Computational Visual Media
DOI 10.1007/s41095-015-0029-x Vol. 1, No. 4, December 2015, 267–278

Review Article

3D indoor scene modeling from RGB-D data: a survey

Kang Chen[1], **Yu-Kun Lai**[2], **and Shi-Min Hu**[1](✉)

Abstract 3D scene modeling has long been a fundamental problem in computer graphics and computer vision. With the popularity of consumer-level RGB-D cameras, there is a growing interest in digitizing real-world indoor 3D scenes. However, modeling indoor 3D scenes remains a challenging problem because of the complex structure of interior objects and poor quality of RGB-D data acquired by consumer-level sensors. Various methods have been proposed to tackle these challenges. In this survey, we provide an overview of recent advances in indoor scene modeling techniques, as well as public datasets and code libraries which can facilitate experiments and evaluation.

Keywords RGB-D camera; 3D indoor scenes; geometric modeling; semantic modeling; survey

1 Introduction

Consumer-level color and depth (RGB-D) cameras (e.g., Microsoft Kinect) are now widely available and are affordable to the general public. Ordinary people can now easily obtain 3D data from their real-world homes and offices. Meanwhile, other booming 3D technologies in areas such as augmented reality, stereoscopic movies, and 3D printing are also becoming closer to our daily life. We are living on a "digital Earth". Therefore, there is an ever-increasing need for ordinary people to digitize their living environments.

Despite this great need, helping ordinary people quickly and easily acquire 3D digital representations of their living surroundings is an urgent yet still challenging research problem. Over the past decades, we have witnessed an explosion of digital photos on the Internet. Benefiting from this, image-related research based on mining and analyzing the vast number of 2D images has been greatly boosted. In contrast, while the growth of 3D digital models has accelerated over the past few years, the growth remains comparatively slow, mainly because making 3D models is a demanding job which requires expertise and is time-consuming. Fortunately, the availability of low-cost RGB-D cameras along with recent advances in modeling techniques offers a great opportunity to change this situation. In the longer term, 3D big data has the potential to change the landscape of 3D visual data processing.

This survey focuses on digitizing real-world indoor scenes, which has received significant interest in recent years. It has many applications which may fundamentally change our daily life. For example, with such techniques, furniture stores can offer 3D models of their products online so that customers can better view the products and choose furniture to buy. People without interior design experience can give digital representations of their homes to experts or expert systems [1, 2] for advice on better furniture arrangement. Anyone with Internet access can virtually visit digitized museums all over the world [3]. Moreover, modeled indoor scenes can be used for augmented reality [4, 5] and can serve as a training basis for intelligent robots to better understand real-world environments [6].

Nevertheless, indoor scene modeling is still a challenging problem. The difficulties mainly arise from two causes [7]: Firstly, unlike outdoor

1 Tsinghua University, Beijing 100084, China. E-mail: K. Chen, chenkangnobel@gmail.com; S.-M. Hu, shimin@tsinghua.edu.cn (✉).

2 Cardiff University, Cardiff, CF24 3AA, Wales, UK. E-mail: Yukun.Lai@cs.cardiff.ac.uk.

Manuscript received: 2015-10-09; accepted: 2015-11-19

building facades, interior objects often have much more complicated 3D geometry, with messy surroundings and substantial variation between parts. Secondly, depth information captured by consumer-level scanning devices is often noisy, may be distorted, and can have large gaps. To address these challenges, various methods have been proposed in the past few years and this is still an active research area in both computer graphics and computer vision communities.

The rest of the paper will be organized as follows. We first briefly introduce in Section 2 different types of RGB-D data and their properties. Publicly available RGB-D datasets as well as useful programming libraries for processing RGB-D data will also be discussed. In Section 3, we systematically categorize existing methods based on their underlying design principles, overview each technique, and examine its advantages and disadvantages. Finally, in Section 4, we summarize the current state of the art and elaborate on future research directions.

2 RGB-D data

"One cannot make bricks without straw." Despite the importance of indoor scene modeling and the fact that RGB-D scanners have been available for decades, it did not become a research focus until the year 2010 when Microsoft launched its Kinect motion sensing input device. Kinect has a more far-reaching significance than as the game controller it was originally released for, because it has a built-in depth sensor with reasonable accuracy at a very affordable price. Such cheap RGB-D scanning devices make it possible for ordinary people to own one at home, enabling development and wide use of 3D modeling techniques for indoor scene modeling. Before discussing modeling algorithms in detail, we first briefly introduce RGB-D data in this section, including different types of RGB-D data and their properties.

2.1 Types and properties

A variety of techniques have been developed to obtain RGB-D data. These include passive techniques such as stereoscopic camera pairs where the depth is derived from disparity between images captured from each camera, and active techniques where some kind of light is emitted to assist depth calculation. The latter are widely used due to their effectiveness (e.g., particularly for textureless surfaces) and accuracy. Currently, light detection and ranging (LiDAR) is the main modality for acquiring RGB-D data. Depending on their working approach, LiDAR systems can be divided into two classes: scannerless LiDAR and scanning LiDAR [8]. In scannerless LiDAR systems, the entire scene is captured with each laser or light pulse, as opposed to point-by-point capture with a laser beam in scanning LiDAR systems. A typical type of scannerless LiDAR system is the time-of-flight (ToF) camera, used in many consumer-level RGB-D cameras (including the latest Kinect v2). ToF cameras are low-cost, quick enough for real-time applications, and have moderate working ranges. These advantages make ToF cameras suitable for indoor applications. Alternatively, some RGB-D cameras, including the first generation of Kinect, are based on structured light. The depth is recovered by projecting specific patterns and analyzing the captured patterned image. Both ToF and structured light techniques are scannerless, so they can produce dynamic 3D streams, which allow more efficient and reliable 3D indoor scene modeling.

Laser pulses in a ToF camera and patterns used for structured light cameras are organized in a 2D array, so that depth information can be represented as a depth image. The depth image along with an aligned RGB image forms an RGB-D image frame which depicts a single view of the target scene, including both the color and the shape. Such RGB-D image frames can be unprojected to 3D space forming a colored 3D point cloud. RGB-D images and colored point clouds are the two most common representations of RGB-D data. RGB-D images are mostly used by the computer vision community as they have the same topology as images, while in the computer graphics community, RGB-D data are more commonly viewed as point clouds. Point clouds obtained from a projective camera are *organized* (also called *structured* or *ordered*) point clouds because there is a one–one correspondence between points in the 3D space and pixels in the image space. This

correspondence contains adjacency information between 3D points which is useful in certain applications, e.g., it can simplify algorithms or make algorithms more efficient as neighboring points can be easily determined. Knowing the camera parameters, organized colored point clouds, and the corresponding RGB-D images are equivalent. If an equivalent RGB-D image does not exist for a colored point cloud, then the point cloud is unorganized (unstructured, unordered). To fully depict a target scene, multiple RGB-D image frames captured from different views are typically needed. As scannerless cameras are usually used, scene RGB-D data captured are essentially RGB-D image streams (sequences) which can later be stitched into a whole scene point cloud using 3D registration techniques.

Depending on the operational mechanism, LiDAR systems cannot capture depth information on surfaces with highly absorptive or reflective materials. However, such materials are very common in real-world indoor scenes, and are used as mirrors, window glass, TV screens, and steel surfaces etc. This is a fundamental limitation of all laser-based systems. Apart from this common limitation, consumer-level RGB-D cameras have other drawbacks caused by their low cost. Firstly, the spatial resolution of such cameras is generally low (512 × 484 pixels in the latest Kinect). Secondly, the depth information is noisy and often has significant camera distortion. Thirdly, even for scenes without absorptive or reflective materials, the depth image may still involve small gaps around object borders. In general, depth information obtained by cheap scanning devices is unreliable, and practical indoor scene modeling algorithms must take this fact into consideration.

2.2 Public datasets

A number of public RGB-D datasets containing indoor scenes have been introduced in recent years. Although most of these datasets were built and labeled for specific applications, such as scene reconstruction, object detection and recognition, scene understanding and segmentation, etc., as long as they provide full RGB-D image streams of indoor scenes, they can be used as input for indoor scene modeling. Here we briefly describe some popular ones (example scenes from each dataset are shown in Fig. 1).

Cornell RGB-D Dataset [9, 10]: this dataset contains RGB-D data of 24 office scenes and 28 home scenes, all of which were captured by Kinect. RGB-D images of each scene are stitched into scene point clouds using an RGB-D SLAM algorithm. Object-level labels are provided on the stitched scene point clouds.

Washington RGB-D Scenes Dataset [11]: this dataset consists of 14 indoor scenes containing objects in 9 categories (chair, coffee table, sofa,

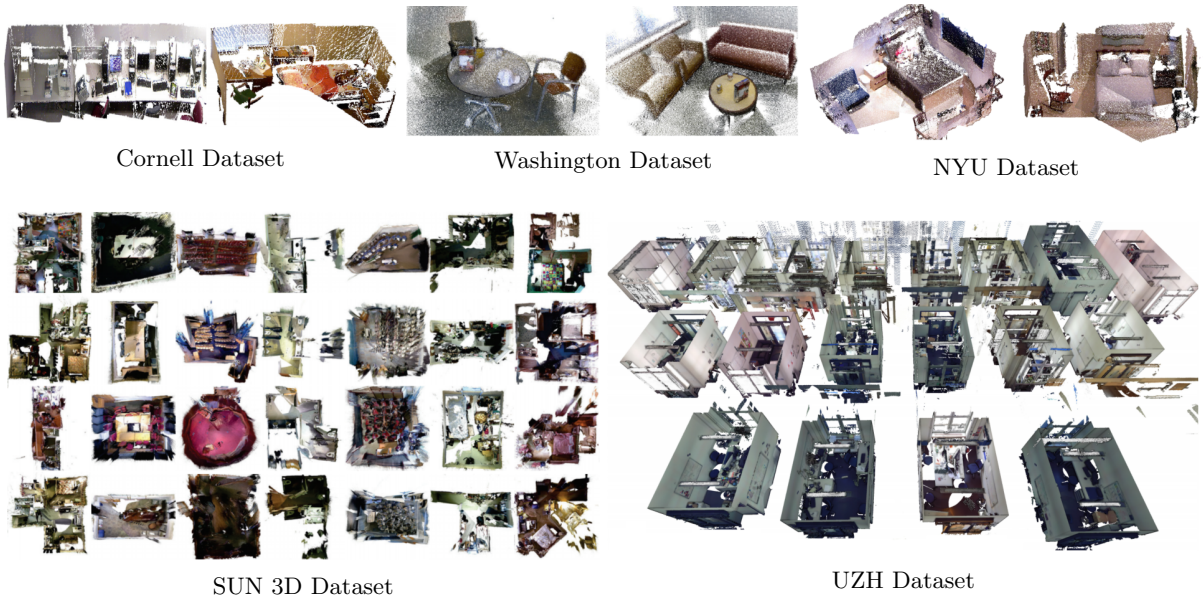

Fig. 1 Example RGB-D data in each public dataset.

table, bowl, cap, cereal box, coffee mug, and soda can). Each scene is a point cloud created by aligning a set of Kinect RGB-D image frames using patch volume mapping. Labels for the background and the 9 object classes are given on the stitched scene point clouds.

NYU Depth Dataset [12, 13]: this dataset contains 528 different indoor scenes (64 in the first version [12] and 464 in the second [13]) captured from large US cities, using Kinect. The scenes are mainly inside residential apartments, including living rooms, bedrooms, bathrooms, and kitchens. Dense labeling of objects at the class and instance level is provided for 1449 selected frames. This dataset does not contain camera pose information, because it was mainly built for single-frame segmentation and object recognition. To get full 3D scene point clouds, users may need to estimate camera poses from the original RGB-D streams.

SUN 3D Dataset [14]: this dataset contains 415 RGB-D image sequences captured by Kinect from 254 different indoor scenes, in 41 different buildings across North America, Europe, and Asia. Semantic class polygons and instance labels are given on frames and propagated through the whole sequences. Camera pose for each frame is also provided for registration. This is currently the largest and most comprehensive RGB-D dataset of indoor scenes.

UZH Dataset [15]: unlike other datasets mentioned above, this dataset was built specifically for modeling. It contains full point clouds of 40 academic offices scanned by a Faro LiDAR scanner, which has much higher precision than consumer-level cameras like Kinect but is also much more expensive.

2.3 Open source libraries

Since the release of the Kinect and other consumer-level RGB-D cameras, RGB-D data has become popular. Publicly available libraries that support effective processing of RGB-D data is thus in demand. The Point Cloud Library (PCL) [16] was introduced in 2011, which is an open source library for 2D/3D image and point cloud processing. The PCL framework contains numerous implementations of state-of-the-art algorithms including filtering, feature estimation, surface reconstruction, registration, model fitting and segmentation. Due to its powerful features and relaxed BSD license (Berkeley Software Distribution), it is probably the most popular library for RGB-D data processing for both commercial and research use.

Another useful library is the Mobile Robot Programming Toolkit (MRPT) [17] which comprises a set of C++ libraries and a number of ready-to-use robot-related applications. RGB-D sensors can be effectively used as "eyes" for robots: understanding real-world environments through perceived RGB-D data is one of the core functions of intelligent robotics. This library contains state-of-the-art algorithms for processing RGB-D data with a focus on robotic applications, including SLAM (simultaneous localization and mapping) and object detection.

3 Modeling techniques

After introducing RGB-D data, we now discuss various techniques for modeling indoor scenes in this section. Based on modeling purpose, these methods can generally be classified into two main categories: geometric modeling (Section 3.1) and semantic modeling (Section 3.2) approaches. The former aims to recover the shapes of the 3D objects in the scene, whereas the latter focuses on recovering semantic meaning (e.g., object types).

3.1 Geometric modeling

Geometric modeling from RGB-D data is a fundamental problem in computer graphics. Ever since the 1990s, researchers have investigated methods for digitizing the shapes of 3D objects using laser scanners, although 3D scanners were hardly accessible to ordinary people until recently. Early works typically start by registering a set of RGB-D images captured by laser sensors (i.e., transforming RGB-D images into a global coordinate system) and fuse the aligned RGB-D frames into a single point cloud or a volumetric representation which can be further converted into mesh-based 3D models. The use of the volumetric representation ensures the resulting geometry is a topologically correct manifold. Figure 2 is a typical geometric modeling result. Based on this pipeline, geometric modeling problems can be split into two phases: registration and fusion. Much research

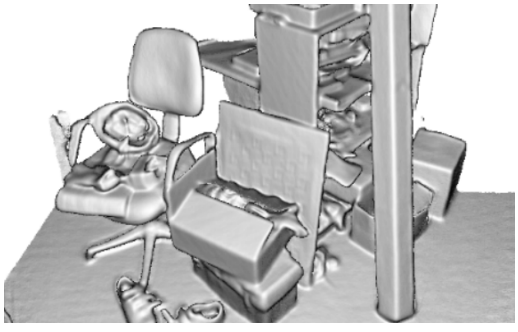

Fig. 2 Geometric modeling result. Reproduced with permission from Ref. [5], © 2011 IEEE.

has been done and theoretically sound approaches have been established for both phases. For the registration phase, iterative closest point (ICP) registration [18, 19] and simultaneous localization and mapping (SLAM) [20] as well as their variants generally produce good solutions. For the fusion phase, the most widely adopted solution is the volumetric technique proposed by Curless and Levoy [21] which can robustly integrate each frame using signed distance functions (SDFs).

Geometric indoor scene modeling methods are extensions of traditional registration and fusion algorithms to indoor scenes. The major difference is that such techniques must take into account the properties of RGB-D data captured by consumer-level RGB-D cameras, namely low-quality and real-time sequences. A well-known technique is the Kinect Fusion system [4, 5] which provides level-of-detail (LoD) scanning and model creation using a moving Kinect camera. As in traditional schemes, Kinect Fusion adopts a volumetric representation of the acquired scene by maintaining a signed distance value for each voxel grid in the memory. However, unlike traditional frame-to-frame registration, each frame is registered to the whole constructed scene model rather than the previous frames using a coarse-to-fine iterative ICP algorithm. This frame-to-model registration scheme has more resistance to noise and camera distortion, and is sufficiently efficient to allow real-time applications. The system has many desirable characteristics: ease of use, real-time performance, LoD reconstruction, etc. Recently, Heredia and Favier [22] have further extended the basic Kinect Fusion framework to larger scale environments by use of volume shifting. However, when used as a modeling system for indoor scene modeling, the volumetric representation based mechanism significantly limits its usage for large and complex scenes due to several reasons. Reconstructing large scale scenes even with a moderate resolution to depict necessary details requires a large amount of memory, easily exceeding the memory capacity of ordinary computers. Moreover, acquisition and registration errors inevitably exist, and can be significant for consumer-level scanning devices. Although frame-to-model registration is more robust than frame-to-frame registration, it is still not a global optimization technique. Scanning larger scenes requires longer moving trajectories. Error keeps accumulating over the long acquisition process and eventually breaks the reconstruction. A typical example is the loop closure problem which causes misalignment when reconstructing large rooms using Kinect Fusion when the camera trajectory forms a closed loop.

Kinect Fusion is designed for real-time online modeling and interaction within relatively small environments. A more general modeling framework is the RGB-D SLAM [23]. As mentioned before, the depth information obtained by cheap scanning devices is unreliable. However, the aligned RGB images can provide important additional information when estimating camera poses. The appearance features from the RGB image and shape features from the depth image can complement each other and together provide much more robust point correspondences between frames. In addition, in a practical scanning process, it is very common to have loop closures in the camera trajectories. Thus, overlaps may exist not only between consecutive frames. Loop closures can be detected and spatial relationship between the corresponding frames offers additional constraints when computing camera poses. The whole sequence of RGB-D frames can be represented as a graph, where each node is a frame and each edge stores the spatial transform between two adjacent nodes. Such graphs are called pose graphs and can be efficiently optimized using SLAM algorithms [20] (see Ref. [24] for various state-of-the-art SLAM algorithms). The general pipeline of the RGB-D SLAM framework is shown in Fig. 3.

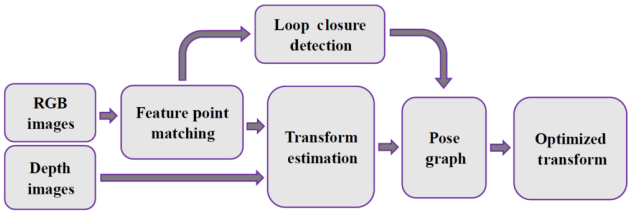

Fig. 3 Pipeline of the RGB-D SLAM framework.

RGB-D SLAM approaches can be divided into two types: sparse mapping and dense mapping. For sparse mapping, only a few sparsely selected key frames are used for reconstruction which can quickly provide a rough structure of the target scene, while for dense mapping, the whole RGB-D stream is used, which can give detailed reconstruction as long as sufficient data is available. In both cases, the key technique is feature point matching, which is the basis for both transform estimation and loop closure detection. Due to the poor quality of depth images obtained by low-cost scanning devices, most sparse mapping systems mainly rely on distinctive feature descriptors detected in RGB images (e.g., SIFT [25], SURF [26], or ORB [27]) to find corresponding point pairs [23]. As real-world indoor scenes usually contain large textureless areas, e.g., painted walls, or repeated patterns, e.g., tiled floors, even state-of-the-art feature descriptors may easily generate falsely matched point correspondences. To reduce the impact of falsely detected point correspondences on reconstruction, the RANSAC (RANdom SAmple Consensus) algorithm [28] is often adopted to determine a subset of correspondences which conform to a consistent rigid transform. RANSAC is an iterative, randomized approach to estimate parameters of a mathematical model (in this case a rigid transform) that fits observed data (sample points) and is robust to outliers (which often occurs in low-quality RGB-D data) [29]. However, this may still fail in challenging cases as repetitive objects or large textureless areas may easily lead to many false correspondences. In practice, manual correction of some falsely estimated transforms is often needed in sparse mapping applications [7]. In contrast, with the help of dense depth streams, a frame-to-frame ICP registration algorithm can provide stronger cues for inferring camera poses. Thus, dense mapping RGB-D SLAM systems [23, 30–32] currently provide more automatic and robust solutions to modeling indoor scenes with consumer-level RGB-D sensors.

3.2 Semantic modeling

The main objective of geometric modeling of indoor scenes is to fully recover 3D geometry. These methods take the target scene as a whole regardless of what it contains, and thus cannot provide a semantic representation of the modeled scene. However, semantic information is of vital importance in modeling for the following reasons. Firstly, semantic information can be used to improve modeling results. For example, in cluttered real-world indoor scenes, it is not practically possible to capture every single corner of the scene due to occlusion. Nevertheless, with simple semantic knowledge, e.g., that desk surfaces are horizontal planes, and chairs have mirror symmetry, we can easily infer the occluded structure. Secondly, semantic representation of the modeled scene is required by higher-level applications. For instance, to understand and interact with the modeled digital scenes, a semantic tag for each object or even part must be known. In fact, for many higher-level applications it can be preferable to lose some geometric precision in exchange for a semantically correct representation, as long as doing so does not lead to confusion in understanding the scene contents. In this spirit, growing attention has been paid recently to semantic modeling methods.

Semantic modeling algorithms focus on reconstructing scenes down to the level of specific objects. Typically, RGB-D data of each semantic region are separated from the surrounding environment and fitted using either existing object models, part models, or even geometric primitives (e.g., planes or cylinders). Semantic modeling has many advantages compared to geometric modeling. Apart from producing a semantically meaningful representation of the modeled scene (e.g., knowledge that the scene contains a table and four chairs) which is beneficial in many applications, the modeling process is much simpler compared to traditional geometric modeling which needs extensive effort for data acquisition, especially when capturing real-world indoor

scenes with low-cost RGB-D sensors. In contrast, semantic modeling systems typically only require sparse RGB-D images because the basic shapes of most interior objects are already known a priori. Hence, semantic modeling techniques are particularly suited to modeling real-world indoor scenes from low-quality RGB-D data. The general pipeline of the semantic modeling framework is shown in Fig. 4.

Clearly, semantic modeling requires sound semantic segmentation of the input RGB-D data. Automatically separating an indoor scene into different kinds of semantic regions is a challenging problem. On one hand, to understand what objects are present in the scene, each object must be separated from its surroundings. On the other hand, recognizing the type and shape of an object is ultimately important for determining whether an adjacent region belongs to the object or not, for effective segmentation. This is an intricate chicken-and-egg problem. To break the interdependency, human prior knowledge is often adopted in the form of semantic or contextual rules. Although many algorithms claim to take advantage of using semantic or contextual information, there are significant differences in terms of what they mean by semantic or contextual information. This is mainly because there is no universal definition of what degree of human prior knowledge can be considered as semantic. Therefore, based on the level of context being used, we classify semantic modeling methods into two categories: primitive-based methods (Section 3.2.1) and model-based methods (Section 3.2.2).

3.2.1 Primitive-based methods

An important observation concerning interior objects is that most of them can be decomposed into a set of geometric primitives (e.g., sphere, cone, plane, and cylinder). Figure 5 gives an example of a semantically modeled scene; note that objects in it are all constructed from basic geometric primitives. Finding proper primitives which best fit the unsegmented noisy input RGB-D scan is the core of primitive-based methods. Thus, primitive fitting algorithms must be capable of reliably distinguishing between inliers and outliers. The state-of-the-art algorithm is based on RANSAC [28] due to its robustness to outliers. However, RANSAC can only estimate one model (i.e., a single primitive) for a particular data set. As for any one-model approach, when two (or more) instances exist, RANSAC may fail to find either one. As an alternative, the Hough transform [33] is often used for robust estimation of models when more than one model instance is present; it finds instances of objects within a certain class of shapes by voting in the parameter space. A major drawback of the Hough transform is that the time complexity increases at a rate of $O(A^{m-2})$ where A is the size of the input data and m is the number of parameters [34]. Thus, it is very time-consuming to detect complex models (large m) in large-scale input scans (large A). Furthermore, the Hough transform is generally more sensitive to noise than RANSAC. As a result, the Hough transform is most often used when we can convert the problem into a 2D parameter space to make the problem tractable [3, 35]. Otherwise, most approaches choose to detect multiple model instances one by one using RANSAC [36–38]. As the primitives are fitted locally from noisy and incomplete data, it is very common to see duplicated primitives where a single primitive is

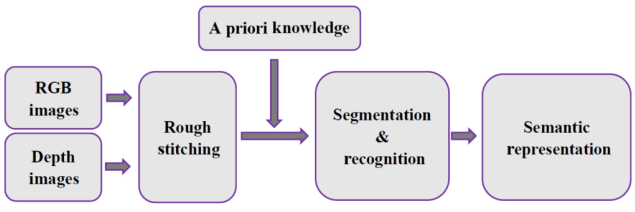

Fig. 4 Pipeline of the semantic modeling framework.

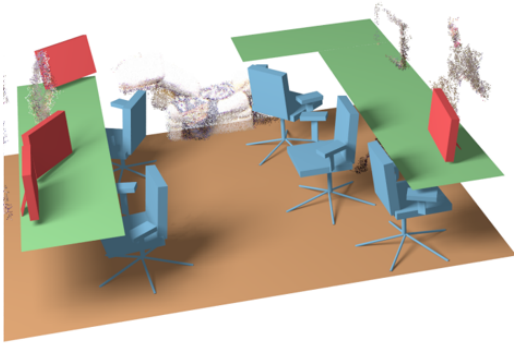

Fig. 5 Primitive-based semantic modeling result. Reproduced with permission from Ref. [39], © 2012 Association for Computing Machinery, Inc.

reported multiple times, or gaps in the set of fitted primitives. Thus global consolidation is also needed to get a consistent scene representation. Depending on the application, different types of global consolidation are used with different a priori assumptions (e.g., regarding typical primitives in the scene).

In large-scale interior architectural modeling, a box assumption is most commonly used. Specifically, walls can be fitted with vertical planar primitives, floors and ceilings can be fitted with horizontal planar primitives, and together they form a strict box. This assumption is the foundation of the state-of-the-art architectural modeling approach [3]. It first segments the scene point cloud into a set of horizontal 2D slices, and points in each slice are projected onto a 2D plane. Line segments are detected in the 2D space, which are then merged into 2D rectangles and combined with other slices to form 3D cuboids. In some cases, convex hull or alpha-shape algorithms are also needed to determine the spatial extent of each planar primitive [36], as they may form general polygonal shapes rather than the more common rectangles.

Many CAD and mechanical models are designed and manufactured by additive or subtractive combination of primitive shapes. Such types of objects can be naturally modeled by primitive-based methods with suitable assumptions. The state-of-the-art method for modeling such objects is proposed by Li et al. [37]. They consider three types of mutual relations (orientation, placement, and equality) and propose an iterative constrained optimization scheme to globally consolidate locally fitted primitives.

Recently, primitive-based methods have been extended to model large-scale indoor scenes, not only for walls or floors but for interior furniture as well. This is based on the observation that furniture items (e.g., chairs, tables, and cabinets) in a large-scale scene usually come from a small number of prototypes and are repeated multiple times. Kim et al. [39] proposed a supervised method which involves two stages. In the offline learning stage, each object of interest is pre-scanned and represented as a set of stable primitives along with necessary inter-part junction attributes. In the online modeling stage, the whole scene is segmented and each segment is fitted with primitives. Then all repeated objects are detected and modeled through hierarchical matching. Variation between object parts can also be handled by specifying degree-of-freedom for each stable primitive in the pre-scanned object, which is the main advantage of this supervised method. Mattausch et al. [15] later introduced an unsupervised method for modeling with high-quality RGB-D data also by detecting repeated objects. They first convert the scene point cloud into a collection of nearly-planar patch primitives. Then, based on geometric similarity and spatial configurations of neighboring patches, patches are clustered in a Euclidean embedding space and repeated objects can thus be detected and modeled. Note that primitives used in these methods are not just meaningless geometric shapes but some kind of semantic abstraction of interior objects or parts, identified from repeated occurrences of instances in the training data, which helps to robustly recover repeated objects from incomplete and noisy data (e.g., chair backs, chair seats, monitors, etc.).

3.2.2 Model-based methods

Despite attempts with certain levels of success as described in the previous subsection, primitive-based methods have fundamental limitations in modeling interior objects. For example, both Refs. [39] and [15] only tackle large-scale public or office buildings with many repeated objects, but in typical home environments many objects only occur once (e.g., a television or a bed). Moreover, many interior objects (e.g., keyboards, desk lamps, and various types of chairs) are too complex to be depicted in detail using a set of simple primitives. Thus, primitive-based methods can only offer an approximation to the target scene.

What happens if we already have a database containing similar 3D models of objects to the ones that appear in the target scene? This is not unrealistic, as for example chairs frequently occur in indoor scenes and it is likely that a chair model similar to the one appearing in the target scene already exists in the database. In this case we no longer need to pre-scan the chair [39] or cluster point cloud regions [15] to learn the underlying

semantic structural knowledge of that chair. More importantly, 3D models are far more flexible and can provide more accurate depiction for chairs in the scene than a set of basic primitives. As long as we have sufficient 3D models in the database, it is much more feasible to get a visually plausible and semantically segmented digital scene by finding and placing correct models and parts, adapted to fit the input scans (see Fig. 6). This is the key spirit of model-based methods. The growing availability of free 3D models online (e.g., in the Trimble 3D Warehouse) has made it possible. Model-based methods thus represent a new trend in scene modeling.

Nan et al. [40] use a search-classify strategy and a region growing method to find independent point clouds from high-quality laser scans, and assign a semantic label for each meaningful object. They first train classifiers for individual pre-defined object categories. In the online stage, they first over-segment the input point cloud. Starting from a seed region in the over-segmentation, the point cloud of an individual object is detected and separated from the background by iteratively adding regions which help to increase classification confidence. After that, a deform-to-fit technique is used to adapt 3D models in the training set to fit the segmented and classified point cloud objects. Their method relies on high-quality scans, to make the problem more tractable.

Shao et al. [41] present an interactive approach to semantic modeling of indoor scenes from sparse sets of low-quality Kinect scans. To avoid problems brought by poor-quality depth images, they rely on user interaction to reliably segment RGB-D images into regions with semantic labels manually assigned. Then an automatic algorithm is used to find the best matched model for each object and arrange them to reconstruct the target scene.

Fig. 6 Model-based semantic modeling result. Reproduced with permission from Ref. [7], © 2014 Association for Computing Machinery, Inc.

For complex scenes with many object instances, Shao et al.'s method [41] requires extensive user assistance for segmentation and labeling to resolve ambiguity due to noise and occlusion. Interior objects normally have strong contextual relationships (e.g., monitors are found on desks, and chairs are arranged around tables). Such relationships provide strong cues to determine semantic categories of each object, and has been used in a number of recognition and retrieval tasks, delivering significant improvements in precision. By utilizing such information, Chen et al. [7] propose an automatic solution to this problem. They exploit co-occurrence contextual information in a 3D scene database, and use this information to constrain modeling, ensuring semantic compatibility between matched models.

The performance of model-based methods relies heavily on the quality, diversity and the number of existing 3D models as well as scenes that represent plausible combinations of models. Novel scenes or scene items without representation in the existing 3D model database are likely to lead to poor results. This is currently the main bottleneck of model-based methods.

4 Conclusions

In this paper, we have presented an extensive survey of indoor scene modeling from RGB-D data. We first briefly introduced some public datasets and programming libraries in this area. We divided methods into two categories: geometric modeling and semantic modeling, and overviewed various indoor scene modeling techniques along with their advantages and limitations in each category. However, from the reviewed methods we can see that robust modeling of real-world complex, cluttered or large-scale indoor scenes remains an open problem because of numerous challenges. Generally, researchers in this area have reached a consensus that utilizing prior knowledge is the right direction to improve

modeling algorithms, especially when the data is incomplete and noisy. In fact, with simple prior knowledge, even traditional geometric modeling methods can benefit significantly. Zhou et al. [42] use an observation that scene parts which have been scanned particularly thoroughly tend to be points of interest (POI). By detecting POI from the scanning trajectory and protecting local geometry in POI, they can significantly improve reconstruction results of complex scenes. Salas-Moreno et al. [43] extend the classic SLAM framework to object level using the prior knowledge that many scenes consist of repeated, domain-specific objects and structures. Therefore, obtaining more human prior knowledge and better using it have become a focus of current indoor scene modeling research. By summarizing a broad area of literature, we hope this survey will give valuable insights into this important topic and will encourage new research in this area.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Project No. 61120106007), Research Grant of Beijing Higher Institution Engineering Research Center, and Tsinghua University Initiative Scientific Research Program.

References

[1] Merrell, P.; Schkufza, E.; Li, Z.; Agrawala, M.; Koltun, V. Interactive furniture layout using interior design guidelines. *ACM Transactions on Graphics* Vol. 30, No. 4, Article No. 87, 2011.

[2] Yu, L.-F.; Yeung, S.-K.; Tang, C.-K.; Terzopoulos, D.; Chan, T. F.; Osher, S. J. Make it home: Automatic optimization of furniture arrangement. *ACM Transactions on Graphics* Vol. 30, No. 4, Article No. 86, 2011.

[3] Xiao, J.; Furukawa, Y. Reconstructing the world's museums. *International Journal of Computer Vision* Vol. 110, No. 3, 243–258, 2014.

[4] Izadi, S.; Kim, D.; Hilliges, O.; Molyneaux, D.; Newcombe, R.; Kohli, P.; Shotton, J.; Hodges, S.; Freeman, D.; Davison, A.; Fitzgibbon, A. 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, 559–568, 2011.

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

[6] Savva, M.; Chang, A. X.; Hanrahan, P.; Fisher, M.; Nießner, M. SceneGrok: Inferring action maps in 3D environments. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 212, 2014.

[7] Chen, K.; Lai, Y.-K.; Wu, Y.-X.; Martin, R.; Hu, S.-M. Automatic semantic modeling of indoor scenes from low-quality RGB-D data using contextual information. *ACM Transactions on Graphics* Vol. 33, No. 6, Article No. 208, 2014.

[8] Iddan, G. J.; Yahav, G. Three-dimensional imaging in the studio and elsewhere. In: Proceedings of the International Society for Optics and Photonics, Vol. 4289, No. 48, 48–55, 2001.

[9] Anand, A.; Koppula, H. S.; Joachims, T.; Saxena, A. Contextually guided semantic labeling and search for three-dimensional point clouds. *International Journal of Robotics Research* Vol. 32, No. 1, 19–34, 2013.

[10] Koppula, H. S.; Anand, A.; Joachims, T.; Saxena, A. Semantic labeling of 3D point clouds for indoor scenes. In: Proceedings of the Conference on Neural Information Processing Systems, 244–252, 2011.

[11] Lai, K.; Bo, L.; Fox, D. Unsupervised feature learning for 3D scene labeling. In: Proceedings of 2014 IEEE International Conference on Robotics and Automation, 3050–3057, 2014.

[12] Silberman, N.; Fergus, R. Indoor scene segmentation using a structured light sensor. In: Proceedings of 2011 IEEE International Conference on Computer Vision Workshops, 601–608, 2011.

[13] Silberman, N.; Hoiem, D.; Kohli, P.; Fergus, R. Indoor segmentation and support inference from RGBD images. In: Proceedings of the 12th European Conference on Computer Vision-Volume Part V, 746–760, 2012.

[14] Xiao, J.; Owens, A.; Torralba, A. SUN3D: A database of big spaces reconstructed using SfM and object labels. In: Proceedings of 2013 IEEE International Conference on Computer Vision, 1625–1632, 2013.

[15] Mattausch, O.; Panozzo, D.; Mura, C.; Sorkine-Hornung, O.; Pajarola, R. Object detection and classification from large-scale cluttered indoor scans.

Computer Graphics Forum Vol. 33, No. 2, 11–21, 2014.
[16] Rusu, R. B.; Cousins, S. 3D is here: Point cloud library (PCL). In: Proceedings of 2011 IEEE International Conference on Robotics and Automation, 1–4, 2011.
[17] Information on http://www.mrpt.org.
[18] Besl, P. J.; McKay, N. D. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 14, No. 2, 239–256, 1992.
[19] Chen, Y.; Medioni, G. Object modeling by registration of multiple range images. *Image and Vision Computing* Vol. 10, No. 3, 145–155, 1992.
[20] Durrant-Whyte, H.; Bailey, T. Simultaneous localization and mapping: Part I. *IEEE Robotics & Automation Magazine* Vol. 13, No. 2, 99–110, 2006.
[21] Curless, B.; Levoy, M. A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 303–312, 1996.
[22] Heredia, F.; Favier, R. Kinect Fusion extensions to large scale environments. Available at http://www.pointclouds.org/blog/srcs/fheredia.
[23] Endres, F.; Hess, J.; Engelhard, N.; Sturm, J.; Burgard, W. An evaluation of the RGB-D SLAM system. In: Proceedings of 2012 IEEE International Conference on Robotics and Automation, 1691–1696, 2012.
[24] Information on http://openslam.org.
[25] Lowe, D. G. Object recognition from local scale-invariant features. In: Proceedings of the 7th IEEE International Conference on Computer Vision, Vol. 2, 1150–1157, 1999.
[26] Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* Vol. 110, No. 3, 346–359, 2008.
[27] Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In: Proceedings of 2011 IEEE International Conference on Computer Vision, 2564–2571, 2011.
[28] Fischler, M. A.; Bolles, R. C. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* Vol. 24, No. 6, 381–395, 1981.
[29] Tsai, C.-Y.; Wang, C.-W.; Wang, W.-Y. Design and implementation of a RANSAC RGB-D mapping algorithm for multi-view point cloud registration. In: Proceedings of 2013 International Automatic Control Conference, 367–370, 2013.
[30] Henry, P.; Krainin, M.; Herbst, E.; Ren, X.; Fox, D. RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments. *International Journal of Robotics Research* Vol. 31, No. 5, 647–663, 2012.
[31] Li, M.; Lin, R.; Wang H.; Xu, H. An efficient SLAM system only using RGBD sensors. In: Proceedings of 2013 IEEE International Conference on Robotics and Biomimetics, 1653–1658, 2013.
[32] Lin, R.; Wang, Y.; Yang, S. RGBD SLAM for indoor environment. In: Proceedings of the 1st International Conference on Cognitive Systems and Information Processing, 161–175, 2014.
[33] Duda, R. O.; Hart, P. E. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM* Vol. 15, No. 1, 11–15, 1972.
[34] Stockman, G.; Shapiro, L. *Computer Vision*. Upper Saddle River, NJ, USA: Prentice Hall, 2001.
[35] Oesau, S.; Lafarge, F.; Alliez, P. Indoor scene reconstruction using feature sensitive primitive extraction and graph-cut. *ISPRS Journal of Photogrammetry and Remote Sensing* Vol. 90, 68–82, 2014.
[36] Sanchez, V.; Zakhor, A. Planar 3D modeling of building interiors from point cloud data. In: Proceedings of 2012 19th IEEE International Conference on Image Processing, 1777–1780, 2012.
[37] Li, Y.; Wu, X.; Chrysathou, Y.; Sharf, A.; Cohen-Or, D.; Mitra, N. J. GlobFit: Consistently fitting primitives by discovering global relations. *ACM Transactions on Graphics* Vol. 30, No. 4, Article No. 52, 2011.
[38] Arikan, M.; Schwärzler, M.; Flöry, S.; Wimmer, M.; Maierhofer, S. O-snap: Optimization-based snapping for modeling architecture. *ACM Transactions on Graphics* Vol. 32, No. 1, Article No. 6, 2013.
[39] Kim, Y. M.; Mitra, N. J.; Yan, D.-M.; Guibas, L. Acquiring 3D indoor environments with variability and repetition. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 138, 2012.
[40] Nan, L.; Xie, K.; Sharf, A. A *search-classify* approach for cluttered indoor scene understanding. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 137, 2012.
[41] Shao, T.; Xu, W.; Zhou, K.; Wang, J.; Li, D.; Guo, B. An interactive approach to semantic modeling of indoor scenes with an RGBD camera. *ACM Transactions on Graphics* Vol. 31, No. 6, Article No. 136, 2012.
[42] Zhou, Q.-Y.; Koltun, V. Dense scene reconstruction with points of interest. *ACM Transactions on Graphics* Vol. 32, No. 4, Article No. 112, 2013.
[43] Salas-Moreno, R. F.; Newcombe, R. A.; Strasdat, H.; Kelly, P. H. J.; Davison, A. J. SLAM++: Simultaneous localisation and mapping at the level of objects. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition, 1352–1359, 2013.

Kang Chen received his B.S. degree in computer science from Nanjing University in 2012. He is currently a Ph.D. candidate in the Institute for Interdisciplinary Information Sciences, Tsinghua University. His research interests include computer graphics and geometric modeling and processing.

Yu-Kun Lai received his bachelor degree and Ph.D. degree in computer science from Tsinghua University in 2003 and 2008, respectively. He is currently a lecturer in visual computing in the School of Computer Science & Informatics, Cardiff University. His research interests include computer graphics, geometry processing, image processing, and computer vision.

Shi-Min Hu is currently a professor in the Department of Computer Science and Technology, Tsinghua University. He received his Ph.D. degree from Zhejiang University in 1996. His research interests include digital geometry processing, video processing, rendering, computer animation, and computer aided geometric design. He has published more than 100 papers in journals and refereed conferences. He is the Editor-in-Chief of *Computational Visual Media*, and on the editorial boards of several journals, including *IEEE Transactions on Visualization and Computer Graphics*, *Computer Aided Design*, and *Computer & Graphics*.