Software Package for Evaluation the Stereo Camera Calibration for 3D Reconstruction in Robotics Grasping System

Alona Vitiuk [1], Anatoliy Doroshenko [1,2]

[1] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prosp. Peremohy, Kyiv, 03056, Ukraine*
[2] *Institute of Software Systems of the National Academy of Sciences of Ukraine, Akademika Glushkova Avenue, 40, Kyiv, Ukraine*

Abstract

The approach for accuracy assessment of the object model a for the problem of stable grasping in the combined system of the proposal of grasping and the reconstruction of the three-dimensional model of the object was considered. A lot of studies indicate that robot learning from demonstration is a promising way to improve grasping performance, but complete automation of the grasping task in unforeseen circumstances remains difficult and it`s accuracy can be affected at each stage of grasp planning task. Combined system allows stable capture of objects of any shape without restrictions on the types of shapes in the training data set. Novel approaches to surface reconstruction of the object are based on restoring the depth of points from a pair of images from two cameras. The quality of the 3D reconstruction is affected by several factors: the movement of the camera and environmental objects, spatial quantization of the image coordinates, correspondence of key points, camera calibration parameters, unaccounted camera distortions, as well as numerical and statistical properties of the selected reconstruction method. Camera parameter errors can be minimized by improving the calibration procedure, so the impact of errors on the quality of the 3D model was investigated. The deviation of the model from the plane is chosen as a metric for quality assessment. For its calculation, the point cloud is processed by plane identification and segmentation. The software package for accuracy estimation was developed. An experiment was conducted to obtain the dependence of the accuracy of the reconstructed planes on the errors of the camera parameters. The impact of calibration errors on 3D reconstruction was evaluated by comparing metrics for individual planes at different levels of artificial error and evaluating the impact of the error on these metrics. Modeling the error of the camera calibration parameters with a given noise level shows that the calibration parameters deteriorate as the noise level increases. In particular, it was established that an increase in error contributes to an increase in the error of estimation of calibration parameters. In addition, orientation parameters (rotation and translation) are more complex and therefore more sensitive to measurement noise than other parameters.

Keywords

three-dimensional reconstruction, camera calibration, stable grip, point cloud, manipulator robot, mobile robot.

1. Introduction

Planning of stable grasping and efficient movement of objects in an unstructured environment is an urgent problem in robotics. In man-made conditions, the process of three-dimensional reconstruction based on multiple images can be used to help robots determine their position in space, build a three-dimensional map of the environment, and recognize surrounding objects. In addition, robust grasp planning can be applied to build a complex system that includes gesture recognition and

13th International Scientific and Practical Conference from Programming UkrPROGP'2022, October 11-12, 2022, Kyiv, Ukraine
EMAIL: alyonavityuk@gmail.com (A. 1); doroshenkoanatoliy2@mail.com (A. 2)
ORCID: 0000-0002-1445-9598 (A. 1); 0000-0002-8435-1451 (A. 2)

prediction [1]. The system for capturing an unknown object by a mobile robot consists of a camera, a robotic arm, and a gripping limb. Based on the video stream from the camera, a map of the environment and reconstruction of the model of the object is constructed. In the process of scanning, the system receives important information about the object's structure, shape, size and orientation in space. Such processing and model building should run in real time.

Recently data-driven approaches that perform grasp planning directly on the basis of sensor data (without intermediate state) have led to significant progress in the implementation of grasps and generalization of their movement planning approaches. Existing methods use convolutional neural network architectures and expect an input RGB-D image [2]. They can be applied in a high-reliability robotic system, but require large volumes of grasping data and 3D object models. The size of this data directly affects the percentage of successful captures. The Contact-GraspNet system [3] uses this approach to efficiently generate the distribution of parallel grasps for 6 degrees of freedom directly from scene depth data. Other systems take a perceptually driven approach and often use a representation of the problem in pixel space, such as learning pixel accessibility maps [4] or constraining capture to the normal of the image plane [5].

Pixel-space representations have obvious computational advantages, but there are also obvious physical advantages to generating full 6-degree-of-freedom grips when interacting with real objects. In addition, in some cases it is useful to have information about the presence of a grasping point that is not directly visible in the observed image, primarily because it represents the best opportunity for grasping (e.g., a cup handle) or because of a grasping constraint (e.g., grasping by only visible points can make it difficult to place the object in the required configuration after grasp). Many existing data-driven methods generate a grasp by selecting a visible pixel as the attachment point of the working limb, limiting the grasping planning to visible points on the object.

This shortcoming can be eliminated using an integrated approach to capture planning with the use of grasping offer subsystems and reconstruction of an object of known shape. However, existing systems using machine learning techniques for two modules (a trained grasper proposal network and a trained shape object reconstruction network [5]) have limitations in grasping objects of unknown shapes that were not used during the training of the 3D reconstruction network. The use of analytical approaches to 3D reconstruction combined with learning approaches for grasp suggestion will allow planning the grasp of objects of any shape to increase the accuracy of the manipulator. A sequence of frames (video) is fed to the input of such a reconstruction system, with the help of which it is possible to obtain an image of the object from different angles and calculate a three-dimensional model of the object of sufficient density.

A monocular camera provides information from sensors in the form of two-dimensional images. Therefore, the depth of each pixel is estimated from the ratio of real-world point coordinates between images from different camera positions. Such correspondences are detected by comparing photometric patterns on neighboring pixels of each individual pixel. When using such an approach, inaccuracies arise: pixels on low-texture areas cannot be accurately mapped on images, and accurate 3D reconstruction is usually limited to areas of images with large gradients. Photometric calibration can significantly improve the performance of direct visual odometry methods to improve reconstruction quality.

This article demonstrates how stereo camera calibration errors affect the overall quality of building a three-dimensional model. After all, both internal and external parameters of the camera can introduce an additional error during the reprojection of pixels. In addition, it must be taken into account that pixels corresponding to the same 3D point may have different intensities in different images due to camera optical vignetting, automatic gain and automatic exposure settings. Existing photometric calibration approaches are reviewed in order to restore image intensity values and establish pixel-to-pixel correspondence for stereo images. An analysis of the reconstruction obtained using a mobile robot positioning algorithm and an analysis of the impact of photometric calibration errors on the quality of direct visual odometry methods as part of a three-dimensional reconstruction system was carried out.

2. Grasp planning system

The ability to autonomously grasp unknown objects can greatly assist robots in performing a wide range of tasks. However, a robot in an unstructured environment may encounter objects about which it has only limited experience or knowledge. In such cases, successful grasp requires complex perception, planning and control. However, because each of these problems is complex, fully autonomous capture of unpredicted objects in an unstructured environment remains an unsolved problem.

This section is devoted to the consideration of the model of the combined system of the offer of grasping and reconstruction of a three-dimensional model of an object of unknown shape for the implementation of stable grasping and subsequent manipulation of the object in the environment.

The system consists of two modules (Figure 1: Combined capture planning system), the results of which are combined by a refinement module for planning capture. The input data is a stereo image. The segmented image from the first camera and the depth map are fed to the capture proposal network for processing, and the stereo pair with camera parameters are fed to the 3D reconstruction module.

For the first module, an implementation of the GPNet capture proposal network was taken, which outputs the capture position relative to the camera frame $T_{CG} \in SE(3)$. The 3D reconstruction subsystem outputs a reconstructed point cloud of the object's surface, providing information about the object's shape and visible parts. The outputs of the two subsystems are combined by projecting the capture proposal T_{CG} onto the nearest point of the reconstructed point cloud, yielding the improved grasp proposal T_{CG}^{+}. Since the position of the camera relative to the robot T_{RG} is known, the grasp in the camera coordinate system can be translated into the robot coordinate system for execution by the manipulator: $T_{RG}^{+} = T_{RC} T_{CG}^{+}$.

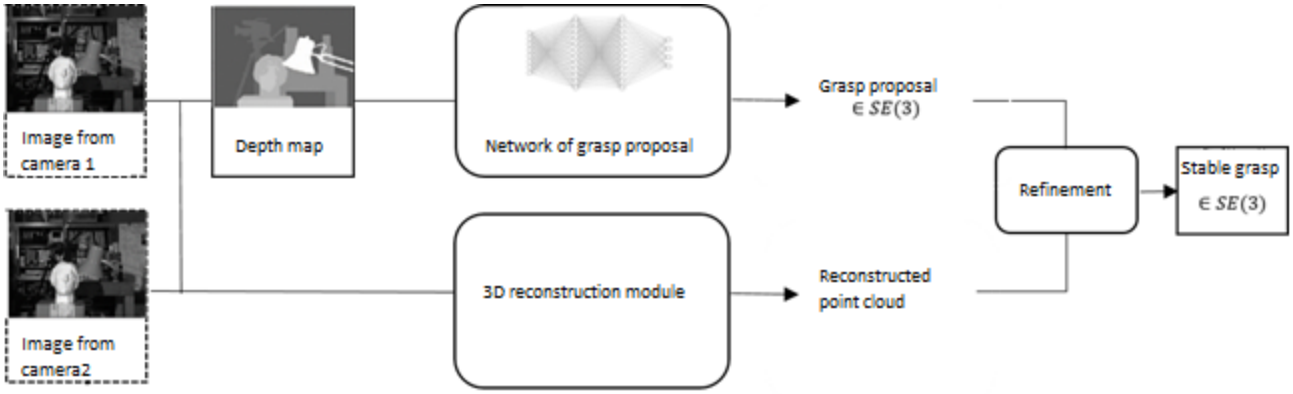

Figure 1: Combined capture planning system

The image acquisition module represents the connection between the main system and the camera. The service sends color images in a format supported by the Robot Operating System (ROS). The image received from the camera is compressed for low bandwidth transmission. In addition, this module also renders the received images.

To rectify the original image, a three-dimensional calibration procedure is performed. It is a calculation of the external and internal parameters of the camera. To find the projection of a three-dimensional point on the image plane, firstly needed to convert the point from the world coordinate system to the camera coordinate system using external parameters (rotation *R* and translation *T*). Next, using the camera's internal parameters, we project a point onto the image plane [6].

Let P be a three-dimensional point with coordinates $\bar{X}$ in the world coordinate system. Coordinate vector of point P in the camera system: $\overline{X_C} = R\bar{X} + \bar{T}$. Here, R is the rotation matrix corresponding to the rotation vector om: R = rodrigues(om). Then we have the coordinates $\overline{X_C}$: $(\overline{X_{C_1}}, \overline{X_{C_2}}, \overline{X_{C_3}})$ = (x, y, z).

The coordinates of the pinhole projection of the point P(a,b) can be represented as:

$$a = \frac{x}{z}$$

$$b = \frac{y}{z}$$

$$r^2 = a^2 + b^2.$$

Let's consider $\theta = \text{atan}(r)$. Then the distortion model [14]:

$$\theta_d = \theta(1 + k_1\theta^2 + k_2\theta^4 + k_3\theta^6 + k_4\theta^8)$$

The coordinates of the distorted point $P`$ represent a vector $\bar{X}`(x`, y`)$:

$$x` = (\frac{\theta_d}{r})a$$

$$y` = (\frac{\theta_d}{r})b$$

The projection of point P on the image is a point $P_{Im}(u, v)$:

$$u = f_x(x` + \alpha y`) + c_x$$

$$v = f_y y` + c_y$$

The purpose of the calibration process is to find the matrix K_{3x3}, the rotation matrix R_{3x3} and the translation vector t_{3x1} using a set of known three-dimensional points (X_w, Y_w, Z_w) and their corresponding image coordinates (u, v). When the values of the internal and external parameters are obtained, the camera is considered calibrated.

There are three types of camera calibration methods: template-based, geometric-keyed, and deep-learning-based. The first approach is to obtain a set of images of a template object with known dimensions and configuration. As the data set is collected, the camera moves and rotates around the pattern to capture images from different viewpoints. This method is best suited for laboratory conditions where it is possible to use a template object manufactured with high geometric accuracy. Various schemes can be used as a template object pattern: checkerboard, circles and more complex ArUco-type markers. The checkerboard patterns are clear and easy to recognize in the image. The corners of the squares on the chessboard are ideal for their localization, as they have sharp gradients in two directions. In addition, these corners are at the intersection of chess lines, and therefore form a repeating structure. All these facts are used to reliably arrange the corners of the squares in a checkerboard pattern.

Calibration by geometric keys is possible when there are geometric clues on the scene under investigation: straight lines, planes or vanishing points of the horizon. Deep learning-based methods are appropriate when sufficient control over the image collection process cannot be exercised (for example, when a single image of a scene is available). The accuracy of methods based on deep learning is much lower.

To obtain the parameters of the camera, it is necessary to collect a set of images of the calibration template in different positions of the camera. Using methods from the OpenCV library, internal camera parameters are obtained and applied to each input image to remove lens distortions. The calibration algorithm is presented in Figure 2.

The world coordinate system is defined by a template object with the image of a chessboard, which is reliably fixed in scene. The three-dimensional object points are the corners of the squares on the chessboard. Any corner of the board can be chosen as the center of the world coordinate system. The axes X_w and Y_w are located along the mounting plane, and the axis Z_w is perpendicular to the plane. Therefore, all points on the chessboard are on the XY plane (ie $Z_w = 0$ for a flat template).

Thus, the camera calibration algorithm has the following inputs and outputs:

- Input data: a set of images with key points where two-dimensional coordinates in the image system and three-dimensional coordinates in the world system are known.
- Output data: camera matrix with internal parameters, rotation and translation of each image.

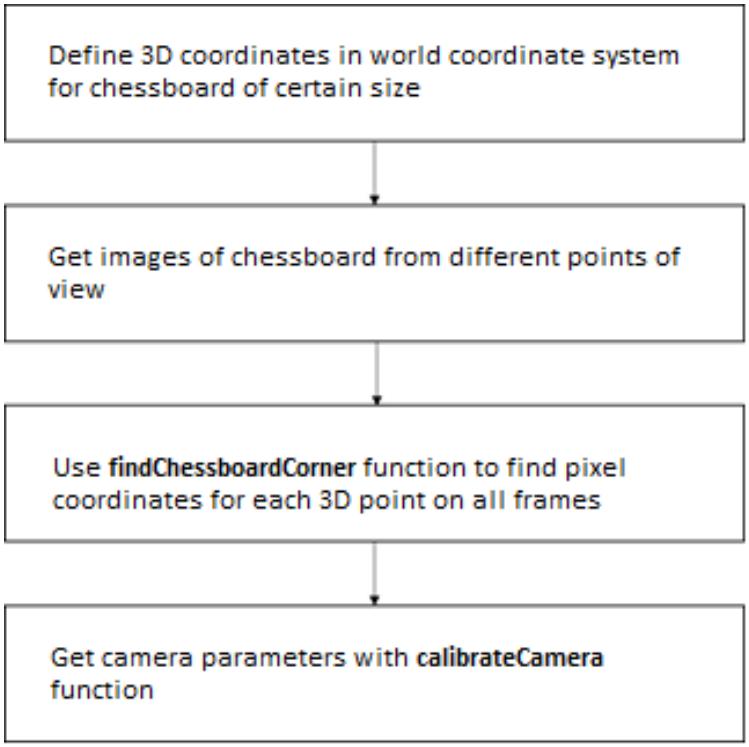

Figure 2: Calibration algorithm using OpenCV

Obtaining three-dimensional information about the scene and estimating the current position of the camera module can be done using the simultaneous localization and mapping module. Implementation of such a system based on LSD-SLAM allows obtaining a dense cloud of points. The average rate of receiving a new key frame in the system and estimating the position is approximately 5 and 10 Hz. As shown in Figure 3, depth estimation occurs mainly on contours in the image. This result is typical for the LSD-SLAM library. The principle of its operation consists in finding the difference in image intensity and finding correspondences on texture contours depending on the contrast of the scene.

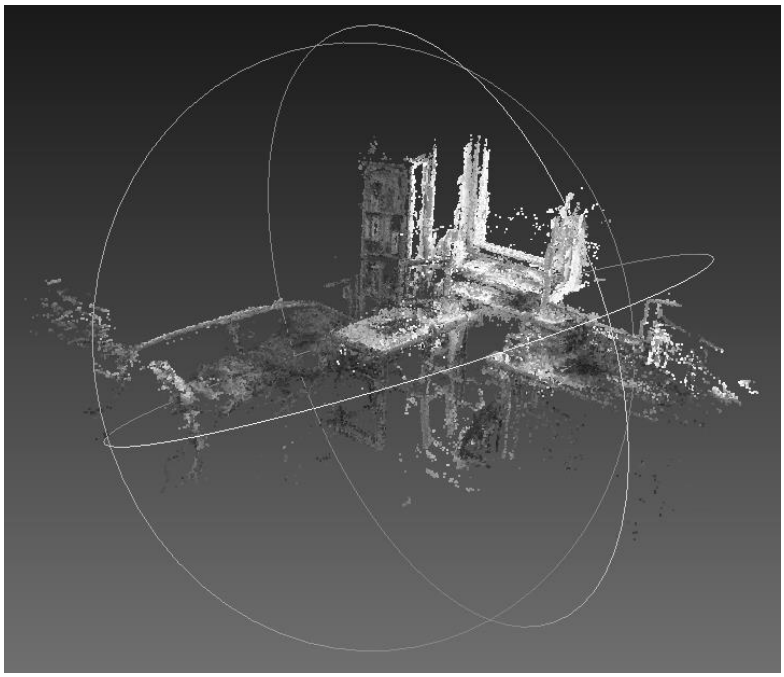

Figure 3: 3D model after reconstruction of the scene

For optimal reconstruction of the scene, it is necessary to ensure its stability. This module provides collection and fusion of reconstructed point clouds from different keyframes. Further optimization of the reconstruction can be done by post-processing the obtained point cloud.

3. Construction of a plane model from a cloud of points and its accuracy

This section deals with the processing of the cloud of points, which represents the surface of the target object. To assess the quality of the built surface model, separate planes were selected, since the deviation of the model from the plane was chosen as the quality criterion. To process the point cloud, planes are identified and segmented, for which an algorithm based on the Random Sample Consensus (RANSAC) method is considered.

Sparse methods for surface reconstruction estimate the surface from a sparse point cloud. It is sparse because it is computed from points of interest that have a non-uniform and sparse distribution in the images. However, the points have decent confidence due to a standard pipeline including point selection, robust and optimal reconstruction using RANSAC and nonlinear least squares.

Sparse methods are useful for its spatial and temporal complexity, especially for obtaining compact models of large-scale environments. This is good for computing with limited hardware or for applications that require high scalability and do not require the level of detail provided by dense stereo. Second, they can initialize dense stereo methods to improve accuracy and obtain more detailed reconstructions if the experimental conditions are favorable to obtain sufficient texture in the images.

A classical method of plane segmentation from a point cloud is the RANSAC algorithm. This method estimates the parameters of a mathematical model for a set of observed data containing a large number of outliers. It randomly selects a minimal set of points to estimate model parameters. From the random samples, it selects the one that best matches the full set of points. According to its general formulation, RANSAC can be easily applied to describe any primitive geometric shapes. However, the basic RANSAC approach assumes that the input data can belong to only one model.

The principle of the RANSAC algorithm is to find the best plane among a 3D cloud of points. At the same time, it reduces the number of iterations, even if the number of points is very large. To do this, it randomly selects three points and calculates the parameters of the corresponding plane. It then detects all points of the seed cloud belonging to the computed plane according to a given threshold. After that, it repeats these procedures N times; in each of them it compares the obtained result with the last saved one. If the new result is better, the saved result is replaced by the new one.

A prioritization function with a soft threshold [6], based on two weighting functions is used to improve the segmentation quality, which takes into account both the distance from the points to the plane and the consistency between the normal vectors. However, this requires estimating the normal vector at each point, which is inefficient in dense point clouds.

According to estimates, the time complexity of RANSAC depends on the size of the subset, the proportion of outliers, and the number of points in the set. RANSAC runtime can be excessively long in some cases. Therefore, a modification of the algorithm is considered for more effective detection of shapes in point clouds – including flat shapes. Octotrees are used to establish spatial proximity between samples and their scoring function considers only a local subset of samples. Local sampling by selecting points inside each node is used to avoid incorrect results.

The quality of 3D reconstruction is affected by several factors: the movement of the camera and environmental objects, spatial quantization of the image coordinates, correspondence of key points, camera calibration parameters, unaccounted camera distortions, as well as numerical and statistical properties of the selected reconstruction method. The impact of calibration errors on three-dimensional reconstruction was evaluated by comparing metrics for individual planes at different levels of artificially introduced error and evaluating the impact of the error on these metrics. For this purpose, the mean square deviation of the cloud of points was calculated and flatness was estimated based on it.

RANSAC algorithm requires four inputs:

- Three-dimensional point cloud;
- Tolerance threshold of the distance t between the selected plane and other points. It is related to the accuracy of the cloud of points;

- The forseeable-support (maximum probable number of points belonging to a single plane) -- it is derived from the density of points and the maximum estimated surface of the plane.
- The probability α (minimum probability of finding at least one good set of observations in N trials) -- it is usually between 0.90 and 0.99.

We have a cloud of points, which is a set of points $P_i(X_i, Y_i, Z_i)$ in the plane coordinate system. This coordinate system is such that the z-axis is perpendicular to the plane. The transformation that translates a point from the global coordinate system of the reconstruction to the local coordinate system of the plane can be represented as $P_L = R * P_G + T$.

We define a point $O(\bar{X}, \bar{Y}, \bar{Z})$ such that:

$$\bar{X} = \frac{\sum X_i}{n}$$
$$\bar{Y} = \frac{\sum Y_i}{n}$$
$$\bar{Z} = \frac{\sum Z_i}{n}$$

Consider the point O as the center of coordinates of the new local system. The coordinates of points in the new system can be represented:

$$x_i = X_i - \bar{X}$$
$$y_i = Y_i - \bar{Y}$$
$$z_i = Z_i - \bar{Z}$$

Let us represent the plane in the local coordinate system as z=ax+by, where a and b can be estimated from the following expressions (assuming that the deviations are measured along the z axis):

$$a = \frac{\sum {y_i}^2 \cdot \sum z_i \cdot x_i - \sum x_i \cdot y_i \cdot \sum z_i \cdot y_i}{\sum {x_i}^2 \cdot \sum {y_i}^2 - (\sum x_i \cdot y_i)^2}$$
$$b = \frac{\sum {x_i}^2 \cdot \sum z_i \cdot y_i - \sum x_i \cdot y_i \cdot \sum z_i \cdot x_i}{\sum {x_i}^2 \cdot \sum {y_i}^2 - (\sum x_i \cdot y_i)^2}$$

Calculate the deviation between the measured points and the segmented plane:

$$e_i = \frac{(z_i - ax_i - by_i)}{\sqrt{a^2 + b^2 + 1}}$$

The flatness deviation (FD) can be determined by the sum of the values of the maximum positive local deviation (TP) and the maximum value of the modulus of negative local deviation (FP):

$$FD = |{e^+}_{max}| + |{e^-}_{max}| = TP + FP$$

Completeness (C1), correctness (C2) and quality (Q) metrics for evaluating the presented method are expressed by the following representations:

$$C1 = \frac{FD}{FD + TP}$$
$$C2 = \frac{FD}{FD + FP}$$
$$Q = \frac{FD}{FD + TP + FP}$$

Here, TP is the number of valid planes that are correctly detected, FN is the number of planes that are unrecognized, FP is the number of incorrectly recognized planes.

4. Results of experiments

This section presents the results of point cloud reconstruction experiments on three different scenes. For each case, the segmentation of the studied plane was carried out and its model was obtained.

The main parameters of the data sets are presented in Table 1.

Table 1
Description of datasets

Index	Length, m	Width, m	Height, m	Number of points	Average density (points/m3)
1	7	5.2	3.2	45361	389
2	8.5	9.4	2.6	38265	184
3	7.3	7.2	2.5	37472	285

As can be seen from Fig. 2, the obtained three-dimensional models have sufficient density. This is due to the fact that the method used by the algorithm for obtaining a cloud of points uses all the information in the image, including contours. This provides high accuracy and reliability in low-textured environments using a single monocamera.

For each scene, plane segmentation was performed and point clouds, which represent sets of measurements belonging to one plane, were isolated. Next, the parameters of the mathematical model of each of the planes were evaluated. An example of a separate cloud of points corresponding to a noisy plane model is shown in Figure 4.

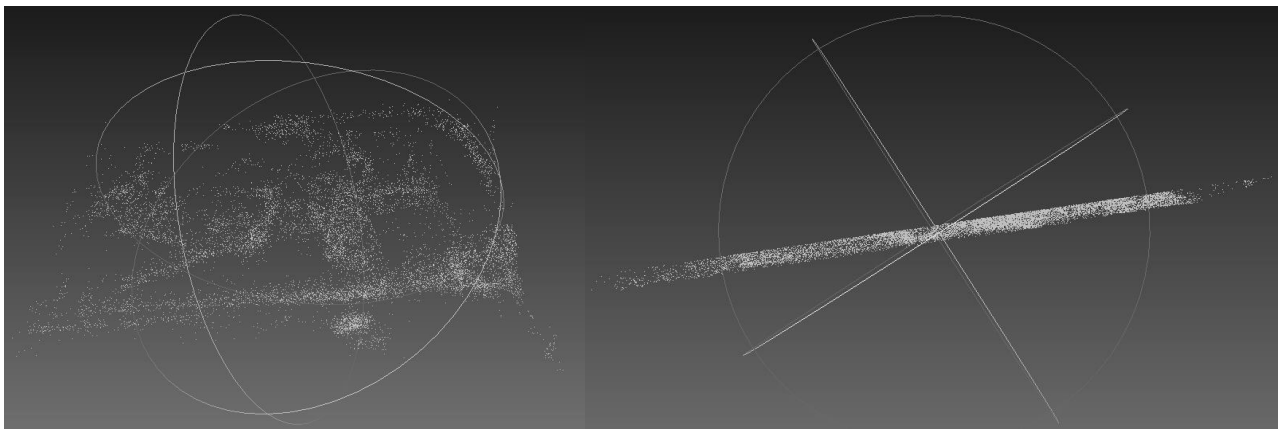

Figure 4: The result of plane segmentation

For each obtained plane and its corresponding point cloud, flatness deviations were evaluated and indicators of metric completeness, correctness and quality were calculated. Next, the effect of changing the calibration parameters on the metric data was investigated. For this, an analysis of the sensitivity of the camera parameters was carried out, in which the pixel values on the image plane were distorted by noise with a standard deviation of 0.05 to 1.0 pixels. Table 2 shows the sample results of the sensitivity analysis. In the simulated system, the camera was positioned in such a way that the direction of the z axis of the global and local coordinate systems coincided.

Table 2
The variance of the calibration parameters as a function of the noise

	0.05	0.1	0.5	1.0
p_x (px)	1.93	4.25	17.63	38.97
p_y (px)	0.43	0.86	4.87	9.75
t_x (m)	0.00	0.00	0.03	0.03
t_y (m)	0.00	0.00	0.02	0.01
t_z (m)	0.00	0.00	0.03	0.02
R_x (deg)	0.00	0.01	0.02	0.02
R_y (deg)	0.06	0.14	0.76	1.48
R_z (deg)	0.08	0.23	1.13	2.27

On the basis of the obtained noisy camera parameters, the process of reconstruction and segmentation of planes from the obtained models was carried out. For each such set of input data, flatness deviations were estimated. The dependence of the flatness deviation on the level of the introduced error for each of the three data sets is presented in the Figure 5.

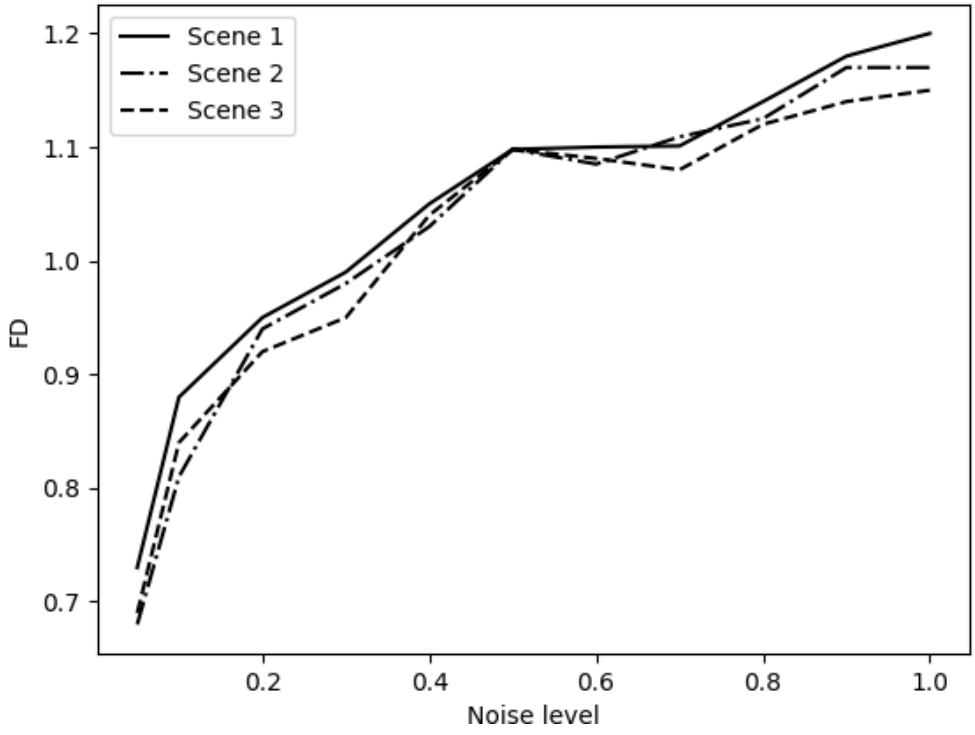

Figure 5: Dependence of flatness deviation on the error value of calibration parameters

This experiment demonstrated the dependence of the accuracy of the reconstructed point cloud on the deviations of the internal and external parameters of the cameras in the stereo system. It was found that increasing the accuracy of camera calibration can provide an opportunity to obtain an increase in the accuracy of a three-dimensional model by up to 60%.

5. Conclusions

The method of assessing the accuracy of the object model for the problem of stable grasping using a combined system of grasping proposal and reconstruction of the three-dimensional model of the object, which allows stable grasping of objects of unknown shape, is considered. The deviation of the model from the plane is chosen as the metric for accuracy assessment, so the point cloud is processed by plane identification and segmentation, for which an algorithm based on the RANSAC method is considered. An experiment was conducted to obtain the dependence of the accuracy of the reconstructed planes on the errors of the camera parameters. The impact of calibration errors on three-dimensional reconstruction was evaluated by comparing metrics for individual planes at different levels of artificially introduced error and evaluating the impact of the error on these metrics. Modeling the error of the camera calibration parameters with a given noise level shows that the calibration parameters deteriorate as the noise level increases. In particular, after analyzing the established flatness metrics, it was established that the error in determining the center of the image is proportional to the measurement error. It follows that an increase in the error contributes to an increase in the error in the estimation of the calibration parameters. In addition, orientation parameters (rotation and translation) are more complex and therefore more sensitive to measurement noise than other parameters.

6. References

[1] A. Doroshenko, O. Novak, Gesture simulator programming using statistical modeling. Problems of programming, 2015, pp. 58-64.

[2] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[3] Sundermeyer, Martin, Mousavian et al., Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes, 2021. URL: https://arxiv.org/pdf/2103.14127.pdf

[4] Zeng, Andy, Song et al, Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching, 2020. URL: https://arxiv.org/abs/1710.01330

[5] B. Staub et al., Dex-Net MM: Deep Grasping for Surface Decluttering with a Low-Precision Mobile Manipulator, 2019 IEEE 15th International Conference on Automation Science and Engineering (CASE), Vancouver, BC, Canada, 2019, pp. 1373-1379, doi: 10.1109/COASE.2019.8842901.

[6] A. Vitiuk, Y. Kornaga, A. Barabash, Capturing unknown objects by a mobile robot using visual information, Scientific notes of V. I. Vernadskyi Tavra National University. Series: Technical sciences, 2018, Vol. 29(68), No. 1(1), p. 93-98.

[7] D. Yang, T. Tosun, B. Eisner et al., Robotic Grasping through Combined Image-Based Grasp Proposal and 3D Reconstruction, 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 6350-6356, doi: 10.1109/ICRA48506.2021.9562046

[8] B. Xu, Z. Chen, Q. Zhu et al., Geometrical Segmentation of Multi-Shape Point Clouds Based on Adaptive Shape Prediction and Hybrid Voting RANSAC. Remote Sens, 2022. URL: https://doi.org/10.3390/rs14092024