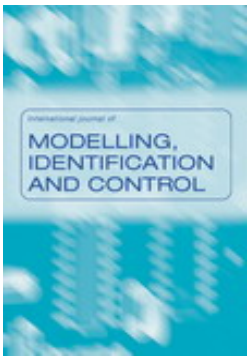


International Journal of Modelling, Identification and Control

ISSN online: 1746-6180 - ISSN print: 1746-6172
https://www.inderscience.com/ijmic

3D indoor reconstruction using Kinect sensor with locality constraint

Peng Zhu, YanGuang Guo

DOI: 10.1504/IJMIC.2023.10053829

Article History:

Received:	12 November 2021
Last revised:	17 January 2022
Accepted:	04 February 2022
Published online:	03 February 2023

3D indoor reconstruction using Kinect sensor with locality constraint

Peng Zhu and YanGuang Guo*

Department of Computer Technology and Information Management,
Inner Mongolia Agricultural University,
Baotou, Inner Mongolia, 014109, China
Email: zhup@imau.edu.cn
Email: luozhangyulong@163.com
*Corresponding author

Abstract: In this paper, an indoor 3D construction is proposed based on RGB-D measurement. It is intentionally designed to solve the traditional issues, such as cloud registration inaccuracy, large computational time. Firstly, potential candidates are extracted by Harris detector, and the SURF method is used to generate the feature descriptors. Afterwards, the correct functional match is selected by RGB and depth measurements with neighbouring constraint. Lastly, 3D clouds are formed through graphical optimisation. In the experiment, the RGB-D sensor is rigidly fixed on the mobile platform to reconstruct the indoor 3D scene, which shows comparable performance in terms of computational time and accuracy.

Keywords: RGB-D; 3D indoor reconstruction; Kinect; point cloud; SURF method.

Reference to this paper should be made as follows: Zhu, P. and Guo, YG. (2023) '3D indoor reconstruction using Kinect sensor with locality constraint', *Int. J. Modelling, Identification and Control*, Vol. 42, No. 1, pp.46–53.

Biographical notes: Peng Zhu received his BS degree in Communication Engineering from Inner Mongolia University of Technology, HuHeHaoTe, China in 2009 and MS degree in Communication and Information Systems from YanShan University, QinHuangDao, China, in 2012. Currently, he is a Lecturer in the Department of Computer Technology and Information Management, Inner Mongolia Agricultural University. His research interests include issues related to digital image processing, embedded development. He is author of a great deal of research studies published at national and international journals, conference proceedings.

YanGuang Guo received her BS degree in Computer Science and Technology from Inner Mongolia Agricultural University, HuHeHaoTe, China in 1998 and MS degrees in Computer Course and Teaching Theory from Inner Mongolia Normal University, HuHeHaoTe, China in 2009. Currently, she is a Professor in the Department of Computer Technology and Information Management, Inner Mongolia Agricultural University. Her research interests include issues related to digital image processing, computer network communication. She is author of a great deal of research studies published at national and international journals, conference proceedings.

1 Introduction

Three-dimensional reconstruction of indoor scenes is a hot research topic in the field of computer vision, Zheng et al. (2020), Zhou et al. (2021) and Yang et al. (2021) propose 3D reconstruction based on different technologies respectively, and it is an important part of mobile robot autonomous navigation and unknown environment model reconstruction.

In recent years, due to the rapid development of the manufacturing technology of image and video shooting equipment, the available video language is becoming more and more diversified and complicated, and its access is becoming more and more convenient. The new opportunities for the development of related computer intelligence application technology have been provided by the rapid development of multi-modal video image synchronisation recording equipment, especially multimedia video security monitoring, and a series of research topics and applications based on multi-modal cameras emerge one after another. Especially after the appearance of RGB-D cameras, people began to try to use a new way (depth information) to solve the traditional problems of computer vision, pattern recognition and computer graphics (Wang et al., 2017; Zhang et al., 2018).

In addition to providing colour images, RGB-D cameras can also provide depth information corresponding to image pixels (Jacob et al., 2020). This makes up for the defects of large amount of computation and poor real-time performance when using binocular cameras to reconstruct scenes, and the lack of depth information when using monocular cameras to reconstruct scenes. And the existing RGB-D cameras, for example, Microsoft's Kinect and Asustek's Axuson cameras are cheap, which are completely suitable for 3D reconstruction of indoor scenes and Dulko

(2018) uses Microsoft Kinect 2.0 for 360° colour and depth mapping of an indoor room research.

However, the range of depth measurement collected by RGB-D camera is limited, and the measured value also contains noise, and its stability is not high. In addition, according to Zhang et al. (2018), the depth image will have a void area. These defects greatly limit the application of RGB-D camera in 3D reconstruction.

In recent years, researchers have applied RGB-D camera to 3D reconstruction of indoor scenes and target objects, and obtained some research results. Such as Atman and Trommer (2018), Nicastro et al. (2019), Li et al. (2021) and Zhu et al. (2019) improved the algorithm of feature point extraction and matching in 3D reconstruction by using feature points extracted from Kinect colour images, and built SLAM system based on Kinect. Unstable feature points are eliminated to make the algorithm more efficient. However, the processing speed is still slow and the accuracy is not high. Cai et al. (2020) designed a fast, efficient and low-cost 3D reconstruction system based on Kinect v2, which is not limited by equipment and hardware, but can only rebuilt for a single static object, and the reconstructed triangular mesh surface is rough and has certain errors. Altmann et al. (2020) proposed a 3D reconstruction algorithm in dynamic scenes. In order to reduce the processing time needed to build histograms, this algorithm proposed for the first time to consider the spatial-temporal structure in pixels of different dynamic scenes. However, the accuracy is greatly influenced by the light intensity of the detected scene and the number of objects being detected. Papadopoulos and Daras (2018) proposed a three-dimensional flow descriptor, which was used to represent the spatial information and surface information of objects, effectively solved the problem of defining three-dimensional direction, and introduced time dimension to encode global motion characteristics. Han et al. (2020) put forward a three-dimensional segmentation scheme based on occupancy of the perception. After clustering samples, excessive segmentation is avoided, thus maintaining high efficiency. Hu et al. (2017) put forward a joint learning model of action recognition based on RGB-D to improve the accuracy of 3D reconstruction, but it has some shortcomings such as slow matching speed of RGB image feature points.

In this paper, an improved RGB-D 3D reconstruction method with constraints is proposed to solve the problems of slow matching speed and insufficient robustness of RGB image feature points between adjacent frames in the process of 3D reconstruction in previous literature.

Main contributions of this paper are as follows:

1 In most 3D reconstruction algorithms, it is difficult to achieve speed and accuracy at the same time. In this paper, in the feature point matching between adjacent frames, the depth information of feature points and local nearest neighbour feature points are used as constraints to ensure the accuracy of feature point matching and can also speed up the process of feature point matching.

2 The carrier attitude estimation information solved by correctly matched feature point pairs can be used as the initial value to introduce into RANSAC sample consistency analysis, which can ensure the rapidity and global optimality of attitude estimation.

3 In previous experiments with indoor 3D reconstruction techniques, comparisons of performance in terms of computation time and accuracy are often missing. An autonomous navigation car system with RGB-D camera is built, which can move and collect scene information in indoor environment, and complete 3D reconstruction of indoor scene and motion trajectory estimation on ROS operating system.

2 RGB-D sensor and pinhole camera model

The RGB-D camera used in this paper is Kinect sold by Microsoft, which is widely used at present (Simonsen et al., 2017). The three lenses of Kinect are 3D infrared structured light emitter, RGB colour image camera and structured light depth sensor. The structured light emitter and the structured light depth sensor jointly solve the depth information of the scene image, as shown in equation (1).

$$D_{IR} = L_{IR} \times f_{IR} / d_{IR} \tag{1}$$

In which D_{IR} is the depth value of scene, L_{IR} is Kinect baseline length, f_{IR} is the focal length of structured light depth sensor, and d_{IR} is parallax.

In this paper, the RGB colour camera adopts pinhole imaging model, and any point in the space is projected onto the imaging plane through the optical centre C of the lens, as shown in Figure 1. Any point P in the world coordinate system, its projection point in the image plane is p, and they are related by the camera matrix:

$$p = C_{3\times4} \cdot P = K_{3\times3} \cdot M_{3\times4} \cdot P \tag{2}$$

$$P = (X, Y, Z, 1)^T, \quad p = (x, y, z, 1)^T \tag{3}$$

P and p are expressed by homogeneous coordinates, and the symbol '~' indicates that the left and right sides of the equation differ by a scale factor; $M_{3\times4}$ indicates the external parameter matrix of the camera; $K_{3\times3}$ represents the internal reference matrix of the camera.

Points under the world coordinate system and points under the camera coordinate system are associated by external parameter matrix $M_{3\times4} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}$ by the rotation matrix R and the translation vector T. Points in the camera coordinate system with the points in the image coordinate system are associated by internal reference matrix.

$$K = \begin{bmatrix} f_\alpha & 0 & \mu \\ 0 & f_\beta & v \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

In equation (4), sum represents the focal ratio of the camera in X direction and (direction), (u, v) indicates the intersection point of the optical centre principal axis in the image plane, that is, the principal point.

Figure 1 Camera pinhole imaging model

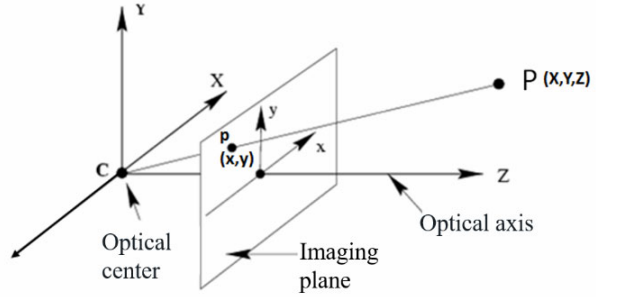
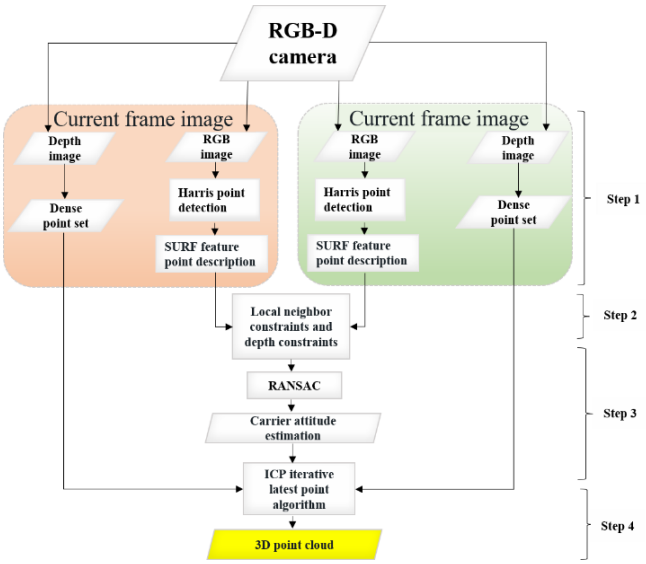


Figure 2 Flow chart of three-dimensional reconstruction algorithm (see online version for colours)



3 Three-dimensional reconstruction algorithm flow

The overall implementation flow frame of 3D reconstruction algorithm in this paper is shown in Figure 2. Firstly, RGB images are acquired by GRB-D camera for greyscale transformation, and Harris corner detector is used to detect the feature points of the images. The speeded up robust feature (SURF) algorithm is used to generate 64-dimensional feature point descriptors for each feature point, which has good orientation invariance and robustness under different illumination conditions. Secondly, feature point matching is performed on the feature point set between two adjacent images by using local nearest neighbour constraint and depth constraint. Thirdly, the matching results are introduced into random sample consistency analysis (RANSAC) to get the camera attitude. RANSAC is a method often used to extract feature data, it shows good effect and robustness in the rule construction of extracting point cloud data. Finally, using the dense point set and the attitude estimation information of the carrier obtained from RGB-D depth image, the final 3D point cloud is generated by graph optimisation method. Key point matching includes three parts: key point detection, feature description and key point feature matching. Each step has a mature algorithm to complete its specific function.

4 Detection and matching of feature points between adjacent frames

Feature extraction is an important part of computer vision, which mainly consists of key point extraction and descriptor calculation. The extraction of keys can be traced back to the 2D key detector. Harris is an efficient angle detection algorithm, which has the advantages of simple calculation and uniform extraction of angle points, but does not have scale invariance, and the extracted feature points have no relevant feature descriptors. SIFT key points are spot detection (blob), which has been proven to be the most advanced method for image classification, location recognition and other tasks, but with high time complexity and long algorithm time. Since then, approximate SURF key points based on the Hessian matrix are proposed for faster detection. Common feature point detection is usually divided into two categories: one is corner detection, such as Harris corner and FAST corner. One is patch feature points, such as SIFT, SURF, CENSURE. Corner detection method has fast calculation speed, but the detected feature points have scale uncertainty. However, the speckle detector is time-consuming, but the feature points found have higher discrimination and scale invariance. Considering that RGB-D is loaded on a mobile navigation car with a relatively fast travelling speed, we use Harris corner detector with a faster computing speed. Harris corner detection algorithm is to design a local detection window in the image. The window moves slightly in different directions, while the change of the average energy value of the window is investigated.

The Harris corner detection algorithm defines autocorrelated functions for each point in the image. The Harris algorithm obtains greyscale change value in each direction. The location of pixels is calculated with significant greyscale changes by an autocorrelation function. Afterwards, a correlation matrix of related functions is constructed. The corresponding corner point location information is obtained by comparing the eigenvalues of the constructed matrix, as shown in equation (5).

$$C(x, y) = \sum\nolimits_{\omega} \left[I(x_i, y_i) - I(x_i + \Delta x, y_i + \Delta y) \right]^2 \tag{5}$$

where $(\Delta x, \Delta y)^T$ is a given displacement, (x_i, y_i) is the point in the window ω.

After local feature detection, there is no direct operation such as image classification and recognition, feature matching, etc. A method must be adopted to describe the local image area, namely feature descriptor. Common feature point descriptors are SIFT, SURF, ORB (Yu et al., 2021), etc. SIFT algorithm has rotation invariance, scale

invariance, brightness invariance and good noise immunity, but the time complexity is high and the algorithm is time-consuming, therefore, good discrimination and easy calculation to generate 64-dimensional feature vectors for feature points are used in this paper. In Luo et al. (2020), SURF is a fast feature point detection algorithm, and the feature extraction process of this algorithm is as follows:

1 Detecting extreme points in scale space: SURF algorithm uses different box filters to process the original image and form an image pyramid. Then, Hessian matrix is used to detect the extreme points of each layer in the image pyramid.

2 Locating feature points: the Hessian matrix is used to find the extremum of the scale image, and the non-maxima are suppressed in the 3×3×3 stereo neighbourhood of the extremum point. Then interpolation operation is carried out in the scale space and image space, and the stable feature points and their scale values are obtained.

3 Determining the main direction: firstly, taking the characteristic point as the centre, calculating Harr wavelet responses of points in the neighbourhood with radius of 6s (s is the scale value of the characteristic point) in horizontal and vertical directions, then giving Gaussian weight coefficients to these response values, then accumulating the responses within 60 degrees to form a new vector, and finally traversing the whole circular area, select the longest vector direction as the main direction of feature points.

4 Generating feature point descriptors: taking the feature points as the centre, rotating the coordinate axis to the main direction, selecting a square area with a side length of 20s (sampling step length) according to the main direction, dividing the window area into 4×4 sub-areas, and calculating the wavelet response within the range of 5s×5s. The horizontal and vertical Haar wavelet responses relative to the main direction are a_x, a_y, and the same response value coefficients are assigned, and then the response coefficients of each subregion and their absolute values are summed to form the vector $V = \left(\sum ax, \sum ay, \sum |ax|, \sum |ay|\right)$.

   Therefore, a 64-dimensional feature description vector is generated for each feature point, and then the vector is normalised so as to be robust to illumination.

5 Depth information and local nearest neighbour feature point constraints

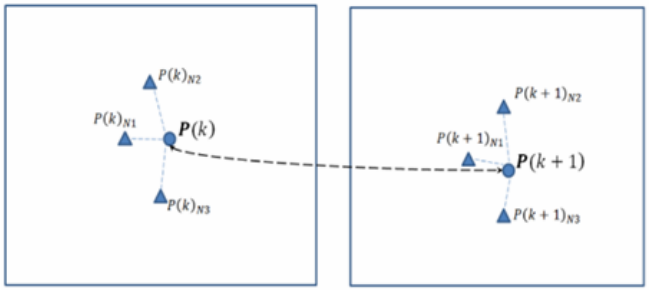
Existing methods usually find the minimum Euclidean distance between feature point descriptors in finding feature point matching pairs. These methods often lead to mis-matching of feature point pairs due to the periodic texture features appearing in the global search of images, which will seriously affect the attitude estimation of subsequent carriers and the 3D reconstruction of indoor scenes, Bhowmik et al. (2020). Considering the disadvantages of the above matching method based only on Euclidean distance, this method combines the local information and depth information of the image to constrain the matching process. This method is based on two assumptions:

Assumption 1 In the K frame image, feature points and their adjacent feature points are still adjacent feature points in the corresponding $K + 1$ frame image

Assumption 2 There is little difference between the two depth values corresponding to the feature point matching pairs of two adjacent frames of images.

The local nearest neighbour constraint means that a local graph is constructed for each feature point, and the set of feature points is first detected from the input image separately and used as nodes in the local graph, and then a star graph is constructed for each feature point using the detected feature points and their n nearest neighbour feature points. In the feature matching process based on the local nearest neighbour graph, the correspondence is obtained by two local graph sets. The neighbour constraint assumption of feature points is shown in Figure 3. In order to better understand the following algorithm, first define some variables, as shown in Table 1.

Figure 3 Distance constraint of nearest neighbour feature points corresponding to feature points, (a) K frame image (b) K + 1 frame image (see online version for colours)



(a) (b)

Assuming that a matching point pair between two adjacent images has been found: ($P(k)$, $P(k + 1)$). There are three neighbouring feature points around the feature point $P(k)$: $P(k)_{N1}$, $P(k)_{N2}$, $P(k)_{N3}$. The feature points corresponding to these three feature points in the $(k + 1)$ image are: $P(k + 1)_{N1}$, $P(k + 1)_{N2}$, $P(k + 1)_{N3}$. Intuitively, if matching point pairs ($P(K)$, $P(k + 1)$) exist, then ($P(k)_{N1}$, $P(k + 1)_{N1}$), ($P(k)_{N2}$, $P(k + 1)_{N2}$), ($P(k)_{N3}$, $P(k + 1)_{N3}$) three point pairs of distances are less than the pre-set threshold. In the same way, the neighbouring feature points of the feature points of the $(K + 1)$ frame image and the corresponding Q points of the (K) frame image also need to satisfy their local constraints.

A depth constraint condition of feature points is a feasible feature matching pair, such as ($P(k)$, $P(k + 1)$) and the difference between the absolute values of the

corresponding depth information values ($Depth(P(k))$, $Depth(P(k+1))$ is less than a pre-set depth threshold.

Table 1 Definition of variables

Variable	*Meaning of representation*
$X(k)_i$	The i-feature point in the K image
$D(X(k)_i)$	The feature point descriptor of $X(k)i$
$C(X(k)_i)$	$X(k)i$ describe the most similar feature points of the $K+1$ frame image
$C^{-1}(X(k+1)_i)$	$X(k+1)_i$ describe the most similar feature points of the K frame image
$X(k)_i^q$	The q neighbour feature point in the K frame image and the i-feature point
m	The smaller value of the number of two sets of matching points between adjacent frames

The established model with local neighbour constraint and depth constraint are as follows:

- *Objective function*

$$F(X(k), X(k+1)) = \min_i \sum_{i=1}^{6} \left\| D(X(k))_i - D(X(k+1))_i \right\|_2 \tag{6}$$

- *Constraints*

Constraint 1:

$$\left\| C\left(X(k)_i^q\right) - X(k+1)_i \right\|_2 \le \delta_N, \quad q = 1, 2, 3 \tag{7}$$

Constraint 2:

$$\left\| C^{-1}\left(X(k+1)_i^q\right) - X(k)_i \right\|_2 \le \delta_N, \quad q = 1, 2, 3 \tag{8}$$

Constraint 3:

$$\left\| Depth(X(k+1)) - Depth(X(k)) \right\|_2 \le \delta_D \tag{9}$$

The most similar pairs of feature points to be sought are represented by the objective function. Where $C(X(k)_i)$ represents the descriptor of the i feature point in the k image. $Depth(X(k)_i)$ represents the depth value of the i feature point in the k image, δ_N represents a neighbour constraint threshold, δ_D indicates the depth constraint threshold. The combination of descriptors of feature points, depth values, nearest neighbour constraint queues, and depth constraint queues together form the constraints. In this paper, the three nearest neighbours of the target feature points are used as the most local constraints.

6 Camera pose and three-dimensional point cloud

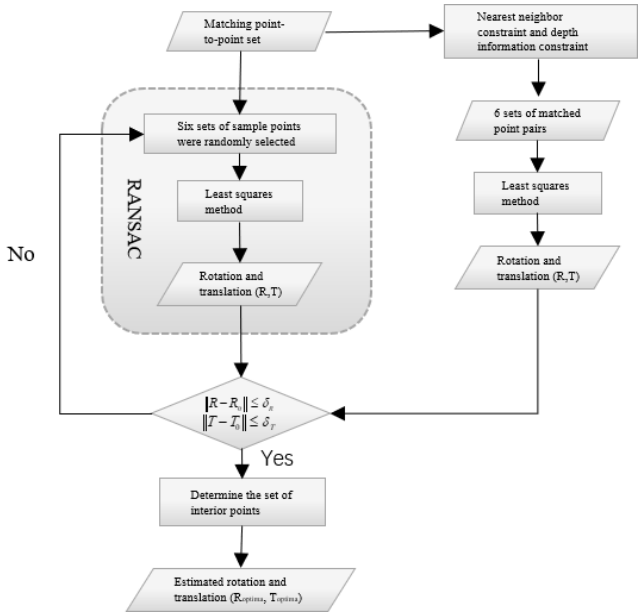
Assuming that the coordinates of point P in the camera coordinate system in the K frame image is $(x, y, z)^T$, the coordinate of point p in the camera coordinate system in the image of $K+1$ frame is $(x', y', z')^T$, and there is a transformation matrix, so that

$$\begin{pmatrix} x' \\ y' \\ z' \\ 1 \end{pmatrix} = \begin{bmatrix} R_K & T_K \\ \vec{0} & 1 \end{bmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{10}$$

R_K represents an orthogonal rotation matrix, T_K represents the translation vector.

The least squares method can be used for the sample set to obtain the camera posture, Elforaici et al. (2018). However, when there are ou1tliers (noise points) in the dataset, the least square method is degraded in parameter estimation. Therefore, RANSAC can be used to remove the outer points before the least square method, and then the least square method can be applied to the inner point set. If that number of points outside in the sample set is large, the RANSAC algorithm performance will be greatly reduced. Therefore, the accurate matching point pairs obtained with the nearest neighbour constraint and depth constraint can be used to assist RANSAC in camera attitude estimation, which can also shorten the parameter estimation time of RANSAC. The flowchart below for the attitude estimation process is given in Figure 4.

Figure 4 Posture estimation flow chart



The final carrier attitude estimation information can be used as the initial value to be introduced into the graph optimisation method to generate the final 3D point cloud，and now the 3D reconstruction of the indoor scene.

7 Experimental results and analysis

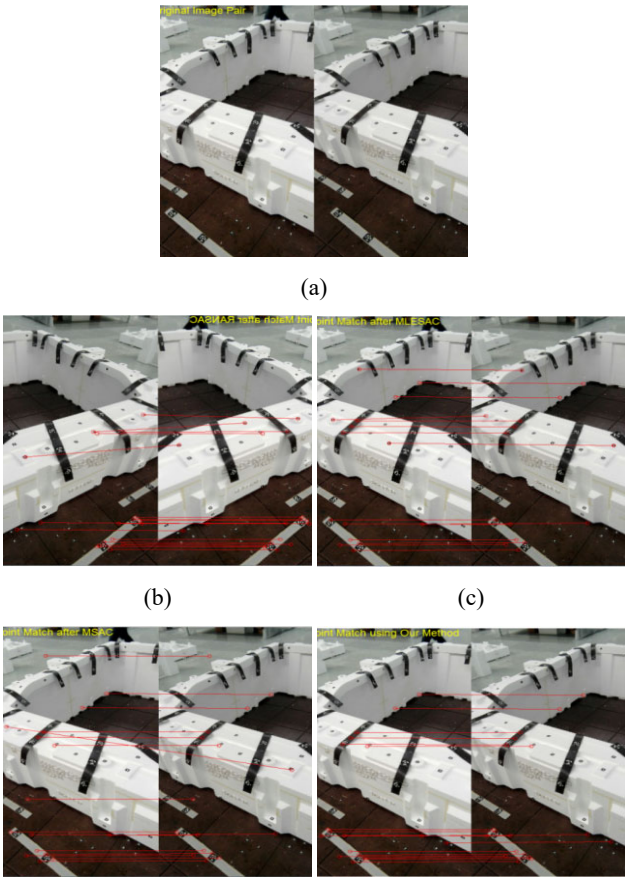
In this paper, two groups of experiments were done, both of which were conducted in the office. In the first group of experiments, the experimenter held Kinect in hand and freely moved indoors to collect the images of indoor scenes.

In the second group of experiments, Kinect was carried on a mobile car to collect RGB images and depth images.

7.1 Matching feature points between adjacent frames

In the step of carrier attitude estimation, after using the local neighbour constraint and depth constraint of image feature points, all the matching point pairs are correct, while the traditional random sampling consistency analysis method is used without removing all the outlier points.

Figure 5 Matching of feature point pairs between adjacent frames (see online version for colours)



(a)

(b) (c)

(d) (e)

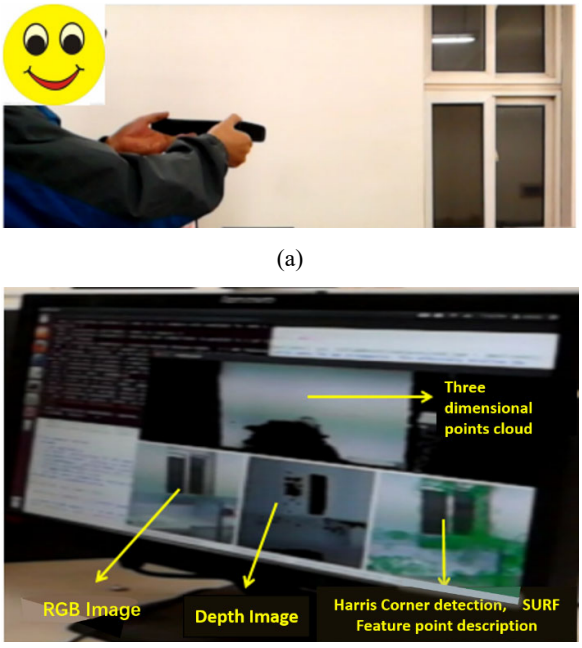
As shown in Figure 5, the left and right colour images in Figure 5(a) are two adjacent frames of images taken with Kinect. Figure 5(b) is the matching result of feature points obtained by RANSAC algorithm, the matching point pairs are marked by red hollow circles and connected by red straight lines. Figure 5(c) is a matching point obtained by using the traditional MLESAC method. Figure 5(d) is a matching point obtained by MSAC method. Figure 5(e) is the matching result of feature point pairs with constraints proposed in this paper. There are still three outer points in the matching point pair obtained by RANSAC method. There are two outer points in the matching point pair obtained by MLESAC method. There are three outer points in the matching point pair obtained by MSAC method. The matching point pairs obtained by the method proposed in this paper have no outer points. In addition, in order to comprehensively analyse the robustness of removing mismatching points, 50 images were taken from different angles. In this paper, 80 sets of images are selected, and the removal of matching points between each set of images is analysed: there are still 121 outliers, 109 outliers and 113 outliers in traditional RANSAC, MSAC and MLESAC methods respectively. With the method of removing outliers in this paper, all mismatching points can be removed.

It can be seen that after introducing local nearest neighbour constraint and depth information constraint, the matching effect of image feature point pairs between adjacent frames is higher in accuracy and robustness.

7.2 Kinect generates 3D point cloud

Three-dimensional point cloud program is implemented under Linux system, using Intel i5 processor and 4G memory. The experimenter freely moves indoors with Kinect, as shown in Figure 6(a). The program can calculate the position, attitude and point cloud information of the camera in real time, and the average depth error of the reconstructed point cloud in X-, Y-directions is (3.75 cm, 2.77 cm).

Figure 6 Hand-held Kinect generates the three-dimensional point cloud in real time (see online version for colours)

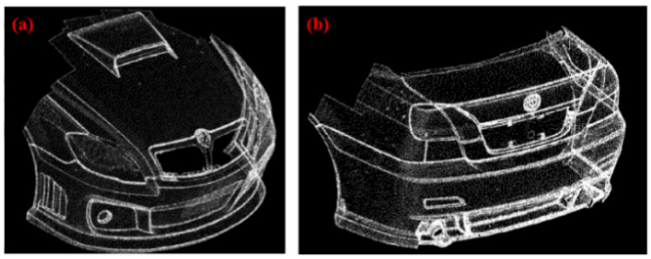


(a)

(b)

The second group of experiments is the 3D point cloud model of Brilliance Zhonghua Automobile reconstructed by this algorithm. As shown in Figure 7, the scene of the front part and the rear part reconstructed by Kinect through layered surround photography. By excluding the blurred images, a complete 3D point cloud process can be obtained by stitching the dense 3D point cloud of each part of the vehicle using Geomagic software, the average depth error of the reconstructed point cloud in the X-, Y-directions is (2.13 cm, 1.85 cm).

Figure 7 3D reconstruction of automobile based on Kinect (see online version for colours)



8 Conclusions

In this paper, the depth image information and colour image information output by RGB-D camera are used to realise real-time 3D reconstruction of indoor scene and attitude estimation of carrier. Firstly, Harris corner detector is used to extract the feature points of RGB images, and then SURF feature point descriptors are used to generate 64-dimensional feature vectors. In the feature point matching between adjacent frames, two kinds of constraints are adopted in this paper: local nearest neighbour constraint of feature points and depth constraint, using these two constraints can obviously increase the matching accuracy of feature point pairs and speed up the matching process of feature point pairs, so that the traditional problems such as inaccurate cloud registration and large computation time are well solved.

In the experiment, Kinect is fixed on a trolley, and real-time indoor scene reconstruction can be realised by using the autonomous mobile trolley. The accuracy and robustness of this method in indoor 3D reconstruction and carrier attitude estimation are verified by these experimental results.

Current research is mostly limited to indoor scenes, in the future, our ultimate goal is to be able to semantically analyse complete 3D scenes from one or more images, which requires joint detection, recognition and reconstruction, and more importantly, capturing and modelling the spatial relationships and interactions between objects and between object parts.

Acknowledgements

This work was supported by Research and Development and application of key technologies of smart agriculture in Inner Mongolia region along the Yellow River under contact 2020GG0033.

References

Altmann, Y., Mclaughlin, S. and Davies, M.E. (2020) 'Fast online 3D reconstruction of dynamic scenes from individual single-photon detection events', *IEEE Transactions on Image Processing*, Vol. 29, No. 5, pp.2666–2675.

Atman, J. and Trommer, G.F. (2018) 'Laser-camera based 3D reconstruction of indoor environments', *IEEE/ION Position, Location and Navigation Symposium*, pp.254–260.

Bhowmik, A., Gumhold, S., Rother, C. and Brachmann, E. (2020) 'Reinforced feature points: optimizing feature detection and description for a high-level task', *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4947–4956.

Cai, J., Yan, F., Wu, Z., Liu, Y. and Liu, Q. (2020) 'Multi-view 3D reconstruction based on Kinect v2', *Journal of Sensing Technology*, Vol. 33, No. 8, pp.1149–1154.

Dulko, F. (2018) '360° color and depth mapping of an indoor room: Microsoft Kinect 2.0', *2018 IEEE Long Island Systems, Applications and Technology Conference*, pp.1–11.

Elforaici, M.EA.., Chaaraoui, I., Bouachir, W., Ouakrim, Y. and Mezghani, N. (2018) 'Posture recognition using an RGB-D camera: exploring 3D body modeling and deep learning approaches', *IEEE Life Sciences Conference*, pp.69–72.

Han, L., Zheng, T., Xu, L. and Fang, L. (2020) 'Occuseg: occupancy-aware 3D instance segmentation', *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2937–2946.

Hu, J., Zheng, W., Lai, J. and Zhang, J. (2017) 'Jointly learning heterogeneous features for RGB-D activity recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 11, pp.2186–2200.

Jacob, S., Menon, V.G. and Joseph, S. (2020) 'Depth information enhancement using block matching and image pyramiding stereo vision enabled RGB-D sensor', *IEEE Sensors Journal*, Vol. 20, No. 10, pp.5406–5414.

Li, C., Guan, T., Yang, M. and Zhang, C. (2021) 'Combining data-and-model-driven 3D modelling for small indoor scenes using RGB-D data', *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 180, pp.1–13.

Luo, J., Qiu, G., Zhang, Y., Feng, S. and Han, C. (2020) 'Research on SURF boundary visual matching algorithm based on adaptive dual threshold', *Journal of Instrumentation*, Vol .41, No. 3, pp.240–247.

Nicastro, A., Clark, R. and Leutenegger, S. (2019) 'X-Section: cross-section prediction for enhanced RGB-D fusion', *IEEE/CVF International Conference on Computer Vision*, pp.1517–1526.

Papadopoulos, G.T. and Daras, P. (2018) 'Human action recognition using 3D reconstruction data', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 8, pp.1807–1823.

Simonsen, D., Spaich, E.G., Hansen, J. and Andersen, O.K. (2017) 'Design and test of a closed-loop FES system for supporting function of the hemiparetic hand based on automatic detection using the Microsoft kinect sensor', *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 25, No. 8, pp.1249–1256.

Wang, S., Zhou. Z., Qu, H. and Li, B. (2017) 'Bayesian saliency detection for RGB-D images', *Acta Automatica Sinica*, Vol. 43, No. 10, pp.1810–1828.

Yang, F., Ding, X. and Cao, J. (2021) '3D reconstruction method of free-form surface based on colored striped structured light', *Acta Optica Sinica*, Vol. 41, No. 2, pp.5–15.

Yu, S., Jiang, Z., Wang, M., Li, Z. and Xu, X. (2021) 'A fast robust template matching method based on feature points', *International Journal of Modelling, Identification and Control*, Vol. 35, No. 4, pp.346–352.

Zhang, Y., Dai, J., Zhang, H. and Yang, L. (2018) 'Depth inpainting algorithm of RGB-D camera combined with colour image', *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference*, pp.1391–1395.

Zheng, T., Huang S., Li Y. and Feng, M. (2020) 'Review of key technologies of 3D reconstruction based on vision', *Acta Automatica Sinica*, Vol. 46, No. 4, pp.631–652.

Zhou, Y., Kuang, H., Mou J. et al. (2021) 'Improved monocular ORB-SLAM for semi-dense 3D reconstruction', *Computer Engineering and Applications*, Vol. 57, No. 8, pp.180–184.

Zhu, Q., Weicun, Z., Zhang, J. and Sun, B. (2019) 'U-neural network-enhanced control of nonlinear dynamic systems', *Neurocomputing*, Vol. 352, No. 4, pp.12–21.