3D INDOOR MAPPING USING A RGB-D CAMERA SYSTEM

VẼ BẢN ĐỒ TRONG NHÀ 3D BẰNG HỆ THỐNG CAMERA RGB-D

Tran Cam Nhan, Nguyen Tan Nhu, Nguyen Thanh Hai
Ho Chi Minh City University of Technology and Education

Received 31/3/2016, Peer reviewed 23/5/2016, Accepted for publication 3/8/2016

ABSTRACT

This research proposes an application of a RGB Depth (RGB-D) camera system in 3D indoor mapping. In order to reconstruct the 3D model of an indoor space where the robot is located, the RGB-D camera system is installed in the robot frame to continuously capture the separated 2D image frames. All corresponding 2D points between the two consecutive image frames are firstly estimated using a Scan Invariant Feature Transform (SIFT) algorithm. Secondly, all pixel coordinates of these matching points are projected to the respective 3D space based on the depth information from the RGB-D camera at each pixel. The current outputs of this stage are the 3D points in the two successive point clouds that are applied to find out the transformation matrix. Finally, the estimated matrix is used to transform the second point cloud to the coordinating system of the first one. The result after repeating the above process with other consecutive pair of image frames and point clouds is the 3D model of the navigational space. The aimed method is relatively low cost configuration; furthermore, its accuracy is acceptable with the indoor robot applications.

Keywords: *RGB-D camera system, 3D indoor mapping, SIFT, transformation matrix, point clouds.*

TÓM TẮT

Nghiên cứu đề xuất một ứng dụng của một hệ thống camera RGB-D để xây dựng bản đồ 3D. Để dựng lại mô hình 3D của một không gian trong nhà, nơi robot hoạt động. Hệ thống camera RGB-D được gắn trên robot để liên tục chụp những ảnh 2D. Tất cả các điểm 2D tương ứng giữa hai ảnh liên tiếp được đánh giá sử dụng phương pháp biến đổi đặc tính quét bất biến (SIFT). Tiếp theo, tất cả các tọa độ điểm ảnh phối hợp được chiếu sang không gian 3D tương ứng dựa vào những thông tin chiều sâu từ camera RGB-D cho mỗi điểm ảnh. Các ảnh ngõ ra hiện tại của giai đoạn này là những điểm 3D trong hai ảnh đám mây điểm được sử dụng để tìm ra ma trận biến đổi. Cuối cùng, các ma trận ước tính được sử dụng để chuyển đổi ảnh đám mây điểm thứ hai cho hệ thống phối hợp. Kết quả sau khi lặp đi lặp lại quá trình trên với cặp ảnh khác và những ảnh đám mây điểm là mô hình 3D của không gian dò tìm. Phương pháp này nhắm đến cấu hình có chi phí tương đối thấp. Hơn nữa, độ chính xác của nó là chấp nhận được với các ứng dụng robot trong nhà.

Từ khóa: *hệ thống camera RGB-D; vẽ bản đồ 3D; thuật toán SIFT; ma trận biến đổi; ảnh đám mây điểm.*

1. INTRODUCTION

Robotic mapping is one of the most vital tasks in automatic robotic applications. The robot has to support the model of the navigational space in order to locate itself

when moving. Moreover, the map is essential for path planning processes in proposing the roadmap to target positions [1].

The robotic mapping is divided into 2D mapping and 3D mapping. The 2D mapping has some disadvantages compared with 3D mapping. For instance, the classical application of sonar in mapping and navigation [2], the 2D mapping strategy based on line segmentation [3], and the 2D mapping of a closed area by a range sensor [4]. One of the great drawback of these is the lack of information in the third space dimension [5]. Realizing the negative trends of that 2D planning methods, the 3D algorithms have been continuously developing with the supports of famous classical findings. The SLAM algorithm, for instance, was applied with 3D mapping methods to get the 3D model of large scale environments [6]. Consequently, the 3D mapping methods have recently become the trend of robotic mapping.

The quality of mapping processes was mainly determined by the methods of acquiring data from surroundings. Firstly, the sonar sensors were used to obtain the ranges from the robot to surrounding obstacles which are used to build a map [2]. The accuracy of sonar sensors were compensated by the advantages of 2d laser scanner [7]. The stereo cameras were also taken advantage of mapping [8], but the time consuming to reconstruct the ranging information from stereo images was costly. The KINECT RGB-D sensor developed by Microsoft in 2012 was considered to be suitable for robotics [9]. This kind of RGB-D sensors has not only the suitable accurateness but also the fast speed of processing time. All above reasons have been considered to be the motivation of developing the 3D mapping algorithm supported by the RGB-D camera system.

The paper introduces the process of concatenating the discrete 3D point clouds acquired from surroundings to form a complete 3D model of navigational spaces. The KINECT RGB-D sensor is firstly used to acquire the pair of consecutive images and their corresponding 3D point clouds. Next, the SIFT algorithm [10] is applied in finding the corresponding points in two 3D images. The 3D corresponding points are estimated by projecting the resultant 2D matching points getting from the previous step to 3D space. Finally, two consecutive 3D point clouds are concatenated by transforming the second point cloud to the first point cloud coordination. Repeating the process with the next two consecutive point clouds, the 3D model of navigational space is becoming larger.

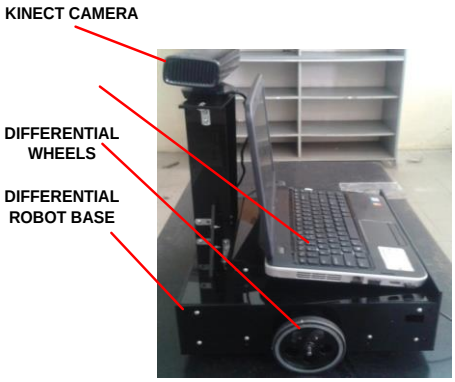

Fig.1 *The process of concatenating two consecutive point clouds*

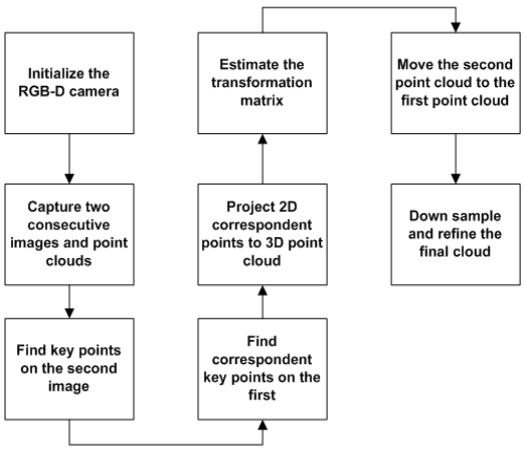

Fig.2 *The process of concatenating two consecutive point clouds*

The configuration of the Robot, presented in Figure 1, is composed of a KINECT camera, a differential robot base, a personal computer, and a differential driver. The KINECT camera is set fixedly in the highest level in order to enlarge the FOV (Field Of Vision), which is about 58.5×46.6 degrees. The RGB images captured by KINECT have the resolution of 640×480, and the resolution of depth images is 320×240. The range of depth images is from $0.8m$ to $4m$. As the robot is able to turn around by controlling the differential drives, the FOV can be even wider. These parameters are relatively suitable for applications in indoor environments.

The next section of the paper sequentially presents the above steps in more detail with some experimental illustrations.

2. METHODS

2.1 The3D mapping blocking diagram

The Figure 2 presents the procedures of concatenating two consecutive point clouds based on the proposed method throughout the paper. After initialized to the predefined parameters, the RGB-D camera is employed to capture both consecutive images and their corresponding point clouds. Two continuous point clouds are concatenated based on the matching information between two successive images. The key points on the first and second images are detected and described by SIFT algorithm before being able to find the correspondences between them. The correspondences in 2D are projected to 3D using geometric transformation method. These corresponding pairs are used to interpolate the transformation matrix, which is applied to the second point cloud to the coordination system of the first point cloud. The next cycle is going to be compiled after the first pair of point clouds is united. The final 3D model of the navigational space is, at last, down sampled to reject the overlapped points.

2.2 SIFT key point detection

In this research, the SFIT algorithm is applied for detecting the pixels coordinator in which their position is independent from scales [10]. Different scales of the original image are firstly computed by convoluting the image with the Gaussian function (1). Thus, scales of the original RGB image are represented by $L(x, y, k\sigma)$ which is estimated Equation (2), where k is called the scale factor. The extreme locations of the Laplace transformation $(\sigma^2\nabla^2 G)$, which is presented in Equation (3), are scale-invariant pixel locations on the RGB image of surroundings. $(\sigma^2\nabla^2 G)$ can also be computed by Equation (4).

$$G(x, y, k\sigma) = \frac{1}{2\pi k^2\sigma^2} e^{-\frac{x^2+y^2}{2k^2\sigma^2}} \quad (1)$$

$$L(x, y, k\sigma) = I(x, y) * G(x, y, k\sigma) \quad (2)$$

$$\frac{\partial G}{\partial \sigma} = \sigma\nabla^2 G \approx \frac{G(x,y,k\sigma) - G(x,y,\sigma)}{k\sigma - \sigma} \quad (3)$$

$$\sigma^2\nabla^2 G \approx \frac{G(x,y,k\sigma) - G(x,y,\sigma)}{(k-1)} \quad (4)$$

The Laplace transform of the Gaussian function, $(\sigma^2\nabla^2 G)$ is approximately equal to the difference of continuous scale versions of $L(x, y, k\sigma)$, shown in the equation (5). The scale invariant pixel locations are finally the minimum or maximum locations of $D(x, y, \sigma)$.

$$D(x, y, \sigma) = \big(G(x, y, k\sigma) - G(x, y, \sigma)\big) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (5)$$

2.3 Key point matching

After located on the 2D image, every key point is described by a 128-dimension vector. Assuming that $\boldsymbol{M}(x_1, x_2, x_3, \ldots, x_{128})$ is a set of key point feature vectors located on the

first image and $\boldsymbol{N}(y_1, y, y_3, \ldots, y_{128})$ is one on the second image. All key points on the first image are found individually their correspondences on the second image. Two key points are consider to be the same if the Euclidean distance d_k between them is less than a pre-determined constant δ as shown the following equation

$$d_k = \sqrt{(x_1 - y_1)^2 + \cdots + (x_{128} - y_{128})^2} \quad (6)$$

2.4 2D-to-3D projection

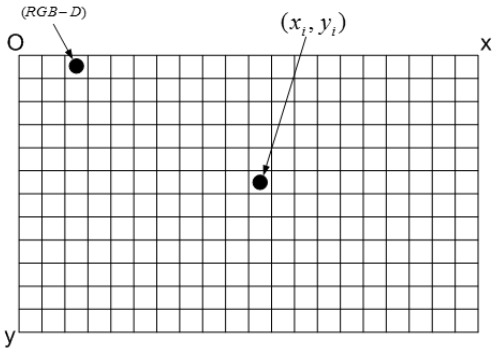

Fig.3*A RGB-D matrix data acquired from the KINECT sensor*

The process of projection has to take advantages of both pixel coordination and depth value from the RGB-D data. Each RGB-D pixel in a matrix data acquired from KINECT sensor, as shown in the Figure 3, is composed of two classes of information that are Red-Green-Blue color parts and depth value.

For calculation of 3D coordinates, every pixel k is only characterized by pixel coordinates symbolized by (x_k, y_k), so it is necessary that the depth information z_k must be employed to obtain the 3D coordinate of that pixel position. The method of the 2D-to-3D projection is illustrated in Equations (7), (8), and (9). As expressed in these equations, an individual 3D point (X_k, Y_k, Z_k) is calculated based on pixel coordination presented as (x_k, y_k), depth z_k, and physical characteristics of the RGB-D camera a. The parameter a is estimated during the calibration process. The origin of 3D coordination system is placed on the center of the 2D image, so the pixel location is shifted to a distance of (x_0, y_0), which is the center coordinates of the 2D image, as shown in Figure 4.

$$X_k = -(x_k - x_0) \times a \times z_k \quad (7)$$

$$Y_k = -(y_k - y_0) \times a \times z_k \quad (8)$$

$$Z_k = z_k \quad (9)$$

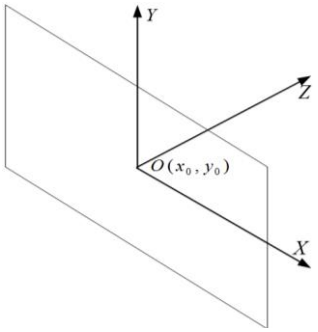

Fig.4*A 3D coordination system on a 2D image*

2.5 3D rotation and transformation

Suggested that there are two sets of 3D points which are $\boldsymbol{P_0}$ and $\boldsymbol{P_1}$ acquired from the first and second view of the RGB-D camera. The point cloud P_0 is based on the first coordination system O_0xyz, and the point cloud P_1 is based on the different coordination system O_1xyz. The rotation and translation matrix $\boldsymbol{M_{10}}$ formed as the equation (10) has its responsibility of moving the point cloud P_1 to the P_0's coordination, which is solved by the formula (11) and (12).

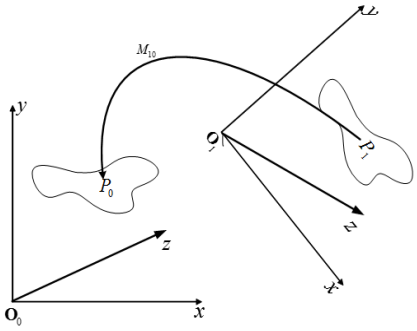

Fig.5 *The* $\boldsymbol{P_1}$*-to-*$\boldsymbol{P_0}$ *transformation matrix*

$$\boldsymbol{M_{10}} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (10)$$

$$\boldsymbol{P_0} = \boldsymbol{P_1} \times \boldsymbol{M_{10}} \quad (11)$$

$$\begin{bmatrix} x_{Mi} \\ y_{Mi} \\ z_{Mi} \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} x_{Ni} \\ y_{Ni} \\ z_{Ni} \\ 1 \end{bmatrix} \quad (12)$$

2.6 3D point cloud concatenation

All 3D point clouds are concatenated together based on the pair transformation matrices. Supposed that there are 4 3D point clouds based on discrete coordination systems, that is P_0, P_1, P_2, P_3 based on the coordination systems of O_0xyz, O_1xyz, O_2xyz, and O_3xyz respectively, as shown in the Figure 6. Moving to the first coordination systems, theP_1has to be applied by the matrix M_{10} like the equation (13). Compiling the same process above in these equation (14), (15), the point clouds P_2, P_3 are able to transformed to the first coordination system. The matrices M_{10}, M_{21}, and M_{32} are inferred from the process of concatenating the two next consecutive point clouds. The result of 3D model of robot's moving space is discussed in the next section of the paper.

$$P_1(O_0xyz) = P_1(O_1xyz) \times \boldsymbol{M_{10}} \quad (13)$$

$$P_2(O_0xyz) = P_2(O_2xyz) \times \boldsymbol{M_{20}} \quad (14)$$

$$P_3(O_0xyz) = P_3(O_3xyz) \times \boldsymbol{M_{30}} \quad (15)$$

$$\boldsymbol{M_{20}} = \boldsymbol{M_{21}} \times \boldsymbol{M_{10}} \quad (16)$$

$$\boldsymbol{M_{30}} = \boldsymbol{M_{32}} \times \boldsymbol{M_{21}} \times \boldsymbol{M_{10}} \quad (17)$$

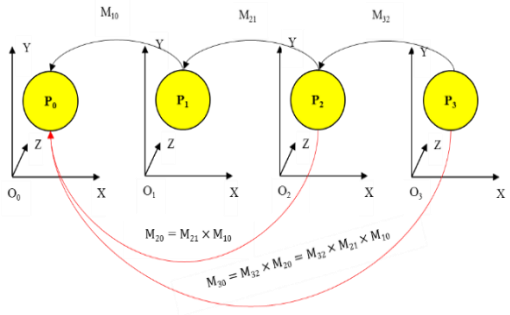

Fig.6 *Different 3D point clouds with individual coordination system*

3. RESULTS AND DISCUSSIONS

This section presents results of cloud image acquisition and then processing image frames to produce 3D images.

3.1 Image and point cloud acquisition

Fig.7 *Two consecutive RGB images captured from KINECT sensor*

KINECT sensor used as the RGB-D camera is considered to be the most suitable input device in the paper. The data acquired from KINECT is comprised of RGB and depth images, whose accuracies are able to be accepted by indoor robotic applications. Moreover, the depth data is pre-processed by KINECT's hardware, so the higher level threads can focus on useful tasks. Figure 7shows two consecutive images captured by KINECT. According the each image, the two continuous 3D point clouds are also taken, as shown in the Figure 8. Two images and point clouds contain both new and old information; consequently, the combination of them is expected to cover more additional data. The next step of concatenating process is to estimate the pixel locations of key points from the two images.

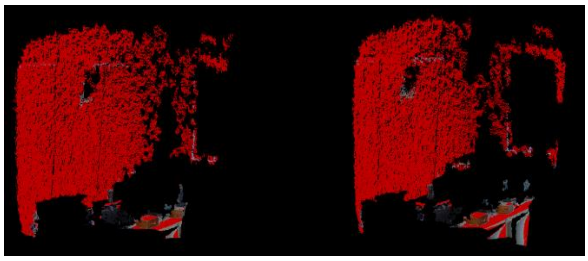

Fig.8 *Two consecutive point clouds captured from KINECT sensor*

3.2 2D key point estimation

The SIFT algorithm is applied to locate the key points in both first and second images whose characteristics are independent from scales and rotations. The Figure 9 shows these key points which are marked in the white points. The key points are mainly focused on the areas where the differences in gray levels are high. Next, each detected key point is described by a 128-dimension vector according to the SIFT method presented in [10] to be able to recognize easily by Euclid's distance equation (6). The two key points are consider to be matched if the distance between their own vectors is less than a pre-defined constant, called ∂.

Fig.9 *Key points on the first and second 2D images*

3.3 3D key point matching

After estimated and described by SIFT, the key points on the second image are matched with their correspondences if they satisfy the criteria in (6) between the two pair of vectors. Figure 10 shows all corresponding key points on the first and second images. Each matching is presented by a green line connected between two key points. The number of matchings is over 200 pairs. Each matching, after that, is projected to the 3D space by using the equation (7), (8), and (9). In some cases, the depth information in a specific pixel on 2D images is not able to determine due to some incorrect refection, so the matching point at that pixel location have to be eliminated. The remaining pairs of matching key points are just over 100 pairs in the case illustrated by the Figure 11 where the matching cases are presented by the red lines.

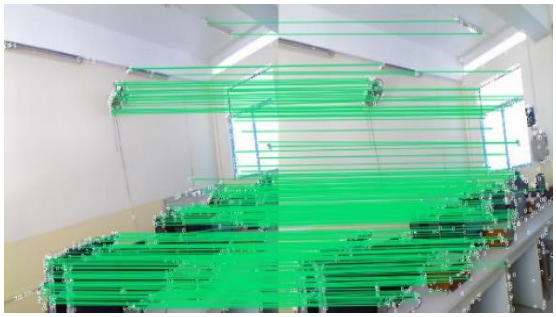

Fig.10 *Matching key points between the first and second 2D images*

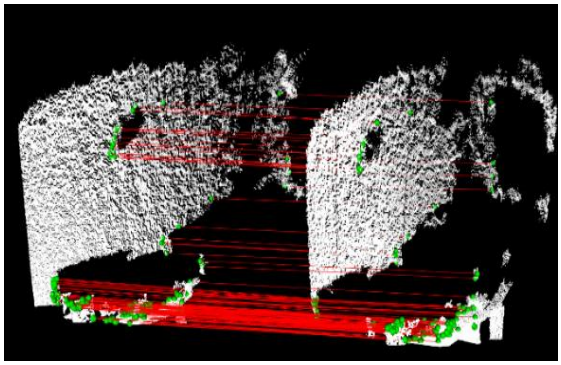

Fig.11 *Matching key points between the first and second 3D point clouds*

3.4 Pair concatenation

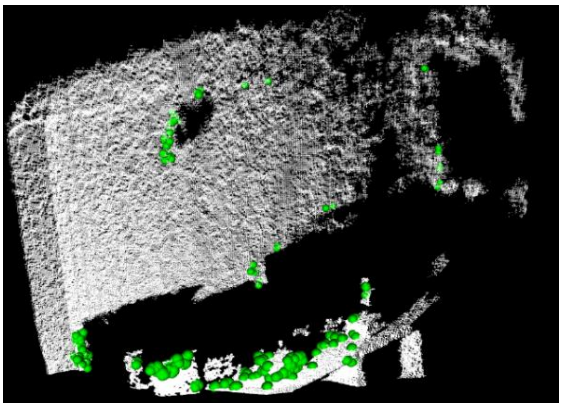

Fig.12 *The concatenated point cloud*

As shown in the equation (10) the transformation matrix has the size of 4×4, in which 12 parameters are unknown; therefore, there must be at least 12 pair of corresponding points in 3D cloud in order to infer the correct values based on the system of Equation (12). However, the number of

correspondences is often larger than 12 due to the errors happened when matching two point clouds. After the transformation matrix is found out, the matrix is applied to the second point cloud so as to have the same coordinating system as the first point cloud. The result of the concatenating point cloud is described in the Figure 12.

3.5 Consecutive concatenation

The Figure 13 shows one frame of the robot's moving space. In this view, there are some artificial landmarks organized randomly in the moving space. When the robot moves straight, the 3D model of this space is illustrated on the Figure 14. In the case of curving, some artificial landmarks are also set statically on the curving line, which is presented on the Figure 15. However, the result of 3D model is not formed very well, as shown in the Figure 16. When the number of landmarks is increased, as shown in the Figure 17, the 3D model of the curving space becomes smoother, which is presented in the Figure 18. The result is also proposed that the 3D model of an entire room is able to be acquired when the robot is static and rotated around this room. The 3D model of an entire room is displayed on the Figure 19.

The measured values between two opposite walls presented in Figure 19 are listed in Table 1. The measured error is 0.023%.

Fig.13 *The first frame of the robot's moving space*

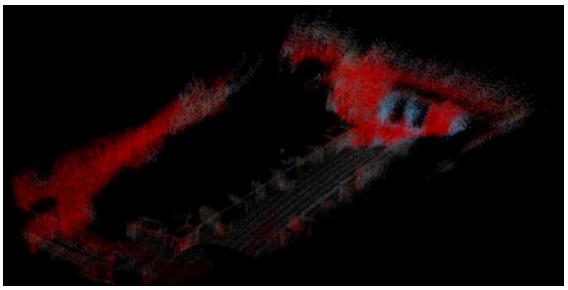

Fig.14 *The second 3D view of the straight moving space*

Fig.15 *The landmarks set on the curving space*

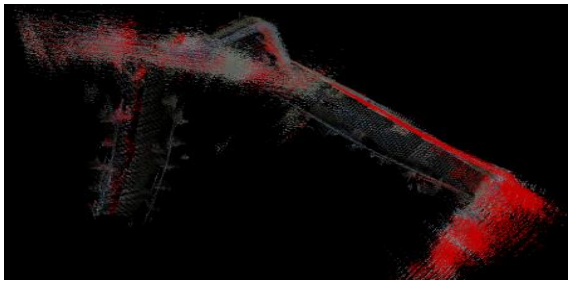

Fig.16 *The interrupted 3D model of curving space*

Fig.17 *The increased number of landmarks*

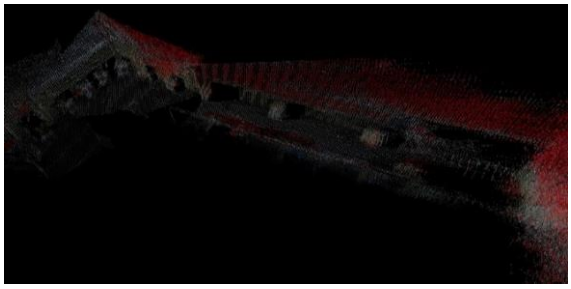

Fig.18 *The smoother curving space when increasing the number of landmarks*

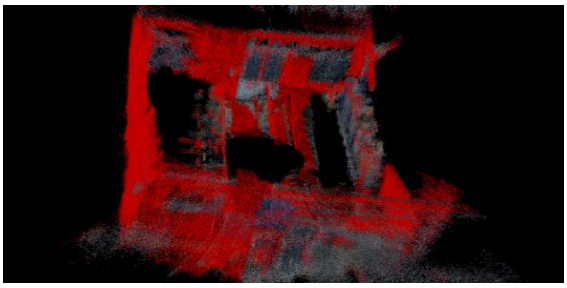

***Fig.19** The first view of a 3D model of the sample room where the robot is rotated*

***Table 1.**The comparison between real and measured width of the 3D model of the room*

Real Values (m)	Measured Values (m)	Real Values (m)	Measured Values (m)
7	6.9	10	9.96
7	7.1	10	9.98
7	6.92	10	10.01
7	7.06	10	9.93
7	6.95	10	9.95
7	6.99	10	9.99
7	7.03	10	10.05
7	7.07	10	10.1
7	6.97	10	9.97
7	7.08	10	10.03

Although there have been various methods of 3D reconstruction recently, the proposed method is still valuable in terms of time consuming and reliability. Stereo cameras could be applied in 3D reconstruction [11], but this method was time-consuming because the depth information had to be converted from stereo images. As the KINECT is an independent system outputting both RGB images and depth images automatically, the processing time is much smaller. Moreover, the depth information is strongly based on the RGB stereo images, which vitally depend on the light conditions, so the accuracy of this method is less reliable than one of the proposed method. Furthermore, the method is more suitable with the data getting from KINECT sensors, as including many noises, than the method having presented in [12], where the separated point clouds are concatenated together just by RANSAC and ICP algorithms. While the RANSAC-ICP algorithms are based only on the 3D data, the method in this paper uses both 2D RGB images and depth images to merger two consecutive point clouds.

4. CONCLUSIONS

In the paper, the 3D model of both navigational spaces in indoor areas was completely reconstructed based on RGB-D image frames obtained from a KINECT system. A SIFT algorithm was employed to detect pixels coordinator to perform the optimal 3D model of navigation spaces. Experimental results proved that the KINECT system could be cheaper cost computation for solution of replacing other stereo cameras with higher cost. Moreover, it showed the effectiveness of the proposed method.

REFERENCES

[1] J. Fuentes-Pacheco, Visual simultaneous localization and mapping: a survey, in *Springer Science+Business Media Dordrecht*, ed, pp. 55–81, 2015.

[2] A. Elfes, Sonar-based real-world mapping and navigation, in *IEEE Journal on Robotics and Automation* vol. 3, ed, pp. 249-265, 1987.

[3] C. P. O. Diaz and A. J. Alvares, A 2D-mapping strategy based on line segment extraction, in *ANDESCON, 2010 IEEE*, ed, pp. 1-6, 2010.

[4] S. Ogawa, K. Watanabe, and K. Kobayashi, 2D mapping of a closed area by a range sensor, in *SICE 2002. Proceedings of the 41st SICE Annual Conference* vol. 2, ed, pp. 1329-1333 vol.2, 2002.
[5] R. Z. M. Dr.Wael R. Abdulmajeed, Comparison Between 2D and 3D Mapping For Indoor Environments, in *International Journal of Engineering Research and Technology* vol. 2, ed, 2013.
[6] J. L. Cras and J. Paxman, A modular hybrid SLAM for the 3D mapping of large scale environments, in *Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on*, ed, pp. 1036-1041, 2012.
[7] L. Hung-Hsing, T. Ching-Chih, H. Ssu-Min, and C. Hsu-Yang, Automatic mapping for an indoor mobile robot assistant using RFID and laser scanner, in *SICE, 2007 Annual Conference*, ed, pp. 2102-2108, 2007.
[8] D. Schleicher, L. M. Bergasa, R. Barea, E. Lopez, and M. Ocana, Real-Time Simultaneous Localization and Mapping Using a Wide-Angle Stereo Camera, in *Distributed Intelligent Systems: Collective Intelligence and Its Applications, 2006. DIS 2006. IEEE Workshop on*, ed, pp. 55-60, 2006.
[9] C. Lim Chee, S. N. Basah, S. Yaacob, M. Y. Din, and Y. E. Juan, Accuracy and reliability of optimum distance for high performance Kinect Sensor, in *Biomedical Engineering (ICoBE), 2015 2nd International Conference on*, ed, pp. 1-7, 2015.
[10] D. G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, in *International Journal of Computer Vision* vol. 60, ed, pp. 91-110, 2004.
[11] M. Wei-wei, L. My-Ha, and J. Kang-Hyun, 3D reconstruction and measurement of indoor object using stereo camera, in *Strategic Technology (IFOST), 2011 6th International Forum on* vol. 2, ed, pp. 738-742, 2011.
[12] X. Huang and M. Hu, 3D Reconstruction Based on Model Registration Using RANSAC-ICP Algorithm, in *Transactions on Edutainment XI*, Z. Pan, D. A. Cheok, W. Mueller, and M. Zhang, Eds., ed. *Berlin, Heidelberg*: *Springer Berlin Heidelberg*, pp. 46-51, 2015.

Corresponding author:
Nguyen Thanh Hai
Ho Chi Minh City University of Technology and Education
Email: nthai@hcmute.edu.vn