Privacy Preserving Dynamic Room Layout Mapping

Xinyu Li[1(✉)], Yanyi Zhang[1], Ivan Marsic[1], and Randall S. Burd[2]

[1] Department of Electrical and Computer Engineering, Rutgers University,
New Brunswick, NJ, USA
{xinyu.li1118,yz593,marsic}@rutgers.edu

[2] Division of Trauma and Burns, Children's National Medical Center,
Washington, D.C., USA
RBurd@childrensnational.org

Abstract. We present a novel and efficient room layout mapping strategy that does not reveal people's identity. The system uses only a Kinect depth sensor instead of RGB cameras or a high-resolution depth sensor. The users' facial details will neither be captured nor recognized by the system. The system recognizes and localizes 3D objects in an indoor environment, that includes the furniture and equipment, and generates a 2D map of room layout. Our system accomplishes layout mapping in three steps. First, it converts a depth image from the Kinect into a top-view image. Second, our system processes the top-view image by restoring the missing information from occlusion caused by moving people and random noise from Kinect depth sensor. Third, it recognizes and localizes different objects based on their shape and height for a given top-view image. We evaluated this system in two challenging real-world application scenarios: a laboratory room with four people present and a trauma room with up to 10 people during actual trauma resuscitations. The system achieved 80 % object recognition accuracy with 9.25 cm average layout mapping error for the laboratory furniture scenario and 82 % object recognition accuracy for the trauma resuscitation scenario during six actual trauma cases.

Keywords: Depth sensor · Occlusion compensation · Object recognition · Privacy preserving · Room layout mapping

1 Introduction

Dynamic room layout mapping is useful and critical to many applications such as activity recognition, virtual reality and home automation. The goal of room layout mapping is to identify and localize objects (furniture, equipment, etc.) in an indoor environment and generate a 2D map with their locations. It is also critical that the room layout mapping system be able to work with people moving in the room. Traditional room layout modeling strategies are usually based on RGB cameras or RGBD cameras [1, 2]. Although many systems have been proposed, these systems are not widely used in real-world applications for several reasons. First, RGB camera raises privacy concerns, especially in privacy-sensitive domains such as medical settings. Second, some previous approaches have made use of multiple cameras or a moving robot to better observe the

A. Mansouri et al. (Eds.): ICISP 2016, LNCS 9680, pp. 61–70, 2016.
DOI: 10.1007/978-3-319-33618-3_7

environment, an approach that may not be cost-efficient and may potentially interfere with the work. In addition, most existing research uses 3D models for template matching for indoor 3D object recognition. The 3D model matching yields good performance in some daily-living scenarios with large furniture, such as beds or dining tables. It does not, however, perform well for furniture or equipment with irregular shapes, such as chairs or medical equipment like a fluid-bag stand. In addition, most template matching methods require an unobstructed view of the environment, which is often not the case because people cause view occlusion. The multi-camera based solution [3] is not cost-efficient and aligning the views from different cameras is difficult and slow. We present a system that uses only Kinect V2 depth sensor. Because we do not use RGB cameras and the depth sensor built in the Kinect V2 cannot capture facial details, user's identity cannot be revealed.

To recognize and track 3D objects, such as furniture pieces or equipment in the room, our system first converts the 3D depth-point-cloud taken by the Kinect depth sensor into a 2D top-view image. The furniture mapping and recognition are based on the pixels in the top-view image. The challenge is that the top-view image is not always clear and representative: people walking in the room may cause view occlusion, which leads to undefined pixels in the top-view image that will cause problems with recognition. In addition, the Kinect depth sensing system also generates random noise, which affects the performance of mapping and recognition system. To restore the occluded view, our system tracks people positions in the room and dynamically updates the top-view image based on their location. We also apply filtering to the top-view images to minimize the influence of random noise caused by the Kinect depth sensor.

We evaluated the system performance in two challenging real-world applications: a laboratory room with four people work in it and an actual trauma room of a level 1 trauma center. We tracked six types of furniture pieces in a laboratory room and achieved average 80 % object recognition accuracy with average layout mapping error of 9.25 cm over a 48 h testing period with four people working in the room. For the trauma room application, we were able to keep tracking and recognize eight medical instrument and furniture sets in the room with an average object recognition accuracy rate of 82 %. The contributions of this paper are:

1. A novel dynamic, privacy-preserving room layout mapping strategy using only a commercial depth sensor.
2. A strategy to restore the missing information in Kinect depth-maps caused by random noise or view occlusion by moving people.
3. The implementation and evaluation of the system in two challenging real-world applications.

2 Related Works

Room layout mapping or indoor 3D objects recognition is a widely used method for applications such as activity recognition, 3D object modeling and virtual reality [2, 4]. Early layout mapping strategies used 2D-image features such as SIFT to map indoor furniture [2]. Solutions based on 2D RGB camera, however, are not feasible for complex and dynamic environments because camera angles, lighting, and distance between the object and camera may change rapidly and may be difficult to control in real-world implementations.

With the recent development of commercial depth sensors, researchers have started using RGB-D camera or depth sensor for room layout modeling and furniture recognition. A common approach is to compare the reconstructed 3D environment with the pre-established 3D CAD models [5–7]. Although approaches based on 3D-model matching achieve good performance, their application is limited because some systems require views from different angles or additional cameras [3] which is not cost effective and may hard to implement in real-world applications. In addition, for systems that require known 3D models of objects as templates [8, 9], it may be impractical to build a 3D CAD model for every object used in real-world applications. Another common approach for recognizing and mapping 3D objects is projecting 3D data onto 2D space [10, 11]. Our method builds on this approach by converting a 3D point-cloud into 2D space, and then performing object mapping and recognition based on 2D top-view image. Rather than using a roaming robot equipped with cameras, which is expensive and potentially unfeasible in a crowded setting, we used a fixed commercial depth sensor.

Most prior work in 3D object recognition does not include people. This omission is problematic for many applications because the information loss due to view occlusion caused by people moving will significantly influence the system's performance. To address this issue, we restore and enhance the images based on people locations in the room. People tracking has attracted a great deal of research and has become more manageable with recent hardware, such as Kinect [12, 13]. Our system achieves comparable layout mapping performance in both stationary environments with no people, and in the dynamic real-world scenarios with multiple moving people.

3 Room Layout Mapping

3.1 Room Layout Mapping Model

A key challenge for room layout mapping is that view occlusion caused by people moving in the room results in the captured image partially missing information. Created by the need to avoid compromising user's privacy, an additional challenge, is to use only depth sensor built in the Kinect, which provides low-resolution depth image without detailed texture information. Both view occlusion and low-resolution depth image makes indoor objects recognition difficult. We designed the system to accomplish room layout mapping in dynamic environment through three steps:

Step1: Generate a point-cloud map of the 3D environment based on the depth sensor and converts it into top-view image.

Step2: Process the top-view image by first restoring the part of view occluded by people in the room and then enhancing the top-view image to eliminate the undefined pixels caused by Kinect sensor's random noise.

Step3: Recognize 3D objects (equipment, furniture, etc.) based on shape and height using 2D template matching and then layout mapping based on recognition results.

3.2 From Point-Cloud to Top View

To have a clear view of the room, we mounted the Kinect H meters above the ground with a tilt angle α so that people and objects in the room are more likely to be seen in camera view (in our application, $H = 2.5$ m and $\alpha = 7$’). Before converting the 3D point-cloud into a 2D top view, each point in the point-cloud needs to be adjusted for the tilt angle α so that the camera space of the Kinect is aligned with the actual setting. The camera space refers to the 3D coordinate system used by the Kinect (Fig. 1), where the x axis grows to the sensor's left, the y-axis grows up and the z axis grows out in the direction the sensor is facing.

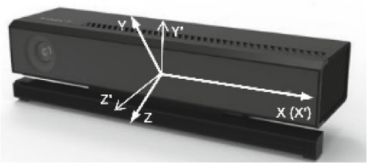

Fig. 1. Coordinate system of Kinect camera space.

If we use *(X, Y, Z)* for each point in the point-cloud in camera space and use *(X', Y', Z')* for the corresponding point in the room, given the tilt angle α, the rotation matrix can be applied as follows:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & -\sin(\alpha) & \cos(\alpha) \end{bmatrix} \times \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} 0 \\ H \\ 0 \end{bmatrix} \tag{1}$$

For objects below the Kinect sensor, the measured y coordinate in camera space is negative. We add H (height of Kinect sensor) to the converted y coordinates to calculate the actual height. By visualizing the point-cloud (Fig. 2) we confirmed that before the tilt-angle adjustment, the ceiling surface (purple pixels) and floor surface (blue pixels) in the point-cloud are not parallel with actual ground due to the tilt angle of Kinect (Fig. 2(b)) After the tilt angle is adjusted, the surfaces are parallel with the actual ground (Fig. 2(c)). Once the tilt angle is adjusted, the top-view image can be generated by

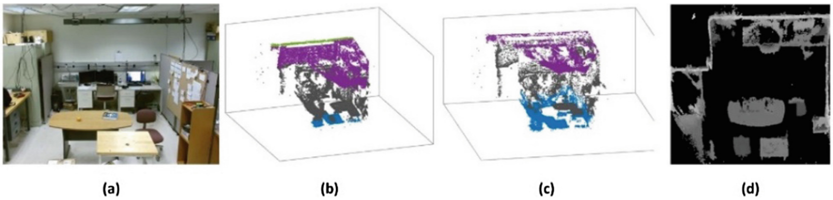

Fig. 2. (a) The picture of the laboratory. (b) The point cloud of the room before tilt angle adjustment. (c) The point cloud of the room after tilt angle adjustment. (d) The top view image of the room. (Color figure online)

projecting all the points in the point-cloud down to the floor plane using the height of the highest pixel at each position in 2D plane as follows:

$$topView(x, z) = max_{Depth}\left(Y' | X' = x, Z' = z\right) \tag{2}$$

Where $topView(x, z)$ denotes the value of pixels in the top-view image and $max_{Depth}\left(Y' | X' = x, Z' = z\right)$ represents the highest point in point-cloud at position (x, z). Because objects are present that we do not want to track, such as lights hanging from the ceiling, we ignore all the points in point-cloud above certain height range to avoid possible confusion. Because the top-view image generated from the adjusted depth image captured by the Kinect, where each pixel represents 1 cm, the pixels in the top-view image reveal the physical dimensions where each pixel represents a 1 cm length in the room. The correlation between pixels and actual distance makes precise room layout mapping possible.

3.3 Top-View Image Enhancement

View occlusion leads to information loss and makes it difficult to recognize 3D objects in top-view images. To address this issue, we restore the missing information in top-view images.

First, the system maintains a dynamic top-view image by selectively updating its pixels. We require that the system start when no people are in the room, which guarantees the initial top-view image free of view occlusion caused by people. The Kinect is set to capture a depth image of the room every 100 ms, which is converted into a top-view image. When a new depth image is captured, the top-view image will not be directly updated unless the room remains empty of people. If a person is detected by the Kinect, the view will be occluded, resulting in information loss in the area of occlusion. Our system only updates the pixels outside the occluded area based on the top-view image and keeps the pixels in the occluded area unchanged from the previous top-view image (Fig. 3).

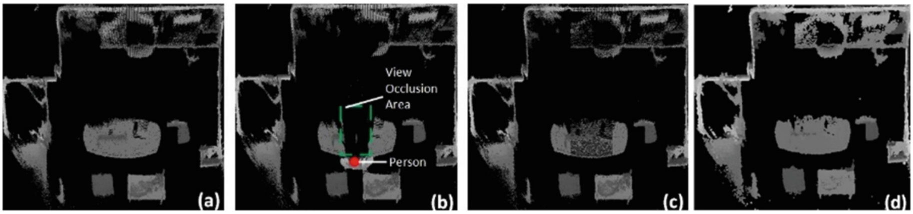

Fig. 3. (a) Top view image with no people. (b) View occlusion caused by people in the room. (c) Compensating for the information loss. (d) Top view image after image enhancement.

Because people blocking the camera view cause view occlusion, occluded areas are defined as rectangular areas behind each person in the room. We used a fixed width for each rectangle (50 cm) based on the average body width of a male adult. The height *d*

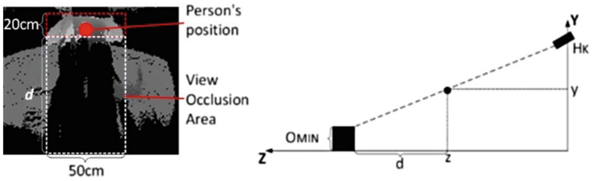

Fig. 4. Left: A person's figure in top view image (red rectangle) and view occlusion area caused by the person (white rectangle). Right: Estimate the length of view occlusion area *d* based on person's position *z* and height *y*. (Color figure online)

of the rectangle is determined from person's location in the room and their estimated height (Fig. 4):

$$d = \frac{(y - O_{min}) \cdot z}{H_k - O_{min}} \tag{3}$$

Here y denotes the height of a person's head-joint provided by the Kinect and z denotes the horizontal distance between a person's head and the Kinect sensor. The height H_k of the Kinect sensor is fixed once it is installed and O_{min} represents the minimum height of tracked objects. Because the top-view of people will act as outliers to the object recognition system, to have a clear top view image, we replaced the rectangle area (in this paper 20 cm × 50 cm) containing the top view of a person with the pixels in previous top view, so that the figure of person will be "erased" in top-view image (Fig. 3(c)).

In addition to restoring the occluded view, we used filtering to eliminate the random noise caused by the Kinect sensor. We used a buffer to store n previous top-view images and calculate the value of each pixel in current top-view image based on both the current top-view image and n previous top-view images. If a pixel $topView(x, y)$ is undefined in current top view image, the system will look back n (n = 5 in our experiments) to previous top view images and assign the value of that pixels using the average value of same pixel in n previous top View images. The image dilate [14] is applied at last to smooth the top-view image. We used the Aforge.net library in our programming [15].

3.4 Object Recognition and Room Layout Mapping

We focused on small furniture and equipment that can be relocated, such as chair, patient-bed or a cart, while large furniture, e.g. metal cabinet, is not likely to be often relocated. Our object recognition is based on template matching on 2D top-view images. We chose template-matching strategy among other pattern recognition approaches for two reasons. First, template matching is fast compared with other strategies, a requirement for real-time room layout mapping because small objects may be frequently relocated. Second, because we are working with top-view images, which are derived from the Kinect depth images, detailed texture information is not available to extract image features.

We generated templates for objects by manually selecting examples of each object from 10 top-view images and averaging them for the template. In practice, the template matching was done using *Aforge.net* library [15], the system will take the area that has maximum matching score with certain template as the location of object. Because, equipment or small furniture pieces might be moved out of the room in real-world applications, we defined a threshold to determine if an object exists in the room. If the maximum matching score of a template is lower than the threshold, the system will decide the certain object is not in the room. The threshold was tuned based on 100 random selected sample images using object recognition accuracy as indicator.

4 Experimental Results

We applied our system in two different application scenarios to test the system performance for normal application and extremely complex and dynamic environment. We first set up the system in a typical laboratory room with four people work in and kept it running for 48 h. None of the individuals were told the system was working in the room in advance so their behavior was not altered by this experiment. We kept tracking six different furniture pieces (Table 1) in the room with very different size and shape, four of the objects were relocated during the experiment. We programmed the system to save a top-view image and a room layout image every minute and later evaluate the system's performance by manually reviewing the top-view image as ground truth and compare it with generated room layout map for evaluation.

The laboratory application scenarios were not crowded and people were working at different place of the room most of time. To further test our system performance, we also applied the system in an actual trauma room in a level 1 trauma center. Deployment and evaluation of the system was approved by the institutional review board (IRB) and was considered do not compromise users' privacy of medical personnel and patients. In this setting, up to 10 people are working simultaneously, an aspect of this setting that leads to view occlusion and increase the difficulty of room layout mapping. We kept tracking eight different piece of medical equipment and furniture sets (Table 1) that varies in shape and size during six trauma resuscitations. Five pieces of medical equipment were relocated. Similar to the experiments we performed in the laboratory, we programmed the system to save a top-view image and a room layout image every minute (Fig. 5). The object recognition accuracy was calculated as follows (Table 1):

$$Accuracy = \frac{number\ of\ images\ that\ an\ object\ is\ correctly\ recognized}{total\ number\ of\ images}$$

We find that the large furniture pieces are easier to track and map in the room: even if few pixels are significantly influenced by the noise or blocked, most of pixels stay similar with template. For smaller objects mapping accuracy is lower because small objects result in few pixels that are easier corrupted by noise. Besides, for objects with irregular shapes, it is hard to generate a representative template for objects with irregular shape. For example, the blood pressure stand has a relatively low tracking accuracy due to its small size and irregular shape (Table 1). Our system is able to achieve performance

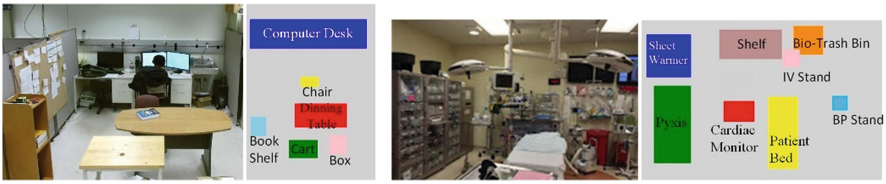

Fig. 5. The photo and room layout mapping results of laboratory and trauma room.

Table 1. Objects used for experiments in laboratory room and trauma room with their dimensions and recognition accuracy. The Grey shaded objects were relocated during the experiments. For the objects with irregular shape, height (H) is measured.

Objects(lab)	Size (inch)	Accuracy (0/4 people)	Objects (Trauma room)	Size (inch)	Accuracy (up to 10 people)
Dinning Table	58×28×30	100%/100%	Patient Bed	31×32×80	92.3%
Cart	28×18×36	100%/87.5%	Shelf	57×42×20	100%
Chair	H=31", irregular shape	100%/42.8%	IV Stand	H=68", irregular shape	67.8%
Box	15×13×19	100%/67.1%	BP Stand	H=43", irregular shape	46.1%
Book Shelf	12×14×16	98.0%/89.1%	Cardiac Monitor Adapter	30×28×17	62.5%
Desk	91×30×29	99.1%/90.2%	Bio-Trash Bin	19×32×16	93.2%
			Pyxis	82×27×78	100%
			Sheet Warmer	30×29×72	99.8%

comparable to previous research [1, 8], however, considering that previous approaches [1, 8] were designed for scenarios with no people moves, our system is considered to more suitable for many real-world scenarios.

When no people are moving, the overall system performance in the laboratory room is around 15 % better than it is when four people are moving in the room. This finding may be because Kinect might not be able to continuously track of people in a crowded and dynamic environment. Loss of tracking of people did not necessarily lead to object detection failure, because it is also depending on the whether the people is standing in front of an objects and the size of view occlusion area caused by people. In the medical application, our system was able to maintain similar performance to laboratory room implementation when number of people doubled. We noticed that the patient bed is blocked by personnel most of the time during the trauma resuscitation, because doctors and nurses usually stand around the patient bed when performing their tasks. Our system is able to make compensation and recognize patient bed with 92.3 % accuracy of patient bed with proposed strategy that is satisfied.

In addition to object recognition accuracy, we evaluated the object mapping error in our laboratory environment without people. Because the time of patient arrival is not predictable, staying in actual trauma room and manually measuring the room layout mapping error may potentially interfere with the room's normal usage. For this reason, we did not perform layout mapping error evaluation. We applied the rotation matrix to the Kinect camera space and used the coordinate system after tilt angle adjustment for mapping error evaluation. The Kinect sensor is used as origin of the coordinate system, the x-axis grows to the Kinect's left, z-axis grows out of the Kinect and parallel with ground. We manually measured the distance of each object in the room to Kinect in both x-axis and z-axis direction and compared the measured distance with generated room layout map. The objects were relocated after each experiment and the entire process was repeated 10 times. We averaged the errors in 10 measurements for each object in the room as room layout mapping error (Table 2).

Table 2. Room layout mapping error in X-axis and Z-axis in cm.

Objects	Dinning table	Cart	Chair	Box	Book Shelf	Desk
Error in X-axis (cm)	13	9.7	20.0	13.0	11.6	11.6
Error in Z-axis (cm)	2.7	13.2	5.5	1.6	2.7	6.4

Our research achieves decimeter-level layout mapping error which is similar to previous research [8]. A difference is that our approach uses only a fixed depth sensor which does not compromise user's privacy. In addition, our approach does not rely on pre-defined 3D CAD models for object recognition, a more practical and convenient for real-world applications. Previous research [10] achieves around 3 cm layout mapping error using laser depth sensor, which is considered to be one of the best indoor layout mapping system. The limitation of this type of system is that the system requires a clear view of the room, view occlusion caused by moving people will significantly influence system performance. The indoor layout mapping strategy proposed in this paper works well with people moving in the room, though the system works with slightly higher layout mapping error, it is considered to be more useful and practical for some dynamic applications.

5 Conclusion and Future Work

We developed a novel room layout mapping system that works in crowded real-world applications that does not compromise the privacy of the people in the view area. Its key feature is eliminating the view occlusion caused by people moving in the room. Due to Kinect limitation, Kinect can track no more than 6 people. If more than six people are in the room, the information lost due to view occlusion might not be restored. In addition, it is possible for Kinect to lose track of people in complex and crowded environments. The system may lose track of some objects or make layout-mapping errors in very crowded environments. Also, template matching will not work well if a small object is blocked or partially blocked by larger objects.

Our future work will use multiple Kinect sensors [16] in different view angles to allow tracking more than six people and restore missing information caused by view occlusion from both people and objects. We will also perform more extensive evaluation of our system in different real-world application scenarios.

References

1. Tomono, M.: 3-D object map building using dense object models with SIFT-based recognition features. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE (2006)
2. Rusu, R.B., et al.: Functional object mapping of kitchen environments. In: IROS 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008. IEEE (2008)
3. Susanto, W., Rohrbach, M., Schiele, B.: 3D object detection with multiple kinects. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part II. LNCS, vol. 7584, pp. 93–102. Springer, Heidelberg (2012)
4. Varvadoukas, T., et al.: Indoor furniture and room recognition for a robot using internet-derived models and object context. In: 2012 10th International Conference on Frontiers of Information Technology (FIT). IEEE (2012)
5. Lin, D., Fidler, S., Urtasun, R.: Holistic scene understanding for 3d object detection with rgbd cameras. In: IEEE International Conference on Computer Vision (ICCV), 2013. IEEE (2013)
6. Camplani, M., Mantecon, T., Salgado, L.: Depth-color fusion strategy for 3-d scene modeling with kinect. IEEE Trans. Cybern. **43**(6), 1560–1571 (2013)
7. Du, H., et al.: Interactive 3D modeling of indoor environments with a consumer depth camera. In: Proceedings of the 13th International Conference on Ubiquitous Computing. ACM (2011)
8. Wittrowski, J., Ziegler, L., Swadzba, A.: 3d implicit shape models using ray based hough voting for furniture recognition. In: 2013 International Conference on 3D Vision-3DV 2013. IEEE (2013)
9. Günther, M., et al.: Model-based furniture recognition for building semantic object maps. Artif. Intell. (2015)
10. Valero, E., Adán, A., Bosché, F.: Semantic 3D reconstruction of furnished interiors using laser scanning and RFID technology. J. Comput. Civil Eng. 04015053 (2015)
11. Salas-Moreno, R.F., et al.: Slam ++: simultaneous localisation and mapping at the level of objects. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2013)
12. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MultiMedia **19**(2), 4–10 (2012)
13. Munaro, M., Menegatti, E.: Fast RGB-D people tracking for service robots. Auton. Robots **37**(3), 227–242 (2014)
14. Chen, S., Haralick, R.M.: Recursive erosion, dilation, opening, and closing transforms. IEEE Trans. Image Process. **4**(3), 335–345 (1995)
15. AForge.Net Library. http://www.aforgenet.com/framework
16. Asteriadis S, Chatzitofis A, Zarpalas D, et al.: Estimating human motion from multiple kinect sensors. In: Proceedings of the 6th International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, p. 3. ACM (2013)