External Joint Calibration of A Novel Multi-Modal Perception System

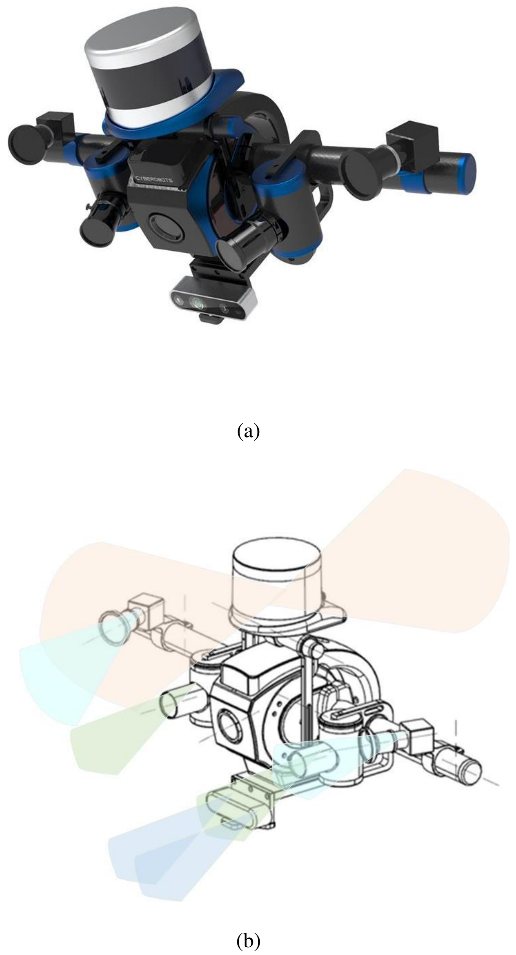
Xiao Han[1,2],Hongpeng Wang[1,2], Chongshan Fan[1,2],Yaojing Li[1,2]

***Abstract*— We design a new multi-modal perception fusion system for the whole scene, which can be divided into two parts. The first part is lidar, rgbamera1, rgbamera2. The second part is RGB camera 3, RGB camera 4, depth camera. The combination of these three modes provides a close range (0.3-10 m) of environmental perception, and provides the texture and distance information of the environment in the cone-shaped blind area of lidar. Our main work is to design the experimental equipment of the proposed multi-modal perceptual fusion system, and to push forward the optimization theory with the re projection error as the objective function and analyze the multi-modal fusion sensing system. The graph structure is established, and the relative pose transformation between various modes is analyzed. In the back-end optimization part, MCPE (Minimally connected pose estimation) and PSE(Pose and structure estimation)are used to analyze and compare the system. In the experimental part, we use matlab to calibrate the binocular camera by minimizing the error of re projection. The lidar, RGB camera and depth camera are calibrated with the help of the multi-modal calibration tool (binding on ROS) proposed . It provides a solid foundation for the future experiments in perception of self-driving, UAV perception, medical field and military field.**

I. INTRODUCTION

With the development and popularization of 5G technology, the research of self-driving technology is hot. The core technology is the joint use of sensors, such as lidar, IMU, color camera, depth camera,[1] the fusion of different multi-source sensor information [2]and real-time detection and recognition of real scene[3]. Because of the cost, we take out the core part and study the key technology. Therefore, the multi-modal perception fusion system proposed in this paper is constructed to study the core technology of unmanned driving.

Different kinds of sensors, i.e. different modes, will provide different types of information. At the same time, different sensors work in different ranges. Therefore, it is very practical and necessary to fuse the information of multimodal sensors. For example, RGB camera and depth camera have high resolution, which can reach pixel level. Depth camera can make up for the depth distance information that RGB camera does not have. However, the depth camera adopts the principle of structured light, and its working range is less than 10 m. Therefore, in order to make up for the distance information in the distance.



(a)

(b)

Fig. 1. a) The general diagram of a new multimodal system with the symbol of cyberrobot placed in the middle b) The sensing area of each sensor mode in the system is rendered by keyshot software.

The major contributions of this study include the following: 1)

1) Design a multi-modal perceptual fusion system according to the requirements
2) Mathematical description and theoretical derivation for multimodal system
3) The multi-modal perceptual fusion system is calibrated with external parameters and tested with ROS open source development tool

This work is supported by National Key R&D Program of China(Grant No.2019YFB1311804), National Natural Science Foundation of China (Grant No.61973173,91848108), Technology Research and Development Program of Tianjin (Grant No. 18ZXZNGX00340)).

[1]College of Artificial Intelligence, Nankai University, Tianjin, 300353

[2]Tianjin Key Laboratory of Intelligent Robotics, Tianjin, 300353

[3]Research Laboratory of Machine Learning and Pervasive Computing, Nankai University, Tianjin, 300353

Email:2120180396@mail.nankai.edu.cn,hpwang@nankai.edu.cn, songyulin@mail.nankai.edu.cn,xinyw@nankai.edu.cn

Hongpeng Wang is the corresponding author

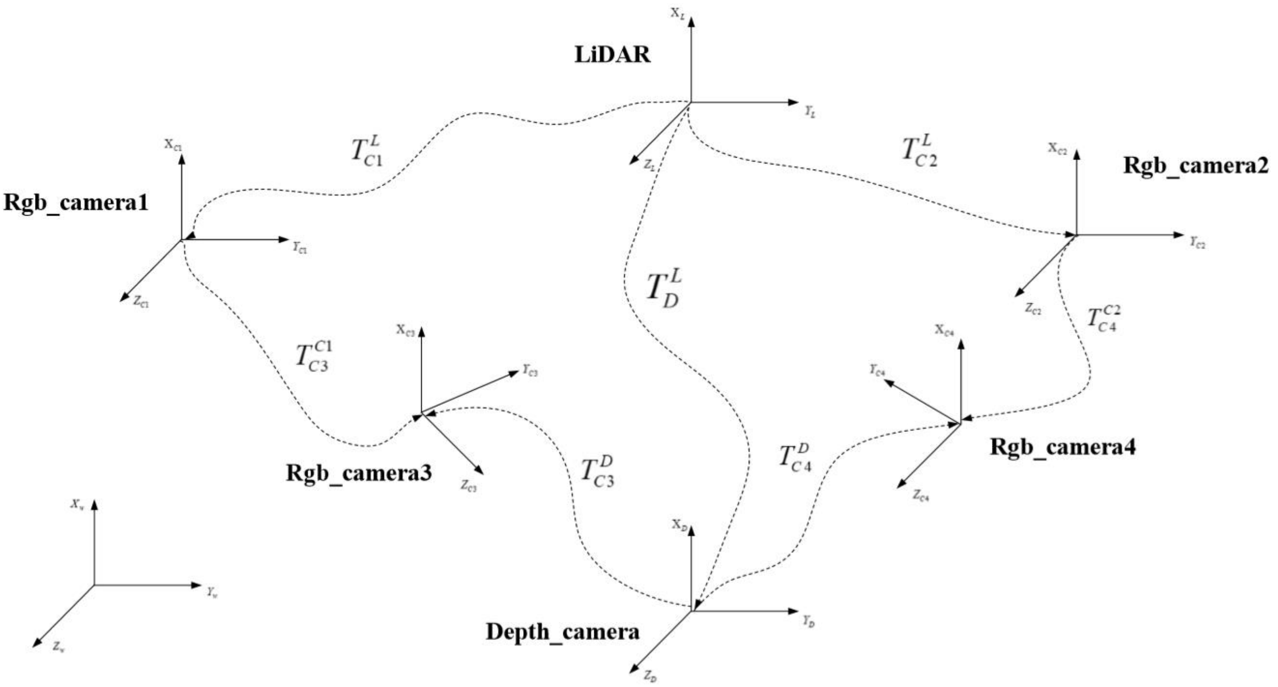


Fig. 2. The diagram structure of multimodal system is shown in. Among them, the vertex represents lidar, camera, IMU and other sensors. Edge represents the relative pose transformation between sensors

The rest of the paper is organized as follows: Section: Related Work comprehensively reviews the multi-model system and Related image processing and optimization algorithm method. The system design and explain are presented in Section:System-Overview, followed by Introduction to multimodal system: components of each mode and overall working principle. Section: Experiment shows the experimental result and compare it to the recent state of art methods. Finally we look forward to further development in Section: Conclusion

II. Related Work

In this section, we briefly introduce the related work in the field of perception and fusion of all terrain unmanned vehicle and multimodal sensor. With the rapid development of driverless driving in recent years, the progress of related directions has been driven. For the field of all terrain unmanned vehicle(UTV).Stanley from the system composition of the car to the analysis of the sensing area, the development of driverless driving is promoted, and the theory to practice is completed.[6]In view of the task of driverless driving in unstructured road, a fault-tolerant perception model of multi vision sensor is proposed, and the system architecture is adjusted to describe the working principle of the whole system.For the field of multimode sensor system,[3] [4]Fusion experiment of lidar and color camera in unstructured environment (Airport).[7]Using outdoor environment (such as tree) to calibrate lidar camera system.[10]Taking a car model as the carrier, the Kalman filtering algorithm is derived. A solid theoretical proof of perceptual fusion.[11] fused lidar and hyperspectral data for urban land-use classification,Using machine learning methods.[12]The Kinect sensor is used, and the internal analysis of the sensor is carried out, which provides a reference for our hardware selection. [13]A four point calibration method is used to calibrate the sensing system.[14]proposes a targetless and automatic camera-LiDAR calibration method. Also extends the hand-eye calibration framework to 2D-3D calibration. [5] and [7]proposed the methods of external parameter calibration [9]use deep neural networks method to register multi model system. [15] propose an end-to-end, automatic, online camera-LIDAR calibration approach.[16]An off-line camera fusion method is proposed to build a dense and accurate 3-D model. And the probability model is constructed to analyze the residual of re projection.Baidu is a network company pushing driverless technology. [17]Pointnet is proposed and a new type of neural network which consumes point cloud directly is designed. It respects the invariance of input points. After the completion of multimodal system design and modeling, we also need to study the data fusion[18] and segmentation, which is the focus of our next work

III. System Design And Problem Formulation

A. System design and procedure

This part briefly describes the causes of multi-modal system, as well as the introduction of each component, and what technology to use, and what fields will be used in the future. First of all, in recent years, the unmanned driving technology is developing rapidly, and the multi-modal sensor information fusion perception is one of the key technologies. So that's why we chose this direction. Secondly, the system we designed includes six sensors. Among them, laser radar

1 adopts radium intelligent c16-151b, four RGB cameras are fixed on the pole, the bottom is a depth camera, and Intel real sense d435i is used. The system can be roughly divided into two subsystems, LiDAR and RGB camera on the outside constitute the subsystem of remote scene perception, and depth camera and RGB camera on the inside constitute the subsystem of close range perception. By fusing texture information and distance information, we can get as much environment information as possible. At the same time, the multimodal system can also be used in UAV, medical, military and other fields.

In order to make the information of lidar color camera, IMU, depth camera and other multimodal sensors fusion perfectly, we need to calibrate the multimodal sensing fusion system designed in this paper and estimate the structure of unknown objects. Among them, calibration includes mechanism calibration and sensor calibration, while sensor calibration requires internal calibration and external calibration based on internal calibration. The calibration procedure contains three steps: 1)

1) Extract the topology structure of the system itself, and analyze the rigid transformation and kinematics
2) Description and mathematical modeling of the use scenario of the system
3) The multimodal system is used for the optimal analysis of the unknown scene's drawing and structure estimation.

B. Problem Formulation

1)

1) {W}denotes the Cartesian world coordinate system.

$$^{W}X_{\mathrm{i}} \in R^{3}$$

represents the i-th calibration point (located at the center of the metal ball i) with respect to W. As a convention, we will use the left superscript indicates the reference coordinate system in this paper.

2) Similarly, l D represents the Cartesian coordinate system of lidar and depth camera respectively. C1 C2 C3 C4 denotes the Cartesian coordinate system of four color cameras respectively. 1,2 are external cameras, 3,4 are internal cameras. Where, the X axis is vertical upward and the Z axis is outward. The y-axis is obtained from the known x-axis and y-axis according to the left-hand rule.
3) The rigid body transformation of j sensor relative to i sensor is Represents as following: T_{j}^{i} where

$$i, j \in \{L, C1, C2, C3, C4, D\}$$

we asume that: 1)

1) The working environment of the whole system has no large-scale change of light exposure, and there is no obstruction
2) After one calibration, there is no relative pose transformation between multimodal sensors, which is very important for back-end data fusion and optimization.
3) The noise of the system obeys zero mean Gaussian white noise

IV. Methodology

A. joint calibration topological

According to the multimode hardware system of Figure 1, we design the topology structure as shown in Figure 2. The relationship between relative position and relative pose of sensors in the whole multi-modal sensing system is described. The vertex represents each mode in the system, and the edge represents the rigid body transformation relationship between adjacent sensor modes.

Next, the theoretical analysis and formula derivation will be carried out from four parts: rigid transformation, system description and optimization (including Minimally Connected Post Estimation and Position And Structure Estimation).

B. Rigid Transformation

After the completion of hardware design of multimodal system, the system needs to be calibrated. Among them, the calibration is divided into mechanism calibration, internal parameter calibration and external parameter calibration. First, the mechanism calibration?Mechanism calibration is mainly to determine the kinematic model of multi-modal perception fusion system. The kinematic model is based on the transformation of 6-DOF rigid body in Cartesian coordinate system.

Due to the limitation of the length of the paper, we only use the rigid transformation of lidar relative to the reference coordinate system (including rotation transformation and translation transformation). The rigid transformation of other modes is the same too.

$$\begin{bmatrix} \mathrm{X}_{L} \\ Y_{L} \\ Z_{L} \\ 1 \end{bmatrix} = \begin{bmatrix} R_{3\times 3} & T_{3\times 1} \\ 0 & 1 \end{bmatrix} \bullet \begin{bmatrix} \mathrm{X}_{w} \\ Y_{w} \\ Z_{w} \\ 1 \end{bmatrix} \tag{1}$$

$$^{L}_{W}R_{3\times 3} = \begin{bmatrix} C_{\alpha}C_{\beta}C_{\gamma} - S_{\alpha}S_{\beta} & -C_{\alpha}C_{\beta}S_{\alpha} - S_{\alpha}C_{\gamma} & C_{\alpha}S_{\beta} \\ S_{\alpha}C_{\beta}C_{\gamma} + C_{\alpha}S_{\beta} & -S_{\alpha}C_{\beta}S_{\gamma} + C_{\alpha}C_{\gamma} & S_{\alpha}S_{\beta} \\ S_{\beta}C_{\gamma} & S_{\beta}S_{\gamma} & C_{\beta} \end{bmatrix} \tag{2}$$

$$^{L}P_{Worg} = \begin{bmatrix} ^{W}X_{L} \\ ^{W}Y_{L} \\ ^{W}Z_{L} \end{bmatrix} \tag{3}$$

Formula (1) describes the homogeneous linear transformation from lidar coordinate system to world coordinate system. The rotation matrix $\mathrm{R}_{3\times 3}$ and translation matrix $T_{3\times 1}$ are represented by formula (2) and formula (3) respectively where:α, β, γ are the angles of rotation around the Y axis, Z axis and X axis of the world coordinate system. Right hand rule for coordinate system

$$|\psi_{i}| \leq \Psi_{\max}, \forall i \tag{4}$$

In intrinsic parameter calibration, we use Zhang's method. Use checkerboard to detect corners. Then the internal parameters of the camera are calculated according to the distortion coefficient. The tilt angle of the chessboard calibration board we used meets the above conditions

C. System Description

After describing the kinematic model and intrinsic parameter calibration of the multimodal system, the system description is carried out. In the field of vision slam, the feature points in unknown environment need to be detected according to sensors. Then, according to the feature point matching, the motion track of the sensor is calculated, and the acquired multiple images are registered. Lead to the purpose of 3D reconstruction and mapping.

$$\begin{bmatrix} S_L \\ S_{c1} \\ S_{c2} \\ S_{c3} \\ S_{c4} \\ S_D \end{bmatrix} = \begin{bmatrix} S_L^1 & S_L^2 & ... & S_L^{n-1} & S_L^n \\ S_{c1}^1 & S_{c1}^2 & ... & S_{c1}^{n-1} & S_{c1}^n \\ S_{c2}^1 & S_{c2}^2 & ... & S_{c2}^{n-1} & S_{c2}^n \\ S_{c3}^1 & S_{c1}^2 & ... & S_{c1}^{n-1} & S_{c1}^n \\ S_{c4}^1 & S_{c4}^2 & ... & S_{c4}^{n-1} & S_{c4}^n \\ S_D^1 & S_D^2 & ... & S_D^{n-1} & S_D^n \end{bmatrix} \tag{5}$$

Equation (5) describes n feature points in unknown environment detected by each sensor mode in multimodal system.

$$\theta^{1,2} = (t_x, t_y, t_z, \upsilon_x \cdot \alpha, \upsilon_y \cdot \alpha, \upsilon_z \cdot \alpha) \tag{6}$$

In order to facilitate the calculation, for each feature point in the location environment, we use the homogeneous representation to represent the 3D point as $(x,y,z,1)^T$.We use $\theta^{1,2}$ parameterization to represent 6-DOF parameters homogeneous representation

$$\varepsilon_i(\theta^{a,b}) = \sum_{p=1}^{4} \left\| y_{y(p)}^{b} - T^{a,b} \cdot y_{y(p)}^{a} \right\|^2 \tag{7}$$

In external parameter calibration, we use four point calibration method. The error function is constructed by summing the error of four center re projection. That is, for the parameter estimation of the ith feature point, sum the difference between the actual observation value and the estimated value of the four centers

$$f(\theta^{1,2}) = \sum_{k=1}^{n} \mu_k^2 \cdot \mu_k^1 \cdot \varepsilon_k(\theta^{1,2}) \tag{8}$$

Taking the joint calibration of two sensor modes as an example, multiple modes can be analogized. Finally, it is considered that not all sensor modes can be detected for the same landmark in the process of moving the multimodal system. So we added two parameters μ_k^1 and μ_k^2.These two parameters can only be taken as 0 or 1. Taking 0 means that the ith sensor mode is not detected to kth landmark; on the contrary, taking 1 means detected.

The above theoretical derivation is to analyze the observation of characteristic points of two sensor modes. Because our system is multimodal. So, on this basis, we extend to s modes (s ¿ 2). Here, our system contains six sensors, that is s = 6. In SLAM back-end optimization, the most commonly used ones are minimal connected post estimation and position and structure estimation.

D. Minimally connected pose estimation

As shown in the first figure in Fig. 3, the working principle of minimally connected post estimation is described, which can be understood as accumulation based on formula (8). Get formula (11)

$$f(\theta) = \sum_{s=1}^{S} \sum_{i=1}^{N} \mu_i^k \bullet \mu_i^1 \bullet \varepsilon(\theta^{1,i}) \tag{9}$$

E. Pose and structure estimation

In the minimal connected post estimation, the basic idea is to optimize the external parameters among the sensor modes according to minimizing the cumulative error.The following is the theoretical derivation of position and structure estimation,As shown in the second figure in Fig.3.

The most commonly used method is the position and structure estimation method, which adopts the framework of probability and statistics theory, and uses Mahalanobis distance in the error function. With the rapid development of machine learning represented by deep learning, the development of probability theory and mathematical statistics is driven. Therefore, this method is very popular, and many theories adopt the advanced technology of deep learning and neural network, which makes this direction very attractive.

$$M = (m_1, m_2, m_3, ..., m_k) \tag{10}$$

Equation (12) indicates the set of all landmarks in the location environment.

$$y_{k(p)}^{i} = T^{M,i} y_{k(p)}^{M} + \eta^i \tag{11}$$

(13) Formula describes the acquisition process of observation samples, where η^i follows the Gaussian distribution with mean value of 0 and variance of Σ^i

$$D_{\Sigma}^2(a,b) = [a-b]^T (\Sigma)^{-1} [a-b] \tag{12}$$

The expression uses Mahalanobis distance as covariance, which can be seen as adding weight to Euclidean distance.

$$\varepsilon(\theta^{M,i}, M) = \sum_{p=1}^{2} \mathrm{D}_{\Sigma}^2(y_{k(p)}^{i}; T^{M,i} y_{k(p)}^{M}) \tag{13}$$

$$f(\theta) = \sum_{s=2}^{S} \sum_{i=1}^{N} \mu_i^s \bullet \mu_i^1 \bullet \varepsilon(\theta^{1,i}) \tag{14}$$

Similar to MCPE, the error function is defined by taking two sensor modes as examples, and then $T^{M,i}$ is estimated by iterative optimization of the error function.Then we extend it to s sensors (we take n = 6 here). Get the objective function $f(\theta)$.

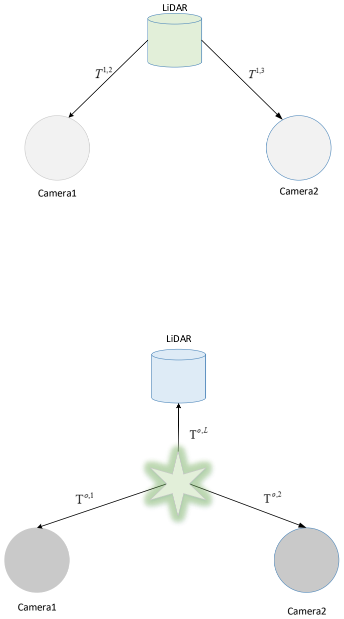


Fig. 3. MCPE and PSE

V. Experiment

A. system consistent design

For our experiments, we use a sensor setup that is mounted on a Tripod.There are also portable handles at the back of the system, which can be held by hand. The multi-modal perception fusion system can be divided into two sub-systems. Three modes on the outside: Lidar and two color cameras constitute the remote sensing subsystem. Two inner color cameras and depth cameras form a close sensing fusion subsystem. The two subsystems cooperate with each other to form the whole scene perception fusion system. Among them, lidar uses c16-151b. The depth camera adopts Intel real sense d435i, which contains IMU. The camera includes two FLIR industrial network interface cameras and two USB3.0 cameras.

B. SIMULATION AND EXPERIMENT RESULTS

Our experimental environment is 16GB ram, Intel (R) core (TM) i7-8700k CPU, gtx1070. Ubuntu LTS 16.04 is selected as the test environment, and C + + is selected as the programming language. Python. Use ROS as middleware. The intrinsic parameters and external parameters of monocular camera and multicular camera are tested by Matlab toolbox. The parameters obtained are shown in Table 1. The open source external calibration tool for lidar, camera and radar, and proposed three con ”guidelines to estimate the sensor positions from simple detections of multiple calibration board locations developed by Joris domhof1, Julian F. P. Kooij and dariu M. gavrila are used for registration between lidar and depth camera.Calibration of lidar module and camera module using four point calibration method.

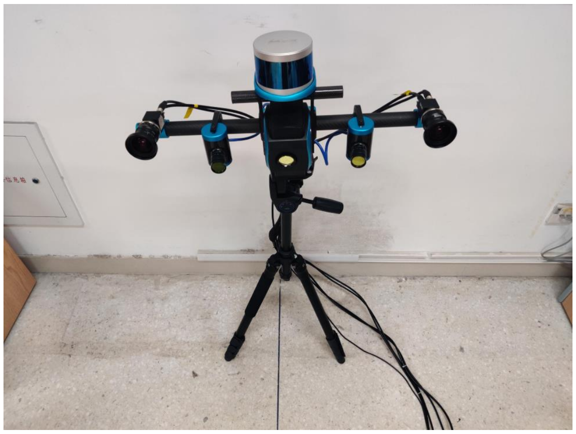


Fig. 4. .The experimental equipment of multi-modal perceptual fusion system designed in this paper can be held or fixed on a tripod

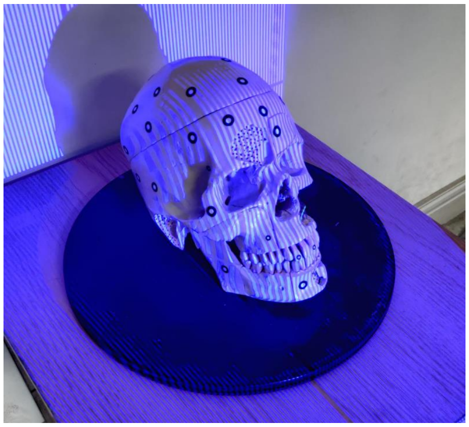


Fig. 5. The object to be scanned is shown as a plastic skull sample

VI. Conclusion

In order to study the detection technology of driverless perception, we extract the core sensor part. And a multi-modal perceptual fusion system is designed and assembled. Then, the mathematical description, theoretical analysis and mathematical derivation of the system are carried out. The optimization techniques of MCPE and PSE are used for comparative analysis. At present, the system design and theoretical derivation have been completed. In the experiment part, we use the opensource external parameter calibration tool proposed in the paper in 2019, and use the four point calibration principle to calibrate the multimodal system. The experimental results show that the ideal effect can be obtained. The multi-modal system can be calibrated to

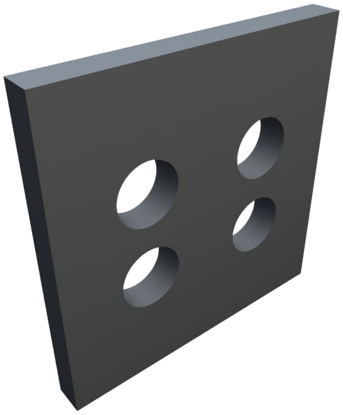


Fig. 6. Checkerboard calibration board

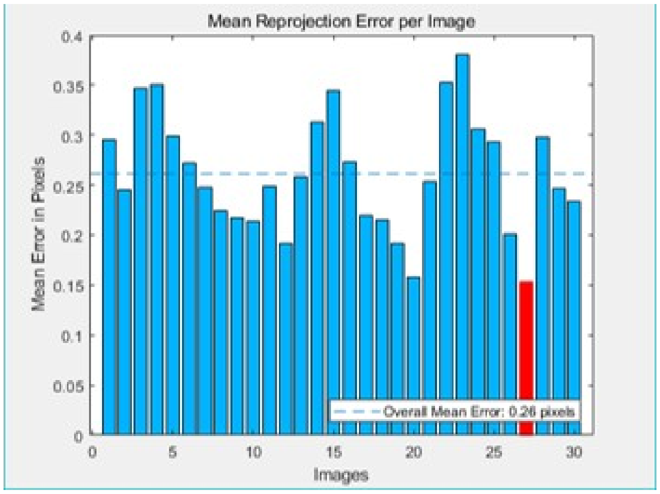


Fig. 7. Calculate the re projection error between the observed value and the estimated value, and optimize it. The optimal parameter estimation is obtained. The method used in the experiment can achieve an error of 0.26 pixels

meet the requirements, so that the multi-modal data can be registered, and provide accurate registration information for the next step of data fusion test. Our next task is to use the system to perform multimodal data fusion in real scenes to highlight the convenience of the system and the advantages of full scene environment awareness

References

[1] Chia-Yen Chen, Hsiang-Jen Chien. On-Site Sensor Recalibration of a Spinning Multi-Beam LiDAR System Using Automatically-Detected Planar Targets[J]. Sensors, 2012, 12(10):13736-13752.

[2] Li, Juan , X. He , and J. Li . "2D LiDAR and Camera Fusion in 3D Modeling of Indoor Environment." 2015 National Aerospace and Electronics Conference (NAECON) IEEE, 2015.

[3] Z. Chai, Y. Sun and Z. Xiong, "A Novel Method for LiDAR Camera Calibration by Plane Fitting," 2018 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Auckland, 2018, pp. 286-291.

[4] B. Yohannan and D. A. Chandy, "A novel approach for fusing LIDAR and visual camera images in unstructured environment," 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, 2017, pp. 1-5.

[5] F. M. Mirzaei, D. G. Kottas, and S. I. Roumeliotis, ?3D LIDAR?camera intrinsic and extrinsic calibration: Identi?ability and analytical least-squares-based initialization,? The International Journal of Robotics Research, vol. 31, no. 4, pp. 452?467, 2012.

[6] C. Guindel, J. Beltran, D. Mart?n, and F. Garc?a, ?Automatic extrinsic calibration for lidar-stereo vehicle sensor setups,? in 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2017, pp. 1?6.

[7] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, ?Automatic extrinsic calibration of vision and lidar by maximizing mutual information,? Journal of Field Robotics, vol. 32, no. 5, pp. 696?722, 2015.

[8] N. Schneider, F. Piewak, C. Stiller, and U. Franke, ?RegNet: Multimodal sensor registration using deep neural networks,? in 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2017, pp. 1803?1810.

[9] Willis A R , Zapata M J , Conrad J M . A linear method for calibrating LIDAR-and-camera systems[C]// 17th Annual Meeting of the IEEE/ACM International Symposium on Modelling, Analysis and Simulation of Computer and Telecommunication Systems, MASCOTS 2009, September 21-23, 2009, South Kensington Campus, Imperial College London. ACM, 2009.

[10] Kasper R , Schmidt S . Sensor-data-fusion for an autonomous vehicle using a Kalman-filter[C]// Intelligent Systems and Informatics, 2008. SISY 2008. 6th International Symposium on. IEEE, 2008.

[11] Man Q , Dong P , Guo H . Pixel- and feature-level fusion of hyperspectral and lidar data for urban land-use classification[J]. International Journal of Remote Sensing, 2015, 36(6):1618-1644.

[12] Pter Fankhauser, et al. "Kinect v2 for Mobile Robot Navigation: Evaluation and Modeling." International Conference on Advanced Robotics (ICAR) IEEE, 2015.

[13] Guindel C , Beltrn, Jorge, Martn, David, et al. Automatic Extrinsic Calibration for Lidar-Stereo Vehicle Sensor Setups[J]. 2017.

[14] Ishikawa R , Oishi T , Ikeuchi K . LiDAR and Camera Calibration Using Motions Estimated by Sensor Fusion Odometry[C]// 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

[15] B. Nagy, L. Kovcs and C. Benedek, "Online Targetless End-to-End Camera-LIDAR Self-calibration," 2019 16th International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 2019, pp. 1-6.

[16] W. Zhen, Y. Hu, J. Liu and S. Scherer, "A Joint Optimization Approach of LiDAR-Camera Fusion for Accurate Dense 3-D Reconstructions," in IEEE Robotics and Automation Letters, vol. 4, no. 4, pp. 3585-3592, Oct. 2019.

[17] Qi C R , Su H , Mo K , et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[J]. 2016.

[18] J. Zhang, M. Kaess, and S. Singh, A real-time method for depth enhanced visual odometry,? Autonomous Robots, vol. 41, no. 1, pp. 31?43, 2017.