Based on the links, they proposed a method to build the model of indoor scenes through low-quality RGB-D sequences. These two methods mentioned above are used mainly to reestablish the indoor furniture, while the reconstruction of indoor structural elements is not involved. Current research on indoor reconstruction does not give much of any method based on RGB-D mapping systems. However, Tang et al. (2019b) came up with a method to rebuild a semantically rich indoor model from low-quality RGB-D sequences. This method provides a quick and automatic way to classify and build the principal indoor structural elements from RGB-D data. These elements may include space, wall, floor, ceilings, windows, and doors.

2 Commercial RGB-D devices

In recent years, RGB-D devices, such as Kinect and Structure Sensors, have gained wide acceptance for SLAM and indoor-mapping applications. The advantage of these types of sensors is that they can capture and update 3D spatial information in real time. Meanwhile, they are more portable and have a lower cost. The measurement accuracy of RGB-D devices decreases with the increase of the measurement distance. Generally, only a depth within 3.5 m is able to be used for frame registration and indoor 3D mapping. This makes it hard to use in large scenes, like airports and underground space. Therefore, due to the limited measurement distance of consumer RGB-D cameras, significant work is needed on measurement error calibration or sensor integration to improve the mapping accuracy and range.

There are typical RGB-D sensors, like Kinect 1, Kinect 2, Azure Kinect DK, and Structure Sensor. In Table 1, we compare the hardware parameters of four types of sensors. The Kinect series devices can be connected to a desktop or a laptop for use, and the sensors themselves carry a depth camera, a near-

Table 1 Comparison of typical RGB-D sensor parameters

Device	Size (height×width×length)	Depth image resolution	Data acquisition frequency (frame/s)	Battery usage	Weight (g)
Kinect 1	64 mm×76 mm ×305 mm	320×240	<30	No battery	1360
Kinect 2	76 mm×165 mm ×350 mm	512×424	<30	No battery	1225
Azure Kinect DK	39 mm×103 mm ×126 mm	640×576, 320×288, 512×512, or 1024×1024	<30	No battery	400
Structure sensor	29 mm×28 mm ×119 mm	640×480	<30	3–4 h continuous mapping and 1000+ h standby	95

Device	Maximum effective distance (m)	View angle	Inertial measurement unit	Interface	Multi-device synchronization
Kinect 1	4.5	57°×43°	×	USB2/ USB3	√
Kinect 2	4.5	70°×60°	×	USB3	×
Azure Kinect DK	4	75°×65°, or 120°×120°	√	USB3	√
Structure sensor	3.5	58°×45°	×	Lightning	×

infrared transmitter, and an ordinary RGB camera each. Structure Sensor devices can work collaboratively with tablets and mobile phones. They carry only depth cameras and near-infrared cameras, and their RGB sensors use RGB cameras from external devices. The working principles of these two series are similar. Both can simultaneously generate 30 frames of 640×480 depth images and RGB images per second.

Kinect 1 and Kinect 2 have the same size, which is about 15 times larger than that of the Structure Sensor. From the point of view of battery usage, the former two do not have their own battery and require an external power supply during the mapping process, while the latter has its own battery, making continuous mapping for 3–4 h possible. Moreover, Kinect is 13 times heavier than Structure Sensor. As for data quality, the data resolution of Structure Sensor is higher, while the difference of the effective measurement distance between these two series is not that much. Based on the comparison above, Structure Sensor has better portability than Kinect series devices. In 2019, Microsoft officially released the Azure Kinect DK device. Its size and weight are much less than those of the previous generation (i.e., Kinect 2 sensor), and the ranging accuracy has also been developed and improved. Azure Kinect DK supports a multi-device synchronization mode. Meanwhile, body tracking software development kit (SDK), computer vision service interface, and voice SDK are embedded. As for the working mode, Azure Kinect DK can switch working modes according to different requirements to obtain depth visual images with different viewing angle sizes and a high resolution.

3 Datasets and modeling techniques

3.1 Public datasets

It is important that the evaluation and comparison of the algorithms should be scientific and objective, as public datasets and benchmarks are introduced. In recent years, public RGB-D datasets with indoor scenes have been released and used widely in many fields, such as scene reconstruction, semantic classification, and object recognition (Fig. 2). Common datasets and explicit evaluation indices can assist in the improvement of most advanced technologies. This view has been proved by several successful examples in the computer vision area. Some public RBG-D datasets are listed as follows, with a brief overview.

1. SUN RGB-D (Song et al., 2015): The benchmark is from Princeton University. There are four different sensors (i.e., Intel Realsense, Asus Xtion, Kinect

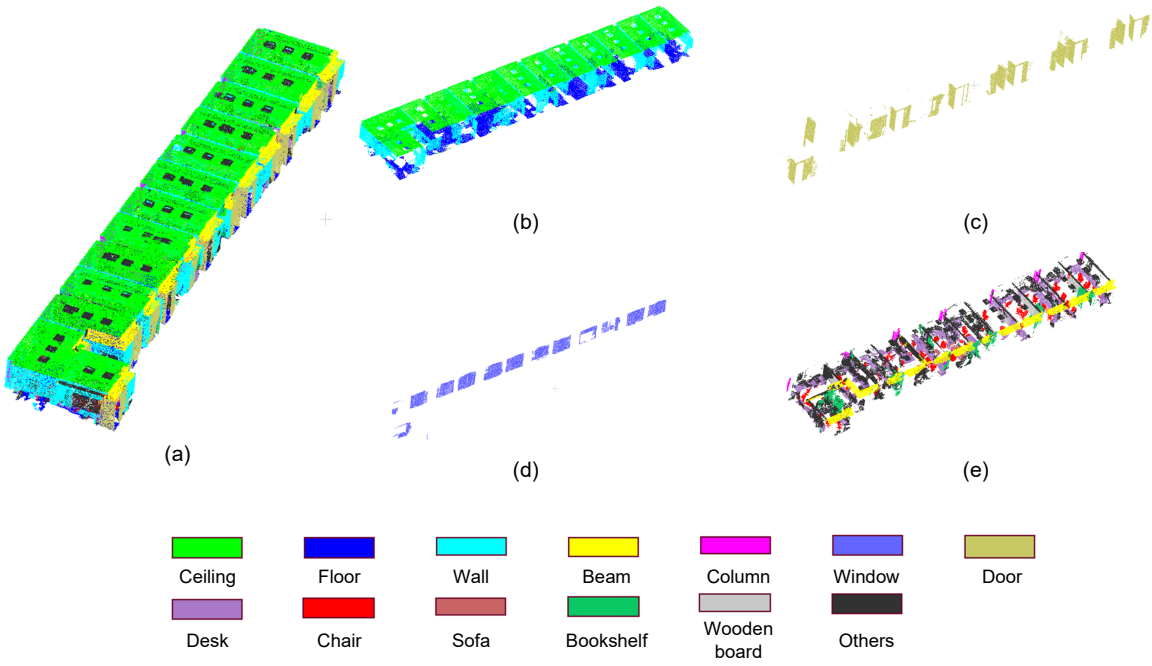

Fig. 2 Automatic semantic modeling of an office scene: the predicted point clouds of the whole scene (a), wall, floor, and ceiling components (b), door components (c), window components (d), and others (e)

1, and Kinect 2) introduced to the collection task. The whole dataset involves 10 335 RGB-D images with dense annotations. To be more specific, it contains 146 617 2D polygons and 64 595 3D bounding boxes with precise object orientations. Three-dimensional room layout and scene category are also included as annotations in each image. The dataset can be downloaded at http://rgbd.cs.princeton.edu.

2. ICL-NUIM (Chen K et al., 2014): The dataset name is the abbreviation of the Imperial College London and National University of Ireland Maynooth. To evaluate the visual odometry, 3D rebuilding, and SLAM algorithm, Chen K et al. (2014) set four interdependent camera trajectories to obtain high-quality RGB-D sequences. A ground-truth surface model was proposed. Through this model, quantitative evaluation of the map results can be carried out. Also, the accuracy of surface reconstruction can be ensured. The dataset also shows us the collection of handheld RGB-D camera sequences with perfect ground-truth poses in synthetically generated environments. This dataset can be downloaded at https://www.doc.ic.ac.uk/~ahanda/VaFRIC/iclnuim.html.

3. TUM RGB-D datasets (Sturm et al, 2012): Sturm et al. (2012) proposed a series of treatments and applications for RGB-D SLAM systems. A motion-capture system was introduced to capture the time-synchronized camera poses. Higher-speed cameras (100 Hz) were applied in this system so that the trajectories can be recorded. Microsoft Kinect was implemented to obtain high-accuracy image sequences, which contain both color and depth images. The resolution parameters were 640×480 (full sensor resolution), and the video frame rate was 30 Hz. The 39 datasets were obtained separately from a single office room and a factory corridor. To compare the simulation results of different approaches, an automatic evaluation tool was provided. This tool was aimed to measure the drifts and pose errors generated from the system mentioned above. The circumstance pictures, annotations of the dataset, codes, and tools listed above can be found at https://vision.in.tum.de/data/datasets/rgbd-dataset/tools.

4. Washington RGB-D scenes dataset (Lai et al., 2014): Fourteen new scenes recorded from a lounge, a coffee room, an office desk, and a meeting area were collected in the RGB-D scenes dataset v.2. These expanded the eight original scenes. Using patch volume mapping, Lai et al. (2014) aligned a set of Kinect RGB-D image frames to create a point cloud in each scene. The scenes contain furniture like chairs, coffee tables, tables, and sofas. The stitched scene point clouds can also provide nine object classes and labels for the background. This dataset is available at http://www.cs.washington.edu/rgbd-dataset.

3.2 Modeling techniques

Recently, many 3D techniques that use RGB-D devices and laser systems have been developed in robotics and computer vision research areas. In this review, we focus mainly on the techniques of modeling the RGB-D system, rather than the typical visual SLAM systems. Typically, frame registration, loop closure detection, and global optimization for the whole trajectories are the required technologies during RGB-D SLAM. We categorize the methods into two types based on the registration styles, including feature-based and dense styles.

First, the feature-based SLAM method is commonly used in the visual SLAM system. The core concept of this method is to achieve camera tracking using the detected feature points, and the pose updating is conducted by minimizing the distances of features. Basically, the pose of frames is obtained by minimizing the residual error of 2D or 3D correspondence through the iterative least squares method. The cost function of a feature-based tracking method is

$$T^{*}=\arg\min_{T}\left(\frac{1}{|A|}\sum_{i\in A}\left|T(P_{\mathrm{r}}^{i})-P_{\mathrm{s}}^{i}\right|^{2}\right), \tag{1}$$

where T^{*} consists of a rotation matrix $\boldsymbol{R}$ and a translation t, T is the initial transformation of the point cloud pairs, A contains the associations between feature points of the frames from two sensors, i is the point index of the point cloud, P_{r}^{i} is the i^{th} point of the reference point cloud, and P_{s}^{i} is the i^{th} point of the source point cloud.

At an early stage, the feature-based RGB-D camera tracking method was proposed by Engelhard et al. (2011). In this method, the features detected by SURFs are used to estimate the camera updating between adjacent frames. To reduce the drift during scanning, a pose graph optimization method is employed for global consistency. Similarly, an RGB-D

SLAM system is used for 3D flight autonomous navigation in a cluttered environment proposed by Huang et al. (2017). Henry et al. (2012) proposed an ICP variant method for camera tracking, in which both color information and depth information are used to enhance the robustness of frame tracking in textureless areas. Instead of evaluating the camera pose directly by minimizing the feature distances, a linear regression method proposed by Steinbrucker et al. (2011) is used to find the best rigid transformation between adjacent frames. As the performance of camera tracking can be influenced highly by feature descriptors, the tracking accuracy, robustness, and processing time with different kinds of features were investigated by Endres et al. (2014). To enhance the tracking robustness in the textureless area, Kerl et al. (2013) proposed a dense direct RGB-D tracking method by minimizing both photometric and depth errors. This is able to provide a high tracking accuracy. The random measurement error of the correspondences can also highly influence the mapping accuracy. There have been many studies on evaluating the theoretical random error of RGB-D devices and then using it to weight the contribution of each 3D correspondence during frame registration. dos Santos et al. (2016) and Vestena et al. (2016) presented the details of the weighed feature-based tracking method. Recently, a complete SLAM called "ORB-SLAM2" was proposed and used by Mur-Artal and Tardós (2017) for monocular, stereo visual cameras, and RGB-D camera tracking. However, the depth measurement error is ignored in this system. Tang et al. (2019a) proposed a vertex-to-edge RGB-D SLAM system to reduce the influence of the depth measurement error. The theoretical error of correspondences is used for weighting purposes, and the error of every pose updating is also used to adjust the edge contribution in global optimization. For the reconstruction part, Fehr et al. (2017) presented a novel 3D reconstruction algorithm based on an extended truncated signed distance function (TSDF), which enables continuous refinement of the static map while obtaining 3D reconstructions of dynamic objects in the scene. Wang et al. (2014) presented a 3D reconstruction approach using a Kinect RGB-D camera. For robust registration, they proposed to use both visual and geometry features combined with a structure from motion (SfM) technique, to enhance the robustness of feature matching and camera pose estimation. In addition, the semantic reconstruction from RGB-D cameras is a hot topic in the RGB-D mapping area (Song et al., 2015; Dai et al., 2017). Dai et al. (2017) proposed a scalable RGB-D acquisition and semantic annotation framework, which includes automated surface reconstruction and crowd-sourced semantic annotation. Ikehata et al. (2015) proposed a novel 3D modeling framework that reconstructs an indoor scene as a structured model from panoramic RGB-D images. This framework is able to recover the structural elements of indoor space, such as rooms, walls, and objects.

For the dense system, the ICP algorithm and relative variants are commonly used techniques (Segal et al., 2009). The ICP algorithm conducts the pose updating by minimizing the whole distances between two sets of point clouds. Compared with the feature-based tracking method, it is time-consuming and costly. Newcombe et al. (2011) first used the ICP method for RGB-D frame registration. The live frames are fused continually into the global volumetric model. In the early vision of the system, the memory cost almost increases linearly with the mapping range or the number of frames. The algorithm can work within an area of only 7 m^2. Therefore, the main challenge of a dense RGB-D system is how to decrease the time and memory cost. To address this issue, an octree-based structure on a graphics processing unit (GPU) was used for voxel representation and to reduce the computing cost during ICP (Zeng et al., 2012). Three-dimensional scenes modeled using this system can be three times larger than those modeled using the original KinectFusion system. However, neither method has addressed the problem of trajectories drift. Similarly, a hierarchical data structure for reducing memory consumption was proposed by Chen JW et al. (2013), and Nießner et al. (2013) explored a new system for large-scale volumetric reconstruction based on a spatial hashing scheme. Their 3D reconstruction hashing scheme supports real-time computation without trading quality or scale (Nießner et al., 2013; Dai et al., 2017). However, all of the works above lack drift correction. Generation of local structured 3D models with rigidity constraints from adjacent frames was proposed by Thomas and Sugimoto (2017) for camera tracking. Instead of using common features of frames, Shi et al. (2018) introduced a coplanar surface detection method. Coplanar surfaces were detected from adjacent RGB-D

frames, and the camera pose was recovered by coplanarity matching. Experimental results showed that the detected coplanarity constraints are able to improve the mapping accuracy significantly.

4 Applications

4.1 Indoor localization

It is universally accepted that locating people precisely in indoor conditions is key to services requiring location perception. Using WiFi or Bluetooth in indoor conditions is a very common solution (Fox et al., 1999; Biswas and Veloso, 2010; Chintalapudi et al., 2010). These techniques are frequently susceptible to the location of the WiFi source, the precision of the signal strength diagram, and the amount of clutter and dynamics in the conditions. Also, the quantity of WiFi makes a difference. Over the past few years, more and more research has been carried out on RGB-D devices, which can calculate the positions of smartphones or tablets with six degrees of freedom (Winterhalter et al., 2015; Cheng et al., 2018; Li et al., 2018).

Fig. 3 shows an example scenario for a method to position an RGB-D smartphone or tablet. As the 2D outline of the environment is typically available from architectural drawings, it can be used as the map for this method. This was first proposed by Winterhalter et al. (2015). To solve the indoor positioning problem in a low-light environment, Chen SN et al. (2017) considered infrared images and depth image pairs as inputs, and listed the most matched images after searching through the existing datasets. Using a Tango device, Li et al. (2018) proposed an approach which combines the feature-based method and the direct method for indoor localization. Since the distance between the Tango camera and the object can be collected by a Tango smartphone directly, the accuracy and success rate of camera pose tracking can be refined. The real scale of the indoor space depends on the direct depth information from the Tango sensor. Also, the direct method in regions with few texture features can generate the motion information of the Tango camera. Using a combination of these methods on visual localization and navigation in a complex space environment, the scale ambiguity issue caused by the monocular vision localization system can be well solved.

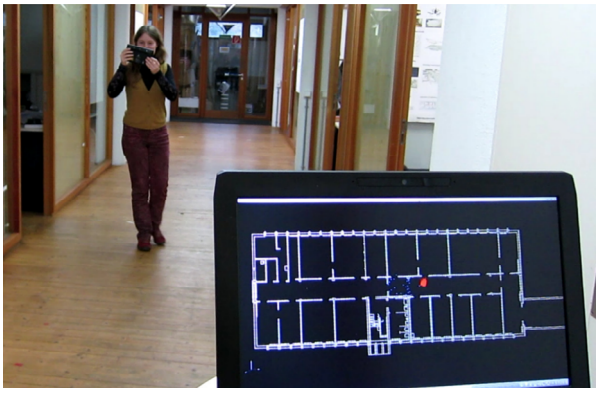

Fig. 3 Localization in positions with six degrees of freedom using only a 2D outline of the environment and an RGB-D Google Tango device
The computer screen shows the floor plan in white and the particle cloud represents the current pose estimate in red (Winterhalter et al., 2015)

4.2 Emergency evacuation

Emergencies such as fires, earthquakes, and terrorist attacks occur frequently. How to ensure safe and efficient evacuation of people in the case of an emergency is a key point in the field of emergency evacuation management. The research hotspot of indoor emergency evacuation is to use simulation technology to restore and quantify the layout of indoor space facilities, the topological structure of escape routes, and the evacuation management strategy, to verify the rationality of the design and management scheme for indoor public areas (Yuan et al., 2017, 2018, 2019). As the building structure is becoming more and more complex, the traditional plan of an evacuation schematic has been found difficult to meet the demand of emergency evacuation guidance. However, RGB-D devices can carry out rapid 3D mapping of indoor space and realize fine modeling of the physical space of an evacuation environment. This physical space model includes obstacle information, exit/channel location and obstacle structure information, firefighting equipment location information, and escape route topology information for emergency evacuation simulation. Thus, RGB-D devices also have a high application value in the field of emergency evacuation. Compared with traditional building design drawings (such as 2D computer-aided design drawings), the indoor evacuation space model based on RGB-D devices contains more diverse information and is closer to objective facts. RGB-D devices can obtain the

real-time locations of facilities, such as seats, desks, and fire-fighting equipment. Even the location and size information of idle items which have a significant impact on the evacuation process can be extracted. In addition, the indoor evacuation space model obtained by RGB-D devices has the location and structure information of the channel and exit, which can be used to identify the key nodes of the personnel escape route, and then the evacuation topology of personnel escape can be obtained.

Using 3D sensors such as RGB-D devices, we can obtain an excellent BIM of buildings. Based on that, some emergency evacuation research has been carried out (Rüppel and Schatz, 2011; Choi et al., 2014; Liu R et al., 2014; Chen AY and Huang, 2015; Santos et al., 2017). Chen AY and Chu (2016) studied the indoor emergency rescue routing. They combined network analysis and BIM, and used medial axis transform for graph construction from BIMs. Then a new path planning model was formed for rescue routes. Ma et al. (2017) built a management platform which contains full integration of BIM technology and virtual reality technology, and timely updates to the daily information. Ma et al. (2017) found that the BIM platform evacuation route information and fire equipment information are very intuitive, and the information transfer is significantly better than that in the 2D plan. Cheng JCP et al. (2018) proposed a simulation model for offshore oil and gas platforms to evaluate different evacuation plans to improve the evacuation performance by integrating the BIM technology and the agent-based model. In their work, some platform information was extracted from BIM and then can be used to model the evacuation environment by integrating the matrix and network models. Cheng JCP et al. (2018) found that the simulation model further improves the simulation performance and safety management on offshore oil and gas platforms. Liu HM et al. (2016) presented a robust keyframe-based monocular SLAM framework, which can reliably handle fast motion and strong rotation of RGB-D devices, ensuring good AR experience. This framework can be useful in emergency response of indoor evacuation. Zou et al. (2016) combined BIM and virtual reality technologies to research the emotional responses of pedestrians in the process of emergency evacuation when immersive virtual environments were constructed. Two immersive virtual environments were developed in Zou et al. (2016), both representing a fire emergency scenario in an apartment but having different levels of realism. Quantitative analysis of the possible negative emotions of the personnel was also made.

Next, we introduce the whole process of emergency evacuation simulation using the BIM acquired by RGB-D devices and give an example. Fig. 4 is a high-precision BIM of an indoor evacuation space, which can be quickly acquired by RGB-D devices. The model is based on an office scene of a research institute in Shenzhen, Guangdong Province, China. BIM contains the information of the initial location, channel, exit, and obstacle structures for emergency evacuation simulation.

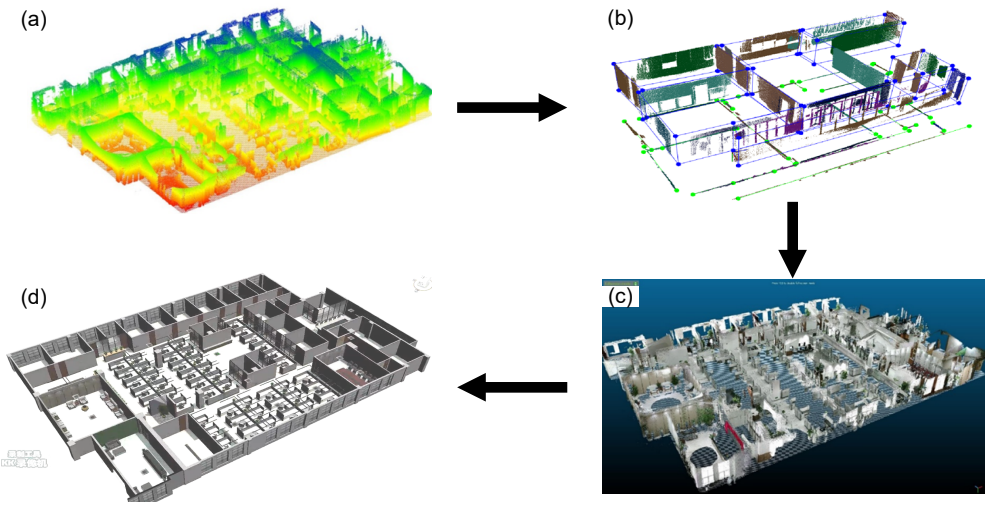

Fig. 4 BIM of the indoor evacuation space obtained by RGB-D devices: (a) fast acquisition of the three-dimensional point cloud in the evacuation scene; (b) automatic recognition of evacuation spatial structures; (c) high precision generation of BIM for the evacuation scene; (d) automatic texture mapping in the evacuation space (BIM: building information model)

As shown in Fig. 5, we imported the BIM acquired by RGB-D devices into the simulation system, and laid out the initial location, exit, navigation area, and other logical areas according to the information in BIM, to achieve the function of laying out pedestrians and key nodes of the evacuation path in simulation. The statistics of evacuation data are shown in Table 2. Then, the design of evacuation topology (blue lines in Fig. 5c) was implemented, and the evacuation path of the current evacuation topology was calculated according to Dijkstra's algorithm (green lines in Fig. 5c). Finally, the simulation of the pedestrian movement process was realized by combining the micro simulation models of pedestrian movement such as social force (Helbing et al., 2000), so that we can obtain pedestrian displacement and the speed data in the process of evacuation, and then analyze the evacuation data using the data we obtained.

Table 2 Statistics of evacuation data

Parameter	Value
Average evacuation time	33.07 s
Total evacuation time	64.5 s
Average evacuation speed	0.95 m/s
Maximum density	1.59 p/m^2
Maximum evacuation path length	62.89 m

5 Conclusions

In this paper, we have presented an extensive survey for indoor scene modeling and applications with RGB-D data. We briefly introduced some public RGB-D datasets and modeling technologies, and divided the technologies for 3D modeling using an RGB-D system into two categories: feature-based and dense styles. The feature-based RGB-D SLAM

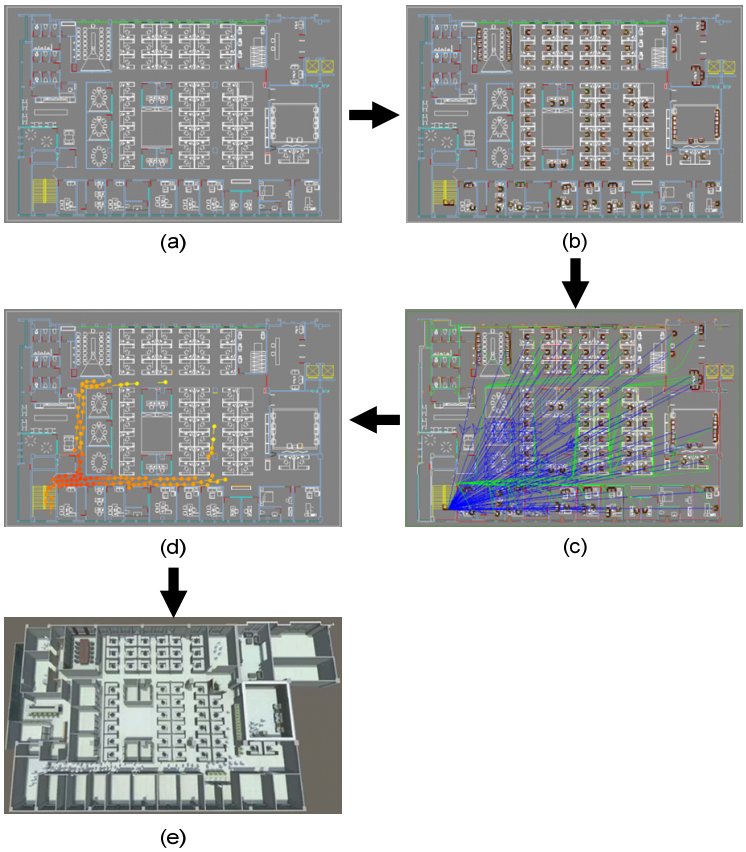

Fig. 5 Emergency evacuation simulation based on the BIM acquired by RGB-D devices: (a) evacuation space BIM imported into the simulation system; (b) layout of personnel location and export areas; (c) evacuation topology structure design and evacuation shortest path automatic generation; (d) visualization analysis of emergency evacuation based on the pedestrian micro simulation model (2D); (e) visualization analysis of emergency evacuation based on the pedestrian micro simulation model (3D)

BIM: building information model. References to color refer to the online version of this figure

system uses a few meaningful points for camera pose estimation, and thus has a lower computational cost than the dense SLAM system. The dense style modeling method uses mainly an ICP algorithm or a modified ICP algorithm for scene fusion, which is time-consuming and has high computing cost. After that, two typical applications including indoor localization and emergency evacuation based on RGB-D devices were introduced. By summarizing a broad spectrum of literature related to RGB-D devices, we hope this work gives some insights into this important topic.

Contributors

Zhilu YUAN designed the research. Shengjun TANG collected the information of RGB-D devices related products, technologies, and databases. You LI and Weixi WANG drafted the manuscript. Ming LI helped organize the manuscript. Renzhong GUO and Shengjun TANG revised and finalized the paper.

Compliance with ethics guidelines

Zhilu YUAN, You LI, Shengjun TANG, Ming LI, Renzhong GUO, and Weixi WANG declare that they have no conflict of interest.

References

Biswas J, Veloso M, 2010. WiFi localization and navigation for autonomous indoor mobile robots. IEEE Int Conf on robotics and automation, p.4379-4384. https://doi.org/10.1109/ROBOT.2010.5509842

Chen AY, Chu JC, 2016. TDVRP and BIM integrated approach for in-building emergency rescue routing. *J Comput Civil Eng*, 30(5):C4015003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000522

Chen AY, Huang T, 2015. Toward BIM-enabled decision making for in-building response missions. *IEEE Trans Intell Transp Syst*, 16(5):2765-2773. https://doi.org/10.1109/TITS.2015.2422138

Chen C, Yang BS, Song S, et al., 2018. Calibrate multiple consumer RGB-D cameras for low-cost and efficient 3D indoor mapping. *Remot Sens*, 10(2):328. https://doi.org/10.3390/rs10020328

Chen JW, Bautembach D, Izadi S, 2013. Scalable real-time volumetric surface reconstruction. *ACM Trans Graph*, 32(4):113. https://doi.org/10.1145/2461912.2461940

Chen K, Lai YK, Wu YX, et al., 2014. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Trans Graph*, 33(6): 208. https://doi.org/10.1145/2661229.2661239

Chen SN, Zheng YL, Luo PP, et al., 2017. Visual search based indoor localization in low light via RGB-D camera. *Int Comput Inform Eng*, 11(3):403-406. https://doi.org/10.5281/zenodo.1129950

Cheng JCP, Tan Y, Song YZ, et al., 2018. Developing an evacuation evaluation model for offshore oil and gas platforms using BIM and agent-based model. *Autom Constr*, 89:214-224. https://doi.org/10.1016/j.autcon.2018.02.011

Chintalapudi K, Padmanabha IA, Padmanabhan VN, 2010. Indoor localization without the pain. Proc 16th Annual Int Conf on Mobile Computing and Networking, p.173-184. https://doi.org/10.1145/1859995.1860016

Choi J, Choi J, Kim I, 2014. Development of BIM-based evacuation regulation checking system for high-rise and complex buildings. *Autom Constr*, 46:38-49. https://doi.org/10.1016/j.autcon.2013.12.005

Dai A, Chang AX, Savva M, et al., 2017. ScanNet: richly-annotated 3D reconstructions of indoor scenes. IEEE Conf on Computer Vision and Pattern Recognition, p.2432-2443. https://doi.org/0.1109/CVPR.2017.261

dos Santos DR, Basso MA, Khoshelham K, et al., 2016. Mapping indoor spaces by adaptive coarse-to-fine registration of RGB-D data. *IEEE Geosci Remot Sens Lett*, 13(2):262-266. https://doi.org/10.1109/LGRS.2015.2508880

Endres F, Hess J, Sturm J, et al., 2014. 3-D mapping with an RGB-D camera. *IEEE Trans Robot*, 30(1):177-187. https://doi.org/10.1109/TRO.2013.2279412.

Engelhard N, Endres F, Hess J, et al., 2011. Real-time 3D visual SLAM with a hand-held RGB-D camera. Proc RGB-D Workshop on 3D Perception in Robotics at the European Robotics Forum, 180:1-15.

Fehr M, Furrer F, Dryanovski I, et al., 2017. TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery. IEEE Int Conf on Robotics and Automation, p.5237-5244. https://doi.org/10.1109/ICRA.2017.7989614

Fox D, Burgard W, Thrun S, 1999. Markov localization for mobile robots in dynamic environments. *J Artif Intell Res*, 11:391-427. https://doi.org/10.1613/jair.616

Handa A, Whelan T, McDonald J, et al., 2014. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. IEEE Int Conf on Robotics and Automation, p.1524-1531. https://doi.org/10.1109/ICRA.2014.6907054

He FN, Habib A, 2018. Three-point-based solution for automated motion parameter estimation of a multi-camera indoor mapping system with planar motion constraint. *ISPRS J Photogr Remot Sens*, 142:278-291. https://doi.org/10.1016/j.isprsjprs.2018.06.011

Helbing D, Farkas I, Vicsek T, 2000. Simulating dynamical features of escape panic. *Nature*, 407(6803):487-490. https://doi.org/10.1038/35035023

Henry P, Krainin M, Herbst E, et al., 2012. RGB-D mapping: using Kinect-style depth cameras for dense 3D modeling of indoor environments. *Int J Robot Res*, 31(5):647-663. https://doi.org/10.1177/0278364911434148

Huang AS, Bachrach A, Henry P, et al., 2017. Visual odometry and mapping for autonomous flight using an RGB-D camera. In: Christensen HI, Khatib O (Eds.), Robotics

Research. Springer, Cham, p.235-252. https://doi.org/10.1007/978-3-319-29363-9_14

Ikehata S, Yang H, Furukawa Y, 2015. Structured indoor modeling. IEEE Int Conf on Computer Vision, p.1323-1331. https://doi.org/10.1109/ICCV.2015.156

Kerl C, Sturm J, Cremers D, 2013. Robust odometry estimation for RGB-D cameras. IEEE Int Conf on Robotics and Automation, p.3748-3754. https://doi.org/10.1109/ICRA.2013.6631104

Konolige K, Agrawal M, 2008. FrameSLAM: from bundle adjustment to real-time visual mapping. *IEEE Trans Robot*, 24(5):1066-1077. https://doi.org/10.1109/tro.2008.2004832

Lai K, Bo L, Fox D, 2014. Unsupervised feature learning for 3D scene labeling. IEEE Int Conf on Robotics and Automation, p.3050-3057. https://doi.org/10.1109/ICRA.2014.6907298

Li M, Chen RZ, Liao X, et al., 2018. A real-time indoor visual localization and navigation method based on Tango smartphone. Ubiquitous Positioning Indoor Navigation and Location Based Service, p.1-6. https://doi.org/10.1109/UPINLBS.2018.8559720

Liu HM, Zhang GF, Bao HJ, 2016. Robust keyframe-based monocular SLAM for augmented reality. IEEE Int Symp on Mixed and Augmented Reality, p.1-10. https://doi.org/10.1109/ISMAR-Adjunct.2016.0111

Liu R, Du J, Issa RRA, 2014. Human library for emergency evacuation in BIM-based serious game environment. Int Conf on Computing in Civil and Building Engineering, p.544-551. https://doi.org/10.1061/9780784413616.068

Ma J, Jia W, Zhang J, 2017. Research of building evacuation path to guide based on BIM. 29th Chinese Control and Decision Conf, p.1814-1818. https://doi.org/10.1109/CCDC.2017.7978811

Mur-Artal R, Tardós JD, 2017. ORB-SLAM2: an open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Trans Robot*, 33(5):1255-1262. https://doi.org/10.1109/TRO.2017.2705103

Newcombe RA, Izadi S, Hilliges O, et al., 2011. KinectFusion: real-time dense surface mapping and tracking. 10th IEEE Int Symp on Mixed and Augmented Reality, p.127-136. https://doi.org/10.1109/ISMAR.2011.6092378

Nießner M, Zollhöfer M, Izadi S, et al., 2013. Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans Graph*, 32(6):169. https://doi.org/10.1145/2508363.2508374

Ortiz LE, Cabrera VE, Gonçalves LMG, 2018. Depth data error modeling of the ZED 3D vision sensor from stereolabs. *ELCVIA Electron Lett Comput Vis Imag Anal*, 17(1):1-15. https://doi.org/10.5565/rev/elcvia.1084

Rüppel U, Schatz K, 2011. Designing a BIM-based serious game for fire safety evacuation simulations. *Adv Eng Inform*, 25(4):600-611. https://doi.org/10.1016/j.aei.2011.08.001

Santos R, Costa AA, Grilo A, 2017. Bibliometric analysis and review of building information modelling literature published between 2005 and 2015. *Autom Constr*, 80:118-136. https://doi.org/10.1016/j.autcon.2017.03.005

Segal A, Haehnel D, Thrun S, 2009. Generalized-ICP. Proc Robotics: Science and Systems, 2:4. https://doi.org/10.15607/RSS.2009.V.021

Shi Y, Xu K, Nießner M, et al., 2018. PlaneMatch: patch coplanarity prediction for robust RGB-D reconstruction. *LNCS*, 11212:767-784. https://doi.org/10.1007/978-3-030-01237-3_46

Song S, Lichtenberg SP, Xiao J, 2015. SUN RGB-D: a RGB-D scene understanding benchmark suite. IEEE Conf on Computer Vision and Pattern Recognition, p.567-576. https://doi.org/10.1109/CVPR.2015.7298655

Steinbrucker F, Sturm J, Cremers D, 2011. Real-time visual odometry from dense RGB-D images. IEEE Int Conf on Computer Vision Workshops, p.719-722. https://doi.org/10.1109/ICCVW.2011.6130321

Sturm J, Engelhard N, Endres F, et al., 2012. A benchmark for the evaluation of RGB-D SLAM systems. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.573-580. https://doi.org/10.1109/IROS.2012.6385773

Tang SJ, Li Y, Yuan ZL, et al., 2019a. A vertex-to-edge weighted closed-form method for dense RGB-D indoor SLAM. *IEEE Access*, 7:32019-32029. https://doi.org/10.1109/ACCESS.2019.2900990

Tang SJ, Zhang YJ, Li Y, et al., 2019b. Fast and automatic reconstruction of semantically rich 3D indoor maps from low-quality RGB-D sequences. *Sensors*, 19(3):533. https://doi.org/10.3390/s19030533

Thomas D, Sugimoto A, 2017. Modeling large-scale indoor scenes with rigid fragments using RGB-D cameras. *Comput Vis Imag Underst*, 157:103-116. https://doi.org/10.1016/j.cviu.2016.11.008

Vestena KM, dos Santos DR, Oilveira EMJr, et al., 2016. A weighted closed-form solution for RGB-D data registration. *Int Arch Photogr Remote Sens Spat Inform Sci*, XLI-B3:403-409. https://doi.org/10.5194/isprsarchives-XLI-B3-403-2016

Wang KK, Zhang GF, Bao HJ, 2014. Robust 3D reconstruction with an RGB-D camera. *IEEE Trans Imag Process*, 23(11):4893-4906. https://doi.org/10.1109/TIP.2014.2352851

Westoby MJ, Brasington J, Glasser NF, et al., 2012. 'Structure-from-motion' photogrammetry: a low-cost, effective tool for geoscience applications. *Geomorphology*, 179:300-314. https://doi.org/10.1016/j.geomorph.2012.08.021

Whelan T, Michael K, Maurice F, et al., 2013. Kintinuous: Spatially Extended Kinectfusion. Technical Report No. MIT-CSAIL-TR-2012-020. https://doi.org/1721.1/71756

Winterhalter W, Fleckenstein F, Steder B, et al., 2015. Accurate indoor localization for RGB-D smartphones and tablets given 2D floor plans. IEEE/RSJ Int Conf on Intelligent Robots and Systems, p.3138-3143. https://doi.org/10.1109/IROS.2015.7353811

Yuan ZL, Jia HF, Liao MJ, et al., 2017. Simulation model of self-organizing pedestrian movement considering following behavior. *Front Inform Technol Electron Eng*, 18(8):1142-1150. https://doi.org/10.1631/FITEE.1601592

Yuan ZL, Jia HF, Zhang LF, et al., 2018. A social force evacuation model considering the effect of emergency signs. *Simulation*, 94(8):723-737. https://doi.org/10.1177/0037549717741350

Yuan ZL, Guo RZ, Tang SJ, et al., 2019. Simulation of the separating crowd behavior in a T-shaped channel based on the social force model. *IEEE Access*, 7:13668-13682. https://doi.org/10.1109/ACCESS.2019.2894345

Zeng M, Zhao FK, Zheng JX, et al., 2012. A memory-efficient KinectFusion using octree. In: Hu SM, Martin RR (Eds.), Computational Visual Media. Springer, Berlin, p.234-241. https://doi.org/10.1007/978-3-642-34263-9_30

Zou H, Li N, Cao L, 2016. Immersive virtual environments for investigating building emergency evacuation behaviors: a feasibility study. 33th Int Symp on Automation and Robotics in Construction, p.1-8. https://doi.org/10.22260/ISARC2016/0040