LOW COST AND EFFICIENT 3D INDOOR MAPPING USING MULTIPLE CONSUMER RGB-D CAMERAS

C. Chen [a, *], B.S. Yang [a], S. Song [a, *]

[a] State Key Laboratory of Information Engineering in Survey, Mapping and Remote Sensing, Wuhan University, No. 129, Luoyu Road, Wuhan, PR China - chenchi_liesmars@foxmail.com, (bshyang, shinesong)@whu.edu.cn

Commission I, WG I/3

KEY WORDS: Indoor Mapping, RGB-D Camera, Kinect, Calibration

ABSTRACT:

Driven by the miniaturization, lightweight of positioning and remote sensing sensors as well as the urgent needs for fusing indoor and outdoor maps for next generation navigation, 3D indoor mapping from mobile scanning is a hot research and application topic. The point clouds with auxiliary data such as colour, infrared images derived from 3D indoor mobile mapping suite can be used in a variety of novel applications, including indoor scene visualization, automated floorplan generation, gaming, reverse engineering, navigation, simulation and etc. State-of-the-art 3D indoor mapping systems equipped with multiple laser scanners product accurate point clouds of building interiors containing billions of points. However, these laser scanner based systems are mostly expensive and not portable. Low cost consumer RGB-D Cameras provides an alternative way to solve the core challenge of indoor mapping that is capturing detailed underlying geometry of the building interiors. Nevertheless, RGB-D Cameras have a very limited field of view resulting in low efficiency in the data collecting stage and incomplete dataset that missing major building structures (e.g. ceilings, walls). Endeavour to collect a complete scene without data blanks using single RGB-D Camera is not technic sound because of the large amount of human labour and position parameters need to be solved. To find an efficient and low cost way to solve the 3D indoor mapping, in this paper, we present an indoor mapping suite prototype that is built upon a novel calibration method which calibrates internal parameters and external parameters of multiple RGB-D Cameras. Three Kinect sensors are mounted on a rig with different view direction to form a large field of view. The calibration procedure is three folds: 1, the internal parameters of the colour and infrared camera inside each Kinect are calibrated using a chess board pattern, respectively; 2, the external parameters between the colour and infrared camera inside each Kinect are calibrated using a chess board pattern; 3, the external parameters between every Kinect are firstly calculated using a pre-set calibration field and further refined by an iterative closet point algorithm. Experiments are carried out to validate the proposed method upon RGB-D datasets collected by the indoor mapping suite prototype. The effectiveness and accuracy of the proposed method is evaluated by comparing the point clouds derived from the prototype with ground truth data collected by commercial terrestrial laser scanner at ultra-high density. The overall analysis of the results shows that the proposed method achieves seamless integration of multiple point clouds form different RGB-D cameras collected at 30 frame per second.

1. INTRODUCTION

Driven by the miniaturization, lightweight of positioning and remote sensing sensors as well as the urgent needs for fusing indoor and outdoor maps for next generation navigation, 3D indoor mapping from mobile scanning is a hot research and application topic. The point clouds with auxiliary data such as colour, infrared images derived from 3D indoor mobile mapping suite can be used in a variety of novel applications, including indoor scene visualization (Camplani et al., 2013), automated floorplan generation, gaming, reverse engineering, navigation, simulation (Gemignani et al., 2016) and etc. State-of-the-art 3D indoor mapping systems equipped with multiple laser scanners (Trimble, 2016) product accurate point clouds of building interiors containing billions of points. However, these laser scanner based systems are mostly expensive and not portable. Low cost consumer RGB-D Cameras provides an alternative way to solve the core challenge of indoor mapping that is capturing detailed underlying geometry of the building interiors.

However, low cost RGB-D Cameras are often not equipped with position and orientation measurement suit, and the visual odometry (Gutierrez-Gomez et al., 2016; Huang A, 2011; Nistér et al., 2006; Whelan et al., 2015) is often used as substitution of active measurement equipment such as IMU. Similar as the IMU, position drift is inevitable when the visual odometry is used. Henry (Henry et al., 2014) developed the RGB-D vision SLAM system to solve the drift problem of visual odometry, which employed the ICP and RE-RANSAC method to process vision points, and optimized the pose graph built by sparse feature categorize in each frame. There are three main steps of classical indoor mapping method based on RGB-data: First, the spatial position transformation was resolved using 2D Image feature detection and tracking techniques to match feature between frames. Then the loop closure detection was applied as constraints for global optimization. Finally, the match errors were minimized using global consistency constraints.

The vision SLAM system has been applied to solve the indoor data problem from RGB-D for a certain degree, however, RGB-D Cameras have very limited field of view resulting in low efficiency in the data collecting stage. Furthermore, lead to incomplete dataset that missing major building structures (e.g. ceilings, walls) (Yang et al., 2015). Meanwhile, the visual odometry does not work properly in no texture region or regions with repetitive textures, which are quite common for indoor images. In general, the FOV of depth camera is smaller than 60 degrees, and the available distance is between 3 to 5 m, which are extremely easy to cause the track failure or match error.

* Corresponding author

Aim to solve the above problem of depth camera and provide an efficient and economical solution for indoor data collection, this paper proposed a novel method using sensor array by combination of multi Kinect sensors, and made a prototype of indoor scanner.

2. METHOD

2.1 Hardware

There are three types of depth camera including stereo camera, structured light camera and TOF (time-of-flight) camera, distinguish by the measurement principal. The stereo camera and structured light camera using parallax theory to calculate depth, while the TOF camera is based on beam distance measurement principle (Sarbolandi et al., 2015), calculate distance from travel time of modulated beam between sensor and object. The precision, resolution and error distribution are better than other two types. Also, as using IR light source, they can mitigate the effect of ambient light. Among the Microsoft Kinect v2, CubeEye and PMD CamCube of the main types of TOF cameras, the Kinect v2 is selected as sensor array components for its wider FOV, higher resolution and longer ranging (Corti et al., 2016), as listed in Table 1.

Items	Details
Depth Image Resolution	512 x 424
Image Resolution	1920 x 1080
Depth Range	0.5-4.5m (Extend to 8m)
Validate FOV(V x H)	~70° x ~60°
Frame Rate	60 Hz

Table 1. Specification of Kinect v2

Compared to outdoor environment, the indoor data collection with depth camera is often impeded by more fend, shorter distance and smaller space, therefore, both horizontal movement and vertical pitch are required to make data complete, which lead to higher risk of tracking lost and more data processing workload. Besides, most RGB-D reconstruct method use the visual odometry to build relatively transformation between frames, which does not work properly in no texture region or regions with repetitive textures which are quite common for indoor images (Yousif et al., 2014). The horizontal vision contains maximum information of three viewpoint but also most unstable and more prone to lost tracking, while downward and upward vision contains less information but with stable textures suitable for system positioning and posing.

According to the characteristics of the indoor environment, we proposed such layout of sensors, as showed in Figure 1. The pitch angles of three Kinect v2 sensors are set at -50, 0 and 50 degrees, and the horizontal and rotational angles are kept consistent. FOVs of each sensor are 10 degree overlapping with adjacent one. The sensor array system could provide 160°x 70° FOV, and cover 100 m^2 for 3 m high building theoretically, which is enough for most indoor buildings.

All sensors are locked in a tripod bar using ball head, and connected to mobile work station by USB3.0 interface. Every sensor is allocated 5Gbps bandwidth with USB3.0 expansion card plugged in separate PCI-E slots, the connection system is shown as Figure 2.

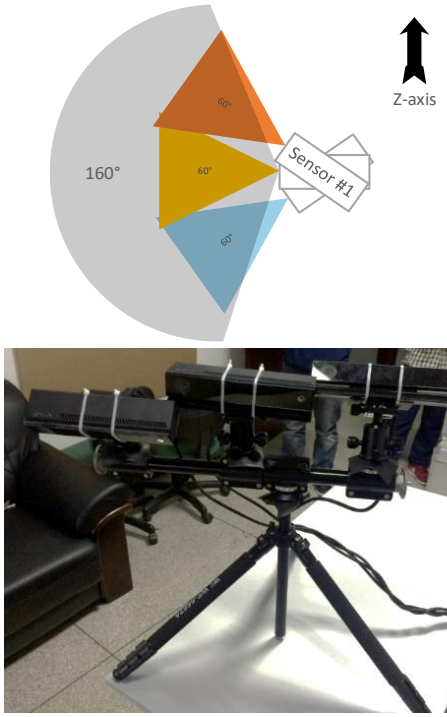


Figure 1. Sensor profile

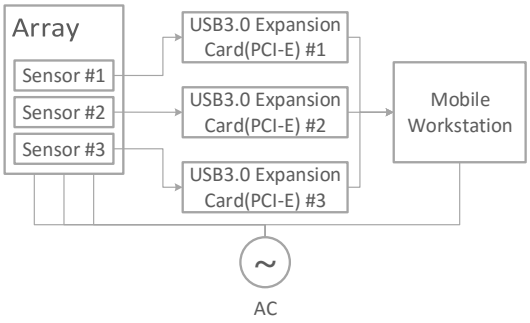


Figure 2. System connection

2.2 Calibration of Sensor Array

Calibration is essential before the sensor array system can be used. The Kinect v2 consists of a RGB camera, an infrared camera and an IR Illuminator (Figure 3). Separate calibration is required to determine the camera intrinsic parameter for each camera, besides, calibration for the relative position and attitude is also needed. Compared to conventional camera model, depth calibration is important for the depth camera. Finally, the relative position and attitude is calibrated for each sensor of the array. Detailed steps are provided below.

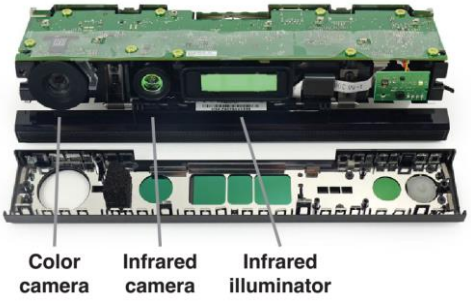


Figure 3. Sensor structure of Kinect v2

2.2.1 Intrinsic Calibration: Precision intrinsic parameters could not only correct distortion of image, and are important to enhance the accuracy of depth and color image fusion (Gui et al., 2014). Intrinsic parameters of pin-hole model for RGB camera including focal length, principal point coordinate and distortion parameters, etc. And, radial distortion and tangential distortion models are employed for lens distortion, as formula 1 and 2.

The image coordinate is transformed with formula 3, in which w=Z, x y denote pixel coordinate and XYZ represents image coordinate. Intrinsic parameters for each camera are obtained with $\text{Intrinsics}_{cam} = \{f\ c_x\ c_y\ k_1\ k_2\ k_3\ p_1\ p_2\}$.

$$\left.\begin{aligned} x_{corrected} &= x(1 + k_1r^2 + k_2r^4 + k_3r^6) \\ y_{corrected} &= y(1 + k_1r^2 + k_2r^4 + k_3r^6) \end{aligned}\right\} \quad (1)$$

$$\left.\begin{aligned} x_{corrected} &= x + [2p_1xy + p_2(r^2 + 2x^2)] \\ y_{corrected} &= y + [p_1(r^2 + 2y^2) + 2p_2xy] \end{aligned}\right\} \quad (2)$$

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \quad (3)$$

2.2.2 Relative Pose of IR and Color Camera: The relative pose of Infrared and Color cameras are calculated to obtain more precise overlay color and depth image for unbiased texture. According to spatial transformation formula 4, the relationship between two cameras could be calculated using 3x3 rotation matrix R and 3x1 translation vector t.

An amount of checkerboard calibration plate images were collected with Infrared and Color cameras, and corner coordinate of plate were extracted respectively, then the $\text{Extrinsic}_{cam} = \{R \quad t\}$ was calculated following Zhang method.

$$\begin{bmatrix} X' \\ Y' \\ Z' \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (4)$$

2.2.3 Depth Correction: TOF camera consists of infrared laser emitter and a series of infrared sensors, and compute distance from travel time of modulated beam between the sensor and object. There are two main TOF technologies are the pulsed and the continuous wave. In the first case, devices compute the distance d using time delay between transmitted pulse and the first echo pulse (Formula 5). This method requires very precise measurement accuracy, which is impossible to be achieved at room temperature.

In the second case, devices takes advantages of continuous modulation ray, such as sine or square wave signal, and calculated the distance from phase displacement between transmitted and echo pulse. Due to periodic signals, this method provides ambiguity distance, and thus restrict the ranging distance of continuous wave TOF cameras.

$$d = c\frac{\Delta t}{2} \quad (5)$$

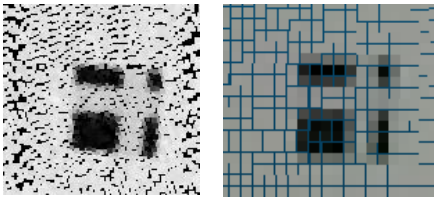
$$d_{amb} = \frac{c}{2f} \quad (6)$$

The continuous wave measurement model is employed to calculated the depth for Kinect v2. As a consumer level RGB-D camera, the calibration of the TOF camera is essential to reduce the system errors. In this paper, the depth drift of Kinect v2 measurement was calibrated by calculating the depth difference with the calibration plate (Fankhauser et al., 2015).

2.2.4 Relative Pose of Sensors: The key factor that the sensor array distinct from multi sensors is that array can be identified as a single one. Toward this purpose, the position of each sensor needs to be calibrated during system integration. The principle of cross-sensor calibration is identical as formula 4, and the spatial transformation is from three dimension to three dimension. The SVD method (Jiyoung et al., 2015) is adopted to calculate the least square rigid transformation matrix in this paper.

Numbers of reference target should be collected for calibration. A single station scanning data from terrestrial laser scanner was obtained as ground truth data, and reference point coordinates were collected to resolve sensor position and orientation. The reference target should be significant and explicit. Limited by the resolution and depth precision, sharp objects are not distinct, therefore, the RGB texture are used to identify reference target. Since the terrestrial laser scanner does not contain RGB information, the reproject errors are inevitable even single lens reflex is mounted to collect texture images.

Taking both the color contrast and reflection intensity into account, the reference target is made of two types of material with intense color contrast and reflection. Images of the reference target in laser point data and TOF cloud data are showed in Figure 4.



(a) (b)

Figure 4. High-contrast Target in (a) Lidar data colorized by intensity (b) Kinect data colorized by texture

The precision of depth measurement are not enough for registration with reference target only (Diaz et al., 2015; Sarbolandi et al., 2015). Moreover, affected by multi-path interference effects of indoor scenes (Jiménez et al., 2014), the center point of the target is bending deformation around the corner when the TOF camera is used. In this case, redundant measurements are required to reduce the errors. The ICP (Iterative Closest Point) is a common method to minimize the distance between two point-cloud data, by iterative correction of transformation parameters between target point cloud and reference point cloud. In this study, the Point-to-Plane ICP is used to match the depth data to TLS point cloud data, and then calculate the $\text{Extrinsic}_{sensor} = \{R \quad t\}$ for each sensor.

3. EXPERIMENT

3.1 Experiment of calibration

The resolution of Kinect v2 is 1920x1080 and 512x424 for RGB camera and infrared camera, respectively. The reference plate is made of A4 size paper, with side length of 0.03 meter, using 5x7 checker board (Figure 5).

1041 pairs of images were collected for calibration using three Kinect v2 sensors, as listed in Table 2. The depth errors after calibration were presented in Figure 6, which shown that there existed varied system errors at the level of -0.02 ~ -0.04m. The

errors were less than ±0.03m for most points, and some random errors were found about ±0.08m.

Figure 5. Color image (left) and IR image (right) for calibration with chessboard 5x7x0.03 pattern

Distribution of errors in the XY plane is presented in Figure 7, and no uniform mode was detected after calibration, which proven these errors belongs to random error and could not be corrected by calibration.

20 high contrast target were set in 2.5 x 2 x 3m regions as indoor calibration field. More than 4 non-coplanar target could be measured for each Kinect sensor. Collection of ground truth data was accomplished using Riegl VZ-400 scanner with angle resolution of 0.02 degree.

Sensors	IR Image	Color Image	Sync Image
Sensor #1	105	112	112
Sensor #2	125	131	133
Sensor #3	109	106	108

Table 2. Images for calibration work

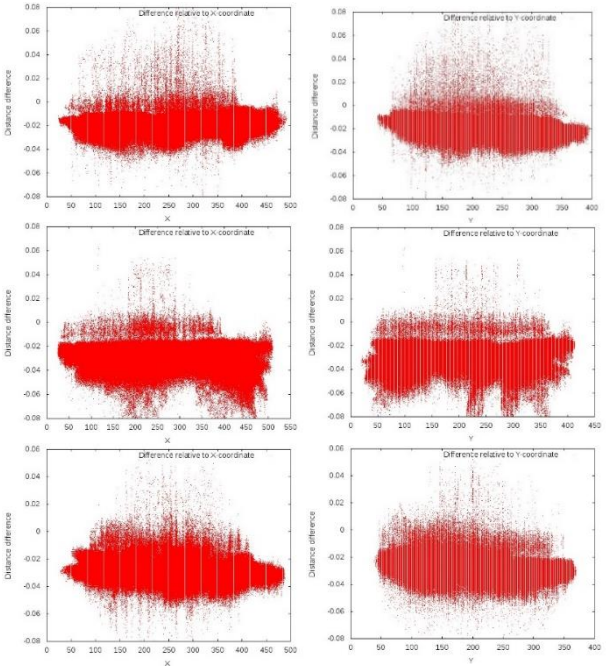


Figure 6. Difference distribution of each sensor relative to X-coordinate (left) and Y-coordinate (right)

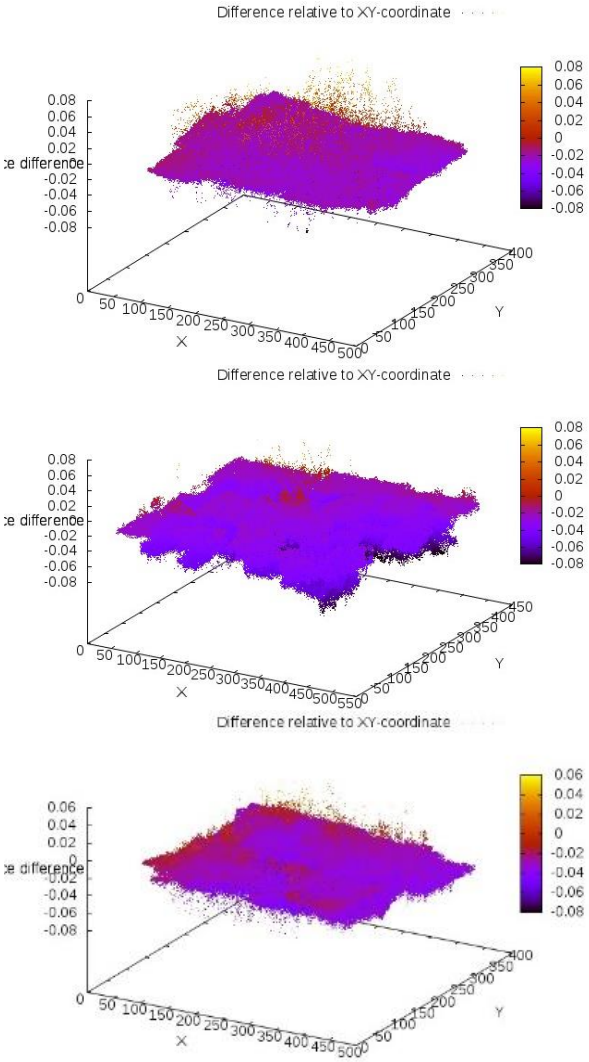
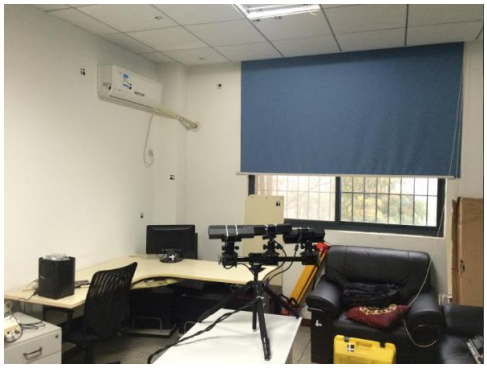
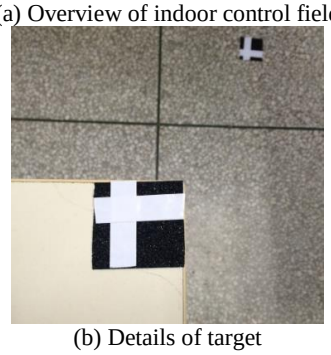


Figure 7. Difference distribution of each sensor in XY-plane



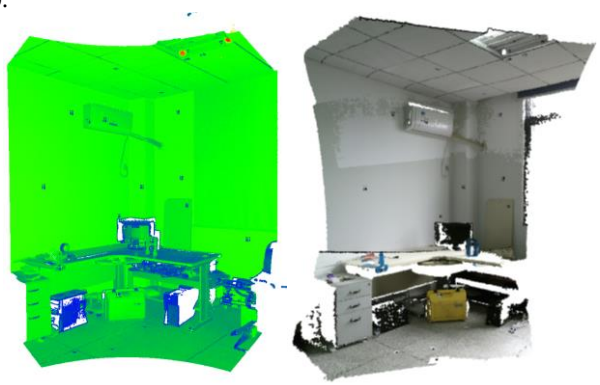
(a) Overview of indoor control field



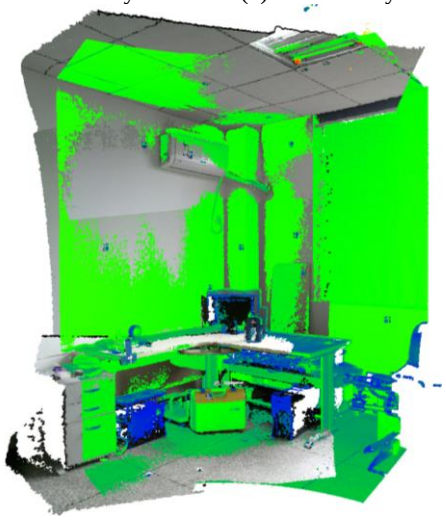
(b) Details of target

Figure 8.Calibrate sensor array with indoor control field

The center point coordinate of reference target were collected simultaneously from both the depth camera and the laser point cloud data, and the {R t} parameters were calculated following 2.2.4, and fitting to get the initial transformation $T_{init,}$ as in Figure 9.



(a) Ground truth by Lidar (b) Scan data by our array



(c) Overlapped

Figure 9. Initial transformation

Although data from different sensors has been transformed in the same coordinate system, errors still need to be reduced between sensor array and ground truth. Affected by system errors and multipath effect of the Kinect, more sample points are required to correct these errors. To minimize the depth errors, the Point-to-Plane ICP method was applied for base match between sensor array and ground truth, results of optimization were provided in Figure 10.

3.2 Experiment of data acquisition

The sensor array proposed in this paper does not equipped with IMU system, however, take advantages of large enough viewing angle, it is a typical SLAM system based on its estimations of the initial tracks using the visual odometry and closed - loop detection system with global optimization.

The Real - Time SLAM System of RTAB-Map (Labbe and Michaud, 2011, 2013, 2014) is adopted in this study, which is based on RGB-D SLAM method. With incremental appearance-based loop closure detector, this module used the BoW to distinguish whether to revisit map. When a closed loop was detected, a constraint is added in the system and overall adjustment is applied using graph optimizer.

However, RTAB-Map is lack of support for multi-source data input, and errors of route calculation are quite large. Therefore, only the parallel components were calculated using the software, and the other route were obtained from exterior orientation parameters of the sensors. Finally, data from key frame were extracted and converted to point cloud.

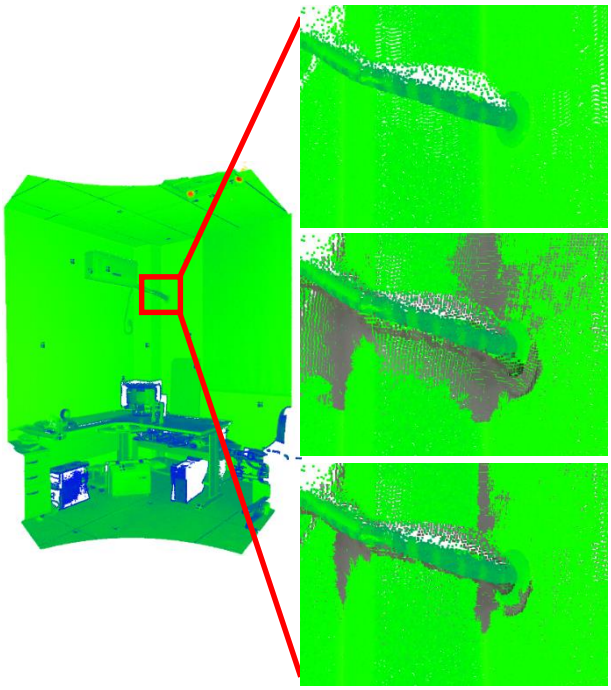


Figure 10. Improvement of ICP process
(from top to bottom are 1. LiDAR data 2. Initial transformation 3. ICP refined transformation)

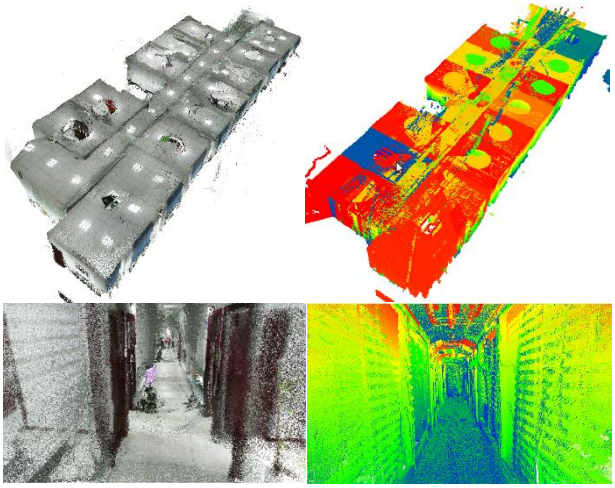


Figure 11. Point cloud captured by out sensor array (left) and LiDAR (right)

4. CONCLUSION

A novel method using multi consumer-level depth cameras for indoor data collection was proposed in this study, and experiment proven the efficiency of the method for indoor RGB point cloud data collection. The system is capable to meet the requirements of indoor mapping, modeling, robot localization and navigation, etc., with low precision demand, though the detailed information is not as good as production-level LiDAR system. In the future, more work will focused on the improvement of the stability of

the SLAM system based on wider angle viewing data, and explore to more application fields.

REFERENCES

Camplani, M., Mantecon, T., Salgado, L., 2013. Depth-Color Fusion Strategy for 3-D Scene Modeling With Kinect. *IEEE Transactions on Cybernetics* 43, 1560-1571.

Corti, A., Giancola, S., Mainetti, G., Sala, R., 2016. A metrological characterization of the Kinect V2 time-of-flight camera. *Robot. Auton. Syst.* 75, Part B, 584-594.

Diaz, M.G., Tombari, F., Rodriguez-Gonzalvez, P., Gonzalez-Aguilera, D., 2015. Analysis and Evaluation Between the First and the Second Generation of RGB-D Sensors. *IEEE Sensors Journal* 15, 6507-6516.

Fankhauser, P., Bloesch, M., Rodriguez, D., Kaestner, R., Hutter, M., Siegwart, R., 2015. Kinect v2 for mobile robot navigation: Evaluation and modeling, *2015 International Conference on Advanced Robotics (ICAR)*, Istanbul,Turkey pp. 388-394.

Gemignani, G., Capobianco, R., Bastianelli, E., Bloisi, D.D., Iocchi, L., Nardi, D., 2016. Living with robots: Interactive environmental knowledge acquisition. *Robot. Auton. Syst*. 78, 1-16.

Gui, P., Qin, Y., Hongmin, C., Tinghui, Z., Chun, Y., 2014. Accurately calibrate kinect sensor using indoor control field, *2014 3rd International Workshop on Earth Observation and Remote Sensing Applications (EORSA)*, Changsha, China, pp. 9-13.

Gutierrez-Gomez, D., Mayol-Cuevas, W., Guerrero, J.J., 2016. Dense RGB-D visual odometry using inverse depth. *Robot. Auton. Syst*. 75, Part B, 571-583.

Henry, P., Krainin, M., Herbst, E., Ren, X., Fox, D., 2014. RGB-D Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments, in: Khatib, O., Kumar, V., Sukhatme, G. (Eds.), Experimental Robotics: T*he 12th International Symposium on Experimental Robotics*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 477-491.

Huang A, R.N., Bachrach A, Henry P, Krainin M, Maturana D, Fox D, 2011. Visual Odometry and Mapping for Autonomous Flight Using an RGB-D Camera, *International Symposium on Robotics Research*, Flagstaff, AZ, USA.

Jiménez, D., Pizarro, D., Mazo, M., Palazuelos, S., 2014. Modeling and correction of multipath interference in time of flight cameras. *Image and Vision Computing* 32, 1-13.

Jiyoung, J., Joon-Young, L., Yekeun, J., Kweon, I.S., 2015. Time-of-Flight Sensor Calibration for a Color and Depth Camera Pair. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 1501-1513.

Labbe, M., Michaud, F., 2011. Memory management for real-time appearance-based loop closure detection, *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, USA, pp. 1271-1276.

Labbe, M., Michaud, F., 2013. Appearance-Based Loop Closure Detection for Online Large-Scale and Long-Term Operation. *IEEE Transactions on Robotics* 29, 734-745.

Labbe, M., Michaud, F., 2014. Online global loop closure detection for large-scale multi-session graph-based SLAM, *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014)*, Chicago, USA, pp. 2661-2666.

Nistér, D., Naroditsky, O., Bergen, J., 2006. Visual odometry for ground vehicle applications. *Journal of Field Robotics* 23, 3-20.

Sarbolandi, H., Lefloch, D., Kolb, A., 2015. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Computer Vision and Image Understanding* 139, 1-20.

Trimble, 2016. Trimble Indoor Mapping Solution. Retrieved from http://www.trimble.com/Indoor-Mobile-Mapping-Solution/Indoor-Mapping.aspx

Whelan, T., Kaess, M., Johannsson, H., Fallon, M., Leonard, J.J., McDonald, J., 2015. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research* 34, 598-626.

Yang, S., Yi, X., Wang, Z., Wang, Y., Yang, X., 2015. Visual SLAM using multiple RGB-D cameras, *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Zhuhai, China, pp. 1389-1395.

Yousif, K., Bab-Hadiashar, A., Hoseinnezhad, R., 2014. Real-time RGB-D registration and mapping in texture-less environments using ranked order statistics, *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on Cybernetics Chicago*, USA, pp. 2654-2660.