IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 14, 2019, accepted April 18, 2019, date of publication May 1, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2913421

Robust Task-Oriented Markerless Extrinsic Calibration for Robotic Pick-and-Place Scenarios

YONG ZHOU[1], QIANG FANG[1], KUANG ZHAO[1], DENGQING TANG[1], HAN ZHOU[1], GUOQI LI[2], XIAOJIA XIANG[1], AND TIANJIANG HU[3], (Member, IEEE)

[1]College of Intelligence Science and Engineering, National University of Defense Technology, Changsha 410073, China
[2]Department of Precision Instrument, Tsinghua University, Beijing 100084, China
[3]Machine Intelligence and Collective RObotics (MICRO) Lab, Sun Yat-sen University, Guanghzou 510275, China

Corresponding author: Tianjiang Hu (hutj3@sysu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Granted No. 61703418 and 61803377.

ABSTRACT Camera extrinsic calibration is an important module for robotic visual tasks. A typical visual task is to use a robot and a color camera to pick an object from a variety of items and place it in a designated area. However, the noise of multi-sensor processing may have a significant impact on the results when running a full-process visual task; in addition, checkerboards are inconvenient or unavailable in pick-and-place scenarios. In this paper, we propose and develop a task-oriented markerless hand-eye calibration method by using nonlinear iterative optimization. The optimization employs a transfer error to construct cost function, which is necessarily observable and estimable for visual tasks. Our method does not require a calibration checkerboard and only uses an available saliency object in the task scene as a marker. It provides an end-to-end method that converts extrinsic parameters into variables that are optimized with the cost function, making it not only robust to sensors with noise but also able to meet the requirements of the tasks' reconstruction accuracy. Different from classic methods detecting a known size calibration pattern, the input of our method is a batch of image points and the corresponding world points. The results show that the accuracy of our extrinsic calibration method is sufficient for the robot's pick-and-place tasks. The experiments of the competition demonstrate that our method is definitely effective in the desired tasks of vision-in-the-loop automatic pick-and-place scenarios.

INDEX TERMS Markerless extrinsic calibration, grasping, reconstruction accuracy.

I. INTRODUCTION

Robotic pick-and-place has a wide range of applications in the industrial fields [1], [2], such as handling and transporting goods in intelligent logistics and warehouse, as well as grasping and classifying objects in clutters scenes. There is an urgent need for robots to complete tasks automatically through machine vision in the field of industry. High accuracy extrinsic calibration is a key issue to precisely get the position of the objects and achieve industrial automation. It is generally called hand-eye calibration, which is usually used to solve two types of problems [3], [4]: one is to establish a mapping between the sensor and the robot workspace frame, *e.g.* robotic pick-and-place an object; the other is to determine the accurate displacement and rotation between two sensors, for example, robot-equipped camera and Inertial Measurement Unit (IMU). Regardless of any types, the goal of hand-eye calibration is to solve the problem of visual measurement directly or indirectly.

In robotic industrial application where machine vision is employed for grasping, there are eye-in-hand system and eye-to-hand system depending on the camera's bearing of the installation as is shown in Figure 1. The classic hand-eye extrinsic calibration method mainly focused on a calibration pattern with known size to establish the mapping between calibration pattern coordinate frame and image pixel coordinate frame. The corner points in calibration pattern coordinate frame can be determined with the pattern size and the corresponding image pixel coordinate can be captured with corner extraction algorithm. Thus, the extrinsic can be optimized by multiple images.

The classic extrinsic calibration method can effectively solve hand-eye issues; the accuracy and efficiency are mostly satisfactory. Certainly, assumptions are inevitably considered

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Moinul Hossain.

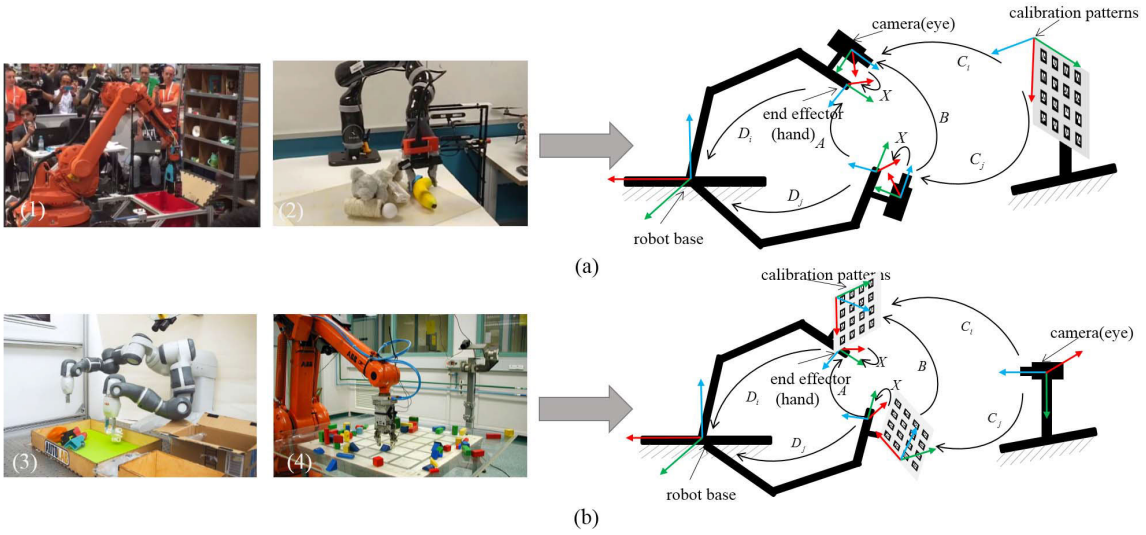


FIGURE 1. Hand-eye system and pick-and-place application. (a) Eye-in-hand system: The camera mounted on the robotic arm, and it will move with the movement of the arm. (1) MIT-Princeton robotic picking system [6]. (2) Queensland University of Technology grasping scene. (b) Eye-to-hand system: The camera is mounted outside the arm. (3) Ken Goldberg's AUTOLAB Grasping Robot. (4) ABB industrial robots.

for practical applications. Typically, in the robotics task-oriented pick-and-place scenarios, the joint angle and end-effector pose of the robot are inaccurate, which cause error propagation and accumulation. In addition, extrinsic calibration methods with checkerboard are sensitive to the accuracy of checkerboard's flatness and size, and ambient light as well. Light reflection will cause corner detection errors, which will directly affect the calibration results.

Previous researches mostly focus on the hand-eye calibration equation of $AX = XB$ or $AX = ZB$ as is described in Figure 1, which is difficult to address the issues mentioned above. Furthermore, it is sometimes practically inconvenient or unavailable to get the checkerboard in the pick-and-place scenarios. This paper aims to explore an end-to-end checkerboard-free extrinsic calibration method that can be effectively to avoid error propagation and accumulation. It constructs the cost function through the physical world and image pixel, and transforms the extrinsic calibration problem into the parameter optimization problem of the function model. We preliminarily focus on the task-oriented robotic pick-and-place scenarios where the objects are rigid body that can be captured and identified placed on a flat surface. To this end, our contributions are summarized as follows:

- We propose an end-to-end extrinsic calibration and optimization method without using checkerboard. It can only use a saliency objects available in hand.
- We use transfer error [5] to construct the cost function. The extrinsic is transformed into optimized parameters of the cost function model. The parameters can be complete extrinsic parameters, in terms of rotation matrix and translation vector.
- We present a nonlinear iteration algorithm to optimize cost function. This optimization algorithm is suitable for strictly convex functions.

The remainder of this paper is organized as follows. In Section II, we review relevant and state-of-the-art related works. The methodology proposed is presented in Section III. The experiments are conducted and corresponding results are presented and discussed as well in Section IV. Section V draws the conclusions and discusses future works.

II. RELATED WORKS

There is extensive scholarly research on hand-eye calibration for robotic grasping. The research of classic methods is mainly concentrating on solving the equation $AX = XB$, such as [7]–[9], but their precision is similar. Moreover, Horaud and Dornaika in their study [10] use the Levenberg-Marquart method to solve the equation. Malti [11] refine hand-eye and the camera intrinsic and distortion parameters simultaneously with epipolar constraints and reprojection errors [12] to minimize respectively. Different from the reprojection error generated from two image plane connected by a homography matrix, our method uses transfer error to minimize the cost function. By constructing the cost function, the extrinsic parameters are transformed into the variables of the cost function and then minimizing it to get optimal values.

Besides dealing with the equation $AX = XB$, much of the hand-eye calibration research has focused on identifying and evaluating the equation $AX = ZB$. One common approach to solve the equation $AX = ZB$ is separable methods, which decompose it into a purely rotational part R_{3x3} and a purely translational part t_{3x1} and then estimates R^X_{3x3} and t^X_{3x1} respectively. Shah in [13] created a separable closed-form solution to resolve the hand-eye calibration problem. He used a special method that involves the Kronecker product and the singular value decomposition. Tobb and Yousef [3], [14] use an iterative method with Euler angles to parameterize the rotation components to reduce the camera reprojection

error and they provide a method with different choices of cost function, parameter choices and separable versus simultaneous solutions. Another typical method is simultaneous solutions. In view of the consideration of both aspects of the rotation and translation simultaneously, this method will avert to propagate the error of the rotation to the translation error. Strobl and Hirzinger [15] use nonlinear optimization methods to estimate hand-eye calibration rotation and translation. Though it can work with both the formulations $AX = XB$ and $AX = ZB$, a metric on the rigid translations and the corresponding error model are required. Horaud and Dornaika [10] simultaneously estimate hand-eye parameters by means of global iterative optimization.

Recently, a growing number of researcher have paid more attention to grope for new extrinsic calibration methods. Iyer *et al.* [16] use a geometrically supervised deep network to estimate the six degree of freedom (6-DOF) rigid body transformation between 3D Light Detection and Ranging (LIDAR) and 2D camera, which give a good inspiration for the hand-eye calibration. Point cloud registration is applied in [17] for hand-to-eye system to estimate the extrinsic parameters through contact-based interaction, which doesn't rely on any fiducial markers or calibration rigs. But they need an additional contact sensor to get contact information. Pachtrachai *et al.* [18] solved unsynchronized hand–eye calibration that the data streams are different capture rates and time delays. They use cross-correlation to synchronize data and use screw constrains to recover data.

Moreover, several systematic studies and reviews of extrinsic calibration optimization have been undertaken. In consideration of sensor noises, according to [19], [20], there are some approaches using Kalman filter to estimate the calibration between camera and IMU. This is a practical method to reduce noise and improve accuracy. Huang and Stachniss [21] systematically studied and compared the accuracy of the three calibration method $AX = B$, $AX = XB$ and $AX = ZB$, showing that in some cases, the motion-based calibration method is superior to the marker-based method. This gives us theoretical support with a markerless approach. Lina and Shen in [22] presents an online markerless approach, which use 5-DOF and nonlinear optimization to calibrate stereo extrinsic. Furrer *et al.* [23] created and provided a collection of datasets for hand-eye calibration and validated different filtering methods on these datasets.

III. PEOPOSED METHOD

A. GEOMETRIC MODEL

In this section, we begin our method by defining geometric notations and coordinate frame. Assuming the camera is a pinhole model [5], we can quickly compute that the point $P(X, Y, Z)^T$ in the 3D space is mapped to the point $p(u, v)^T$ on the image plane by similar triangles. Denote the focal length of camera as f; then we have the equation as follows:

$$(X, Y, Z)^T \mapsto \left(fX/Z, fY/Z\right)^T \tag{1}$$

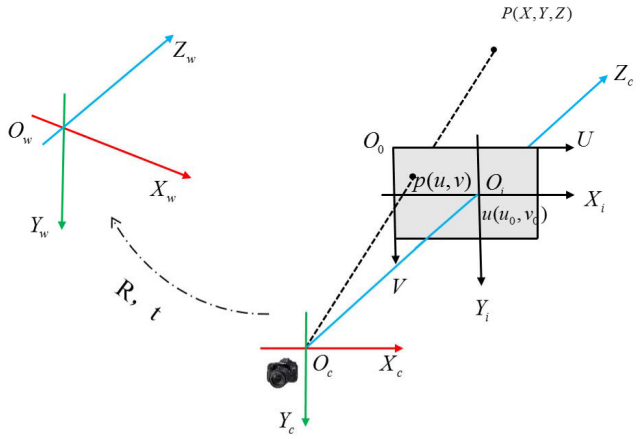


FIGURE 2. Coordinate frames.

Based on the pinhole camera model, we define four coordinate frames as shown in Figure 2: pixel coordinate frame $(O_0 - UV)$, which is located in the image plane, with principal point O_0 in the top left corner of the image, pixel for unit; image coordinate frame $(O_i - X_iY_i)$, which is paralleled with Pixel coordinate frame, but its principal point O_i is the center of the image, meter for unit; camera coordinate frame $(O_c - X_cY_cZ_c)$ located in the optical center, with coinciding with optical axis, X and Y axis both perpendicular to Z axis, in line with the right-hand coordinate frame; and world coordinate frame $(O_w - X_wY_wZ_w)$ is located in the three-dimensional physical space, which is custom coordinate frame. All of the coordinate frames satisfy right-hand rule. Generally, points in space are expressed in the Euclidean coordinate frame, known as world coordinate frame. In the case where the camera intrinsic matrixes are known, we can establish the mapping relationship between the pixel points on the image and the camera coordinate frame. Further, to determine the coordinates of the image pixel in the world coordinate frame, the transformation between $O_c - X_cY_cZ_c$ and $O_w - X_wY_wZ_w$ needs to be known. The two coordinate frames are related with a rotation matrix R and a translation vector t. According to Zhang [24], a point $(X_w, Y_w, Z_w)^T$ in 3D physics space is mapped to pixel coordinate $(u, v)^T$. We can get the homogeneous coordinates as described in the equation

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = M\begin{bmatrix} R & t \end{bmatrix}\begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{2}$$

where s is an arbitrary scale factors, (R, t) is extrinsic parameters which we denote as $\Phi = (R, t)$; M is camera intrinsic matrix.

B. FORMULATION

Reprojection error is used to represent the error between the projected point and the measured one of the images. As is shown in Figure 3 (a), p_i and p_i' are different projections of the

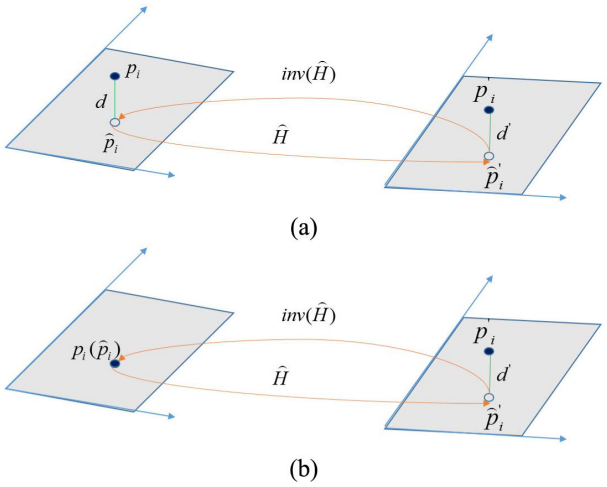


FIGURE 3. (a) Reprojection error. (b) Transfer error.

same point in space. We have the reprojection error equation as follows:

$$\varepsilon = \sum_i \left(d^2 \left(p_i, \hat{p}_i \right) + d^2 \left(p_i^{'}, \hat{p}_i^{'} \right) \right) \tag{3}$$

where $\hat{p}_i^{'} = \hat{H}\hat{p}_i$, $\hat{H}$ is a homography between two images; and d represents the Euclidean distance of two points.

If we consider error only in the second image with the first is measured perfectly, the above expression is degenerated to transfer error defined as Eq. (4). A comparison between reprojection error and transfer error is shown in Figure 3.

$$\varepsilon = \sum_i d^2 \left(p_i^{'}, \hat{p}_i^{'} \right) \tag{4}$$

As described in Eq. (2), 2D pixel (u, v) and 3D world point (X, Y, Z) is converted through rigid-body motion and perspective projection with the parameters of extrinsic and intrinsic respectively. Assume the world points are in a plane; thus this model is transformed into a map of image plane and world plane. We denote world plane as π, world points as $W(x)$, and image plane as $\pi^{'}$, image points as $I(x)$. Thus, we have $W(x) \in \pi$ and $I(x) \in \pi^{'}$. Then, the map from image plane to world plane is represented as:

$$\pi^{'} \mapsto \pi \tag{5}$$

Their equation is given by:

$$\hat{W}(x) = g(M, \Phi, I(x)) \tag{6}$$

Rotation of extrinsic is represented by the Euler angles as

$$R = R(\varphi, \theta, \psi) \tag{7}$$

Thus, the rotation angles are $R(\varphi, \theta, \psi) \in R^3$ and translation vector is $t \in R^3$. Then, cost function is constructed as:

$$g(x) = \sum \left\| W(x) - \hat{W}(x) \right\|^2$$
$$s.t.\ \hat{W}(x) = g(M, \Phi, \hat{I}(x)) \tag{8}$$

Consequently, extrinsic is transferred as the parameters of objective function to be optimized. We can optimize the 6DOF parameters that is given by Eq. (9) or we only optimize the rotation $R(\varphi, \theta, \psi)$ and just initialize the translation t as shown in Eq. (10).

$$[R(\varphi, \theta, \psi) | t] = \arg\min g(x) \tag{9}$$
$$R(\varphi, \theta, \psi) = \arg\min g(x) \tag{10}$$

The classic hand-eye calibration methods for $AX = XB$ or $AX = ZB$, estimate the homography $\hat{H}$ with two or multiple sets of corresponding images. In consideration of error in each of the two images, an alternative method of quantifying error involves estimating a correction for each correspondence, known as reprojection error. However, the proposed algorithm establishes the relationship between the coordinates of the plane marker points in the world frame and the corresponding points on the image. The coordinates of the marker points in world frame can be strictly accurately measured. In this case, the appropriate quantity to be minimized is the transfer error.

C. ITERATIVE MINIMIZATION METHOD

Our proposed extrinsic calibration method is derived from task-oriented robotic pick-and-place tasks, which is simultaneously in consideration of extrinsic calibration and reconstruction as shown in Figure 4. When calibration datasets have been collected, world points and image points are input in objective function, where image points are converted to the world coordinate with the process of perspective projection and rigid-body motion. Then, extrinsic becomes part of decision variables. Certainly, we can select part of extrinsic as the decision variables. In this paper, we studied the model with optimizing translation (w/) and without optimizing translation (w/o). Experiments show their difference in Section IV. In addition, constraint conditions are given by Eq. (8); **Algorithm1, 2** are employed to optimize the objective function as iteration method. In the part of reconstruction, we use pixel points from images and calibrated extrinsic to estimate their position in physics space. Naturally, reconstruction accuracy is used as the accuracy evaluation metric.

Algorithm 1 Optimizing Extrinsic

Input: $W(x_1, x_2, \ldots, x_n)$, world points
$I(x_1, x_2, \ldots, x_n)$, image points
while convergence condition not satisfied
if rotation and translation to be optimized **then**
$[R(\varphi, \theta, \psi) \mid t] = FS(g(x))$;
end
If rotation-only to be optimized **then**

$$R(\varphi, \theta, \psi) = FS(g(x));$$

end
end
return $[R(\varphi, \theta, \psi) | t]$ or $R(\varphi, \theta, \psi)$

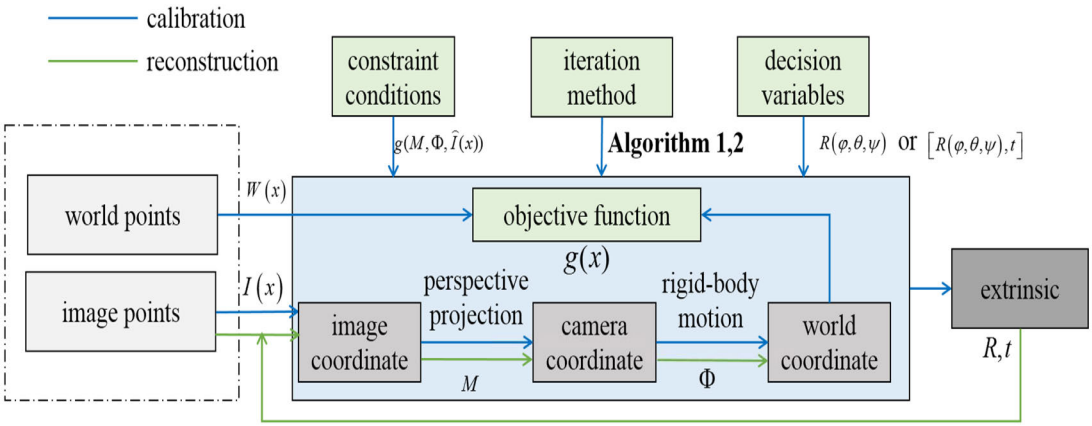


FIGURE 4. Overall system framework.

Algorithm 2 One Iteration With Simplex FS(·)

Denote $\phi(i)$ as the list of points, $i = 1, 2, \ldots, n+1$,.
Order the points from lowest function value $g(\phi(1))$ to highest $g(\phi(n+1))$.
$m = \sum \phi(i)/n, i = 1, 2, \ldots, n$;
$r = 2m - \phi(n+1)$;
if $g(\phi(1)) \leq g(r) < g(\phi(n))$ **then**
  $\phi(n+1) = r$;
else if $g(r) < g(\phi(1))$ **then**
  $s = m + 2(m - \phi(n+1))$;
  if $g(s) < g(r)$ **then**
    $\phi(n+1) = s$;
  else
    $\phi(n+1) = r$;
  end if
else if $g(r) \geq g(\phi(n))$
  if $g(r) < g(\phi(n+1))$ **then**
    $c = m + (r - m)/2$;
    if $g(c) < g(r)$ **then**
      $\phi(n+1) = c$;
    else
      $v(i) = \phi(1) + (\phi(i) - \phi(1))/2, i = 1, 2, \ldots, n+1$;
      The next iteration is $\phi(1), v(2), \ldots, v(n+1)$;
  else
    $c = m + (\phi(n+1) - m)/2$;
    if $g(c) < g(r)$ **then**
      $\phi(n+1) = c$;
    else
      $v(i) = \phi(1) + (\phi(i) - \phi(1))/2, i = 1, 2, \ldots, n+1$;
      The next iteration is $\phi(1), v(2), \ldots, v(n+1)$;
  end if
end if

We propose an iteration method to optimize extrinsic in **Algorithm1.** For the input of world points and image points, it is optional to optimize both rotation and translation, or rotation-only. When the convergence condition is dissatisfied, **Algorithm1** will search a better solution within each iteration. For each iteration step, **Algorithm2** is used to search with simplex [25] properties. After initialization, this algorithm constructs a simplex of $n + 1$ points and orders these points from lowest function value $g(\phi(i))$ to highest. The simplex updates according to the procedure of the algorithm. At each step, the algorithm use another point to replace current worst point $\phi(n + 1)$ in the simplex.

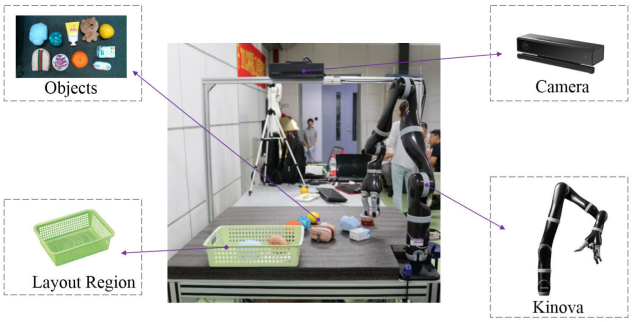


FIGURE 5. Experimental scenario and data acquisition.

IV. EXPERIMENT AND RESULTS

A. EXPERIMENTAL SETUP

A classic object pick-and-place vision task is presented in Figure 5 with a camera installed over the table and various objects on the table to be grabbed to the desired regions. On the right is a 7 DOF Kinova robotic arm. The basket next to Kinova is utilized to place the object grabbed. After the objects are designated, the robot automatically grabs the objects one by one from the table and puts them in the basket. Undoubtedly, recognizing objects and knowing where they are is one of the most important steps. In our experiment, we use camera to capture images and Kinova to read the position and orientation of the end-effector from the robot controller.

The specific steps of estimating the relationship between the camera coordinate frame and robot base coordinate frame are as follows:

(1) Define the camera coordinate frame and the world coordinate frame as Figure 2. The world coordinate frame is exactly the robot base coordinate frame.

(2) Calibrate camera intrinsic parameters and distortion coefficient and measure the translation t between the two coordinate frames.

(3) Place an available object on the workbench and move the end-effector of the arm in the state of gripping the object. Then capture the camera image and record the end-effector's position in the world coordinate frame.

(4) Repeat step (3) at least 15 times. Then we can obtain the datasets of images and the corresponding end-effector's position. The end-effector's position is regarded as world points $W(x)$. The image points $I(x)$ that are synchronized with world points are obtained with vision-based detection algorithm.

(5) Input the image points and corresponding world points into the proposed algorithm and initialize the rotation angle; then we can get the optimal value of the rotation in the form of Euler angles.

To increase the scientific rigor and reliability of the method, we performed multiple quantitative experiments to validate our proposed method. Before showing the experimental results, we will introduce the error metrics first.

B. ACCURACY METRICS

To analyze and compare the results of our proposed method, reconstruction accuracy error is used to evaluate the results of our method. The reconstruction accuracy is to determine the position of the target in the world coordinate frame from the pixel coordinates of the target image by the obtained extrinsic of the camera and the robot, and compare the distance from the real position. The position of the target in world coordinates from the image pixels can be determined simply with the methods of pinhole camera model and rigid body coordinate transformation.

When robots are used to perform vision tasks, we are more interested in how the tasks are completed. Extrinsic is mainly used to solve two types of problems as described in section I, so starting from task drive, we use reconstruction accuracy error to evaluate the effect of task completion. For example, when we use the robotic arm to grasp an object, we want to know the distance between the target position determined by visual methods and its ground truth position. If we denote n and i as the total number of samples and component of it respectively, the reconstruction accuracy error is defined as the average Euclidean distance between calibration object points $\vec{\chi}_i$ and the estimated pints $\widehat{y}_i$. Namely,

$$\eta = \frac{1}{n}\sum_{i=1}^{n}\left\|\vec{\chi}_i - \widehat{y}_i\right\| \tag{11}$$

C. RESULTS AND DISCUSSION

1) ALGORITHM ITERATION AND RECONSTRUCTION ACCURACY

In this section, the methods previously presented are validated with real datasets. The datasets of images are collected with a color camera and a vision-based detection algorithm was used to detect the calibration object to get image points $I(x)$. The camera's intrinsic parameters are pre-calibrated. The ground truth of world points $W(x)$ was obtained from the robotic arm.

Figure 6 shows the loss of iterative optimization function with and without optimizing translation in the color of green and cyan respectively. In both of the models, the initial value of Euler angles are set as $euler = [\text{pi}\ \ 0\ \text{pi}/2]^T$, where pi is the ratio of circumference to diameter. In the model of w/o, the displacement of the two coordinate frame is set to a fixed value t, denoted $t = [-0.52\ \ -0.59\ \ 0.7]^T$; while the model w/ set t as initial value. The convergence satisfies $\varepsilon < 1e-4$; the maximum number of iteration is not limited. We used 30 batch size of data for the iterative optimization.

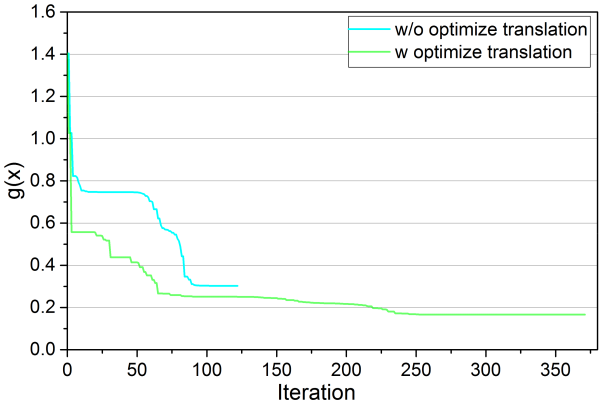


FIGURE 6. Loss accuracy of cost function with optimizing translation and without optimization translation.

TABLE 1. Iteration results.

Parameter	W/ Optimize Translation	W/O Optimize Translation
Iteration	371	121
Time/s	162.2	69.5
Final Value	0.166	0.303
η/m	0.00898	0.00791

In both of the two models, iterative optimization starts from the same point. The first one is optimized for a three-dimension vector that stops after iterations more than a hundred times; the second includes rotation and translations of a six-dimension vector that converges after more than three hundred iterations, and its convergence value is lower than the first one as well. As TABLE 1 shows, there is a significant difference between the two models, including the number of iterations, consumed time, η, and the final value of the cost function. Datasets, including optimization samples and test samples, used in both of the two models are the same. What is interesting about the data in this table is that although the number of iterations 371 takes more time and the cost function drops more, its reconstruction accuracy doesn't perform better in test samples. This suggests that higher dimensional optimization variables do not necessarily give better results.

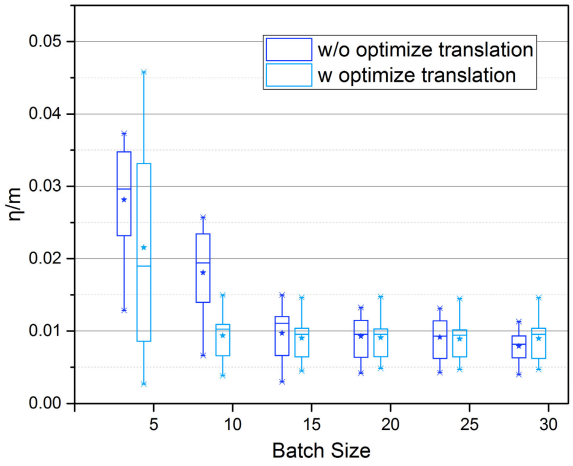


FIGURE 7. Reconstruction accuracy error of without optimizing translation and with optimizing translation.

The proposed algorithm can only optimize the rotation, or optimize the rotation and translation at the same time. To further explore their differences and effects on reconstruction accuracy, a comparative experiment was conducted and their results are presented in Figure 7. As is illustrated in Figure 7, we selected every 5 batch size of data as iterative samples, from 5 to 30. For each experiment, we can get an estimated extrinsic result. Then the optimized results of translation and Euler angles are used to perform reconstruction accuracy error testing in the test samples that contain 20 batch size of data. It is apparent from Figure 7 that as the batch size increases, there is a clear trend of decreasing of reconstruction error in the two models of with and without optimizing translation decreases. Although η is high when the number of points is small, η drops significantly as the number of iteration samples increases, and finally in both of the two models is less than 1 cm. When the number of iteration samples is more than 10, reconstruction accuracy of w/ remains dynamic stable, but the model of w/o is still a slow downward trend. Considering each independent pick-and-place task, the maximum of reconstruction accuracy error seems to be more convincing. Through numerous experiments, we observed that when η is less than 15mm, it is sufficient for our picking tasks. If the number of iteration samples is over 15, it can meets our pick-and-place indicator requirements.

Comparatively analyzing as presented in TABLE 1 and Figure 7, considering efficiency and accuracy, performance of w/o is better. Especially when there are a large number of robots and sensors that need calibration, such as robot swarm, w/o has a greater advantage. In addition, in order to maintain better accuracy and stability, or in other words, in order to make the variance of reconstruction accuracy smaller and the mean lower, the model with optimizing translation should use more than 10 iteration samples and the model w/o should use more than 15.

Figure 8 shows the estimated reconstruction results under the robot base coordinate frame space, including optimization samples and test samples. Optimization samples are used for iteration and optimization, while test samples are used for testing the proposed algorithm's reconstruction accuracy.

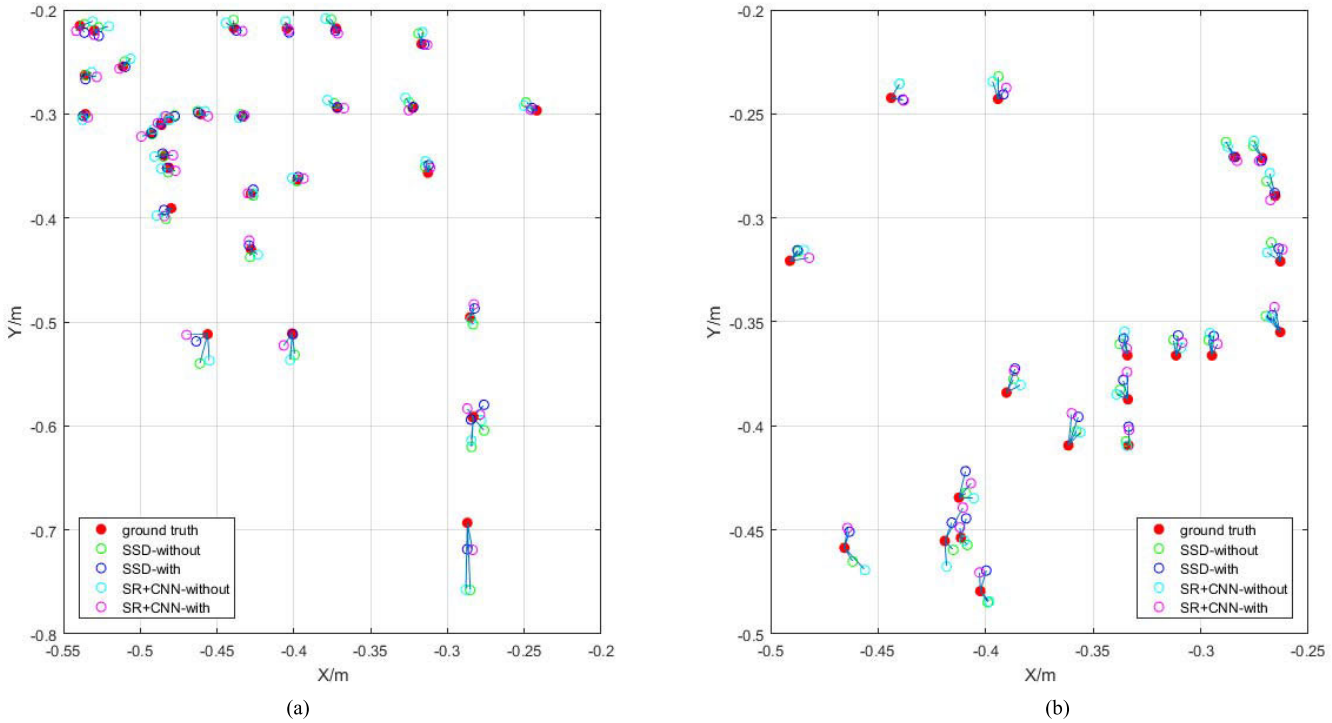


FIGURE 8. Reconstruction results of optimization samples (a) and test samples (b). The blue line connecting the ground truth and the estimated value indicates the reconstruction accuracy error.

As a key part of automatic acquisition of image points, we use SSD [26] and saliency detection (SR + CNN) [27] algorithms to detect the calibration object for experiments. Considering that it is also necessary to recognize objects for vision-based grasping tasks, training an object detection algorithm in the process of extrinsic calibration never brings additional work. The average reconstruction accuracy error of SSD-without, SSD-with, SR + CNN-without and SR + CNN-with is 9.22, 6.92, 10.4 and 8.10 mm. Although their image detection results are different, they all can satisfy the accuracy requirements of the object grasping. Additionally, no matter the model of w/o or w/, test samples show a relatively stable reconstruction error, which is very beneficial for grasping tasks. Because for the pick-and-place tasks, each picking action can have a certain range of acceptable errors. But too much error can lead to the failure of current picking task, even in the next picking try, the reconstruction error is quite small. It is experimentally found that reconstruction accuracy error less than 15mm is sufficient for completing a grasping task. Therefore, the results of test samples demonstrate the effectiveness of our method for automatic pick-and-place scenarios.

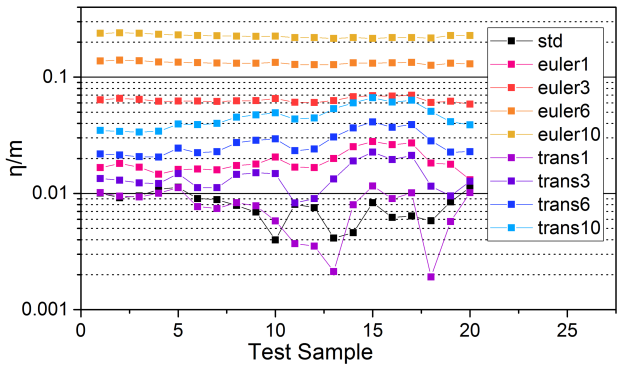


FIGURE 9. Rotation and translation noise's effects on reconstruction accuracy.

2) WEIGHTS COMPARISON BETWEEN ROTATION AND TRANSLATION

Actually, rotation and translation of the extrinsic parameters have different weights on the reconstruction accuracy. Figure 9 displays a comparison of their differences in terms of reconstruction accuracy. We get the optimal value of the extrinsic as the standard value, and then artificially add different percentages of disturbance to the translation and rotation respectively, including 1%, 3%, 6%, 10%. Five percent of noise is a remarkable error, but in order to compare the different weights of the rotation and translation, we artificially added a maximum disturbance of 10%. Intuitively speaking, in our experiment, the translation introduces a 1% error of approximately 5 mm, while the rotation Euler angle introduces a 1% error of approximately 1.8 degrees. An error of 5 mm in length is more difficult to occur than an error of 1.8 degrees in angle in the process of measurement. Or, under the most condition generally, the measurement error will be less than 5 mm in length and greater than 1.8 degrees in angle. However, from the test results of η, it can be seen roughly that under the same proportion of disturbance, the reconstruction accuracy of the Euler angle is much larger than the translation. Even when the translation increases by 1% of disturbance, the average η of the test sample is 7.65 mm, which is even smaller than the standard value of 7.91 mm as far as reconstruction accuracy. Considering that the translation is generally at the meter level and the 1% disturbance is at the centimeter level, the proposed extrinsic calibration methods have a good robustness within a controllable disturbance range.

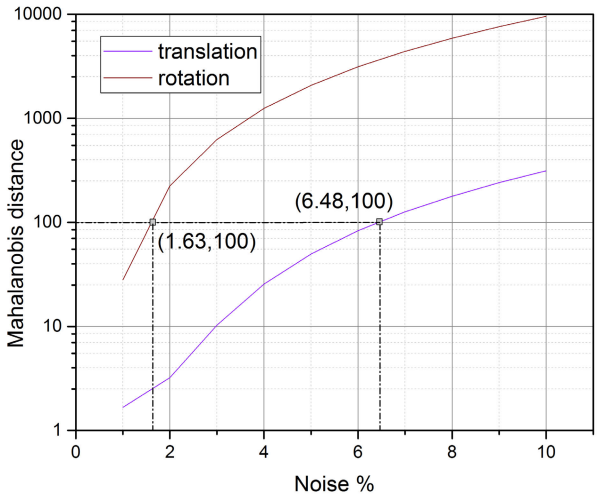


FIGURE 10. Mahalanobis distance of reconstruction accuracy.

Although the weight of the effects of rotation and translation through the disturbance percentage can be roughly estimated, however, because of the different dimensions of length and angle, it cannot be directly compared. In order to observe the effect of rotation and translation on the reconstruction accuracy of extrinsic, we introduced the Mahalanobis distance. Mahalanobis distance is an effective method for calculating the similarity of two unknown sample sets, which is unitless and scale-invariant. Figure 10 shows the relationship between noise intensity and Mahalanobis distance. In the Mahalanobis distance space, the curves of translation and rotation are significantly diverse. In the Euclidean space, the difference between the two may not be obvious, but there is a noteworthy difference in the Mahalanobis space. For example, when the Mahalanobis distance is 100, the translation brings 6.48% error, while the rotation only brings 1.63% error. In other words, the rotation only needs to introduce 1.63% of the noise, which can produce Mahalanobis distance of 100. As the noise intensity increases, the difference between the two is increasingly obvious. A possible explanation for this might be that the translation's error only increases the error of the same scale in the reconstruction accuracy, while the rotation's error will be amplified as the depth increases. As a result, the scale of the error is related

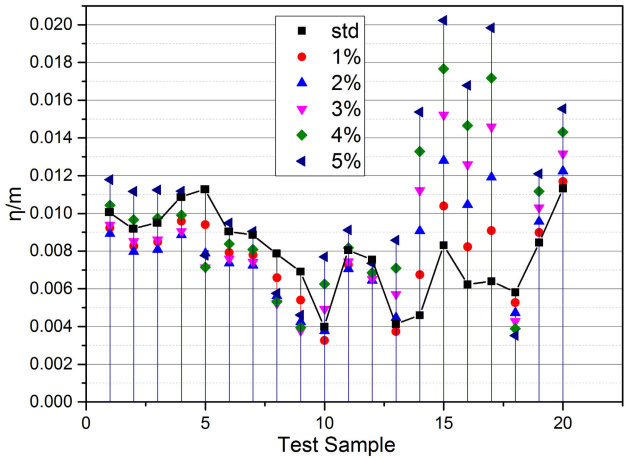


FIGURE 11. Reconstruction accuracy of test samples that introduce disturbance of camera intrinsic matrix and distortion coefficient.

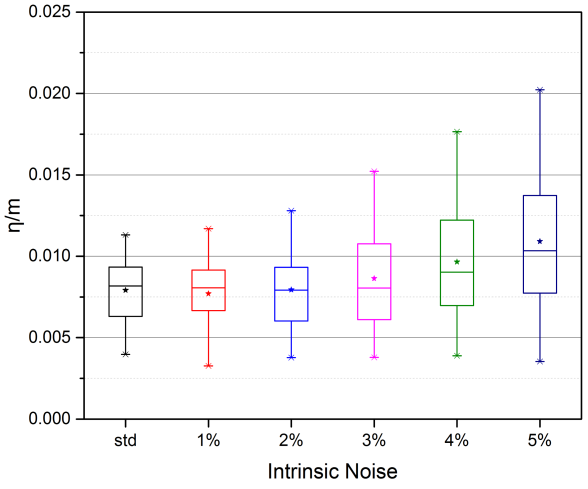


FIGURE 12. Average reconstruction accuracy of test samples that introduce different percentages noise of intrinsic matrix and distortion coefficient.

to the distance between the camera and the working plane of the robotic arm.

3) ROBUSTNESS TEST AGAINST CAMERA INTRINSIC NOISE

Camera's intrinsic matrix and distortion coefficients are generally essential for vision tasks or reconstructions. However, there are certain errors in the commonly used camera intrinsic parameters methods [24]. Existing extrinsic calibration methods hardly take the camera intrinsic noise into account. However, the intrinsic parameter error is a non-negligible factor in vision tasks. From the perspective of task-oriented, our method considers the intrinsic parameters noise of the camera. As presented in Eq. (6), the camera intrinsic parameters are optimized together to improve the accuracy of vision localization in the process of extrinsic optimization. Figure 11 and Figure 12 show the reconstruction accuracy results of the test samples by optimizing the extrinsic parameters when the camera intrinsic parameters are added disturbance. We use the intrinsic parameters obtained by [24] as the standard value, and then add noise from 1% to 5% based on this standard value. In the optimization process and the reconstruction process, we all use the same set of camera intrinsic parameters, with or without noise in the uniform percentage simultaneously. Figure 11 evidently shows the reconstruction accuracy results for each test sample. Compared to the standard model, there is no obvious trend in adding disturbance. On some test samples, it is normal for the reconstruction accuracy result of adding noise to perform better. But on other test samples, the reconstruction accuracy with disturbance added even perform better. This discrepancy could be attributed to that the standard camera parameters actually are not the strictly precise values, but rather we regard them as standard values. When the intrinsic parameters noise are added 5%, the reconstruction accuracy of the test samples is almost the largest. These results show that when the disturbance gradually increases, the results will be greatly affected. But in a certain disturbance range, the proposed methods can optimize the error introduced by the camera intrinsic parameters.

When the noise percentage is lower 3%, the reconstruction accuracy is in a relatively small range; 1% of the noise of the reconstruction accuracy is even lower than the standard value. However, when the intrinsic noise reaches 4% and 5%, their reconstruction accuracy increases by 1.7 mm and 3.4 mm respectively. Although the absolute error is not remarkable, the percentage of error increases by 22% and 37.8% respectively. Considering the accuracy requirements of vision tasks, it seems that we need to pay attention to the camera intrinsic parameters error when using the proposed method.

V. CONCLUDING REMARKS

The proposed hand-eye calibration method is demonstrated as an effective approach for robotic pick-and-place tasks. Compared with the classic hand-eye calibration that uses a checkerboard to assist in calibrating extrinsic, our method only uses an available object in the task scene. In addition, the proposed method is robust to measurement errors. Through experimental observation, we found that the reconstruction accuracy of 15 mm is sufficient for our pick-and-place tasks. Practically, the maximum and the average reconstruction accuracy error are 11.3 mm and 7.91 mm respectively. For each picking try, the end-effector of robotic arm moves to the appropriate position, ensuring the finger's successful gripping of the object. From the engineering point of view, our method is not only easy to access and implement in practice, but also is effective for our tasks.

Actually, our method is not limited to the extrinsic calibration and optimization of the hand-eye system; it can be used to optimize the extrinsic calibration between two sensors as well. For the following works, we would try to use this method to calibrate the extrinsic parameters between cameras and IMU.We intend to use the drone's onboard camera to visually locate the ground target and apply the method to the camera and IMU calibration.

ACKNOWLEDGMENT

The authors would like to thank Tsinghua University that provided the experimental platform for the competition. They would also like to thank those who provided guidance and revisions to this article. *(Yong Zhou and Qiang Fang contributed equally to this work.)*

REFERENCES

[1] E. Klingbeil, D. Rao, B. Carpenter, V. Ganapathi, A. Y. Ng, and O. Khatib, "Grasping with application to an autonomous checkout robot," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 2837–2844.

[2] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[3] A. Tabb and K. M. A. Yousef, "Solving the robot-world hand-eye (s) calibration problem with iterative methods," *Mach. Vis. Appl.*, vol. 28, nos. 5–6, pp. 569–590, 2017.

[4] S. Remy, M. Dhome, J. M. Lavest, and N. Daucher, "Hand-eye calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, vol. 2, Sep. 1997, pp. 1057–1065.

[5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[6] A. Zeng, K. T. Yu, S. Song, D. Suo, E. Walker, and A. Rodriguez, "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge," presented at the IEEE Int. Conf. Robot. Automat. (ICRA), Marina Bay Sands, Singapore, 2017.

[7] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," *IEEE Trans. Robot. Autom.*, vol. 5, no. 3, pp. 345–358, Jun. 1989.

[8] F. C. Park and B. J. Martin, "Robot sensor calibration: Solving AX=XB on the Euclidean group," *IEEE Trans. Robot. Autom.*, vol. 10, no. 5, pp. 717–721, Oct. 1994.

[9] K. Daniilidis, "Hand-eye calibration using dual quaternions," *Int. J. Robot. Res.*, vol. 18, no. 3, pp. 286–298, Mar. 1999.

[10] R. Horaud and F. Dornaika, "Hand-eye calibration," *Int. J. Robot. Res.*, vol. 14, no. 3, pp. 195–210, 1995.

[11] A. Malti, "Hand–eye calibration with epipolar constraints: Application to endoscopy," *Robot. Auto. Syst.*, vol. 61, no. 2, pp. 161–169, 2013.

[12] Y. S. Hung and W. K. Tang, "Projective reconstruction from multiple views with minimization of 2D reprojection error," *Int. J. Compute Vis.*, vol. 66, no. 3, pp. 305–317, 2006.

[13] M. Shah, "Solving the robot-world/hand-eye calibration problem using the Kronecker product," *J. Mech. Robot.*, vol. 5, no. 3, pp. 031007-1–031007-7, 2013.

[14] A. Tabb and K. M. A. Yousef, "Parameterizations for reducing camera reprojection error for robot-world hand-eye calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep./Oct. 2015, pp. 3030–3037.

[15] K. H. Strobl and G. Hirzinger, "Optimal hand-eye calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2007, pp. 4647–4653.

[16] R. K. R. Ganesh Iyer, J. Krishna Murthy, and K. Madhava Krishna, "CalibNet: Geometrically supervised extrinsic calibration using 3D spatial transformer networks," presented at the IEEE/RSJ Int. Conf. Intell. Robots Systemsa (IROS), Madrid, Spain, 2018.

[17] T. A. Oliver Limoyo, F. Maric, L. Volpatti, and J. Kelly, "Self-calibration of mobile manipulator kinematic and sensor extrinsic parameters through contact-based interaction," presented at the IEEE Int. Conf. Robot. Automat. (ICRA), Brisbane, QLD, Australia, 2018.

[18] K. Pachtrachai, F. Vasconcelos, G. Dwyer, V. Pawar, S. Hailes, and D. Stoyanov, "CHESS—Calibrating the hand-eye matrix with screw constraints and synchronisation," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2000–2007, Jul. 2018.

[19] J. D. Hol, T. B. Schön, and F. Gustafsson, "Modeling and calibration of inertial and vision sensors," *Int. J. Robot. Res.*, vol. 29, nos. 2–3, pp. 231–244, 2010.

[20] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *Int. J. Robot. Res.*, vol. 30, no. 1, pp. 56–79, 2011.

[21] K. Huang and C. Stachniss, "On geometric models and their accuracy for extrinsic sensor calibration," presented at the IEEE Int. Conf. Robot. Automat. (ICRA), Brisbane, QLD, Australia, 2018.

[22] Y. Ling and S. Shen, "High-precision online markerless stereo extrinsic calibration," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2016, pp. 1771–1778.

[23] F. Furrer, M. Fehr, T. Novkovic, H. Sommer, I. Gilitschenski, and R. Siegwart, *Evaluation of Combined Time-Offset Estimation and Hand-Eye Calibration on Robotic Datasets*. Cham, Switzerland: Springer, 2018, pp. 145–159.

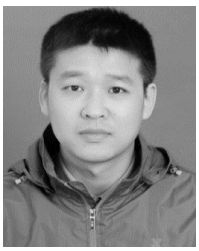
[24] Z. Zhang, *A Flexible New Technique for Camera Calibration*. Washington, DC, USA: IEEE Computer Society, 2000.

[25] J. C. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, 1998.

[26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[27] K. Zhao, Y. Zhou, Z. Zhou, D. Tang, Y. Chang, Q. Fang, H. Zhou, and T. Hu, "Onboard smart surveillance for micro-UAV swarm: An experimental study," presented at the 8th Annu. IEEE Int. Conf. Cyber Technol. Automat., Control, Intell. Syst., Tianjin, China, 2018.

YONG ZHOU was born in Shaoyang, China, in 1995. He received the B.S. degree in automation from the National University of Defense Technology (NUDT), Changsha, China, in 2017, where he is currently pursuing the master's degree with the College of Intelligence Science and Engineering. His research interests include machine vision and unmanned aerial vehicles, focusing on sensor calibration, autonomous navigation and localization, and pose estimation.



QIANG FANG received the B.Eng. degree in automation from Xidian University, in 2007, and the M.S. and Ph.D. degrees in control science and engineering from the National University of Defense Technology, in 2009 and 2013, respectively, where he is currently a Lecturer with the College of Intelligence Science and Engineering. His research interests include the areas of robotics and unmanned aerial vehicles, with a focus on state estimation, autonomous navigation, and deep learning.

KUANG ZHAO received the bachelor's degree from the National University of Defense Technology (NUDT), China, in 2016, where he is currently pursuing the master's degree with the College of Intelligence Science and Engineering. His research interests include computer vision, machine learning, and robotics.

DENGQING TANG received the B.S. and M.S. degrees in control science and engineering from the National University of Defense Technology, Changsha, China, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the National University of Defense Technology, where he focuses on the vision-based dynamic object pose estimation, visual SLAM, and object detection. His M.S. thesis was on the vision-based UAV localization.

HAN ZHOU received the Ph.D. degree in control science and engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2015.

She was with the University of Essex, U.K., from 2011 to 2012, as a Joint Ph.D. Candidate, financially supported by the Chinese Scholarship Council. Since 2015, she has been a Lecturer with NUDT. Her research interests include robotics, dynamics analysis, and learning control.

GUOQI LI received the B.Eng. degree from the Xi'an University of Technology, China, in 2004, the M.Eng. degree from Xi'an Jiaotong University, China, in 2007, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2011.

From 2011 to 2014, he was a Scientist with the Data Storage Institute and the Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore. Since 2014, he has been with the Department of Precision Instrument, Tsinghua University, Beijing, China, where he is currently an Associate Professor. He has authored or coauthored more than 80 journal and conference papers. His current research interests include brain-inspired computing, complex systems, machine learning, neuromorphic computing, and system identification.

Dr. Li serves as a Reviewer for a number of international journals. He has been actively involved in Professional Services, such as serving as an International Technical Program Committee Member and a Track Chair for international conferences. He is an Editorial Board Member and a Guest Associate Editor of the *Frontiers in Neuroscience* (Neuromorphic Engineering Section).

XIAOJIA XIANG received the B.E., M.S., and Ph.D. degrees in automatic control from the National University of Defense Technology (NUDT), Changsha, China, in 2002, 2007, and 2016, respectively. Since 2002, he has been with the College of Mechatronic Engineering and Automation, NUDT. He is currently an Associate Professor with the College of Artificial Intelligence, NUDT. His research interests include mission planning, autonomous, and cooperative control for unmanned systems.

TIANJIANG HU (S'07–M'10) received the B.Eng. and Ph.D. degrees in robotics and automatic control from the National University of Defense Technology, China, in 2002 and 2009, respectively. He is currently a Full Professor with Sun Yat-sen University, and has originated Machine Intelligence and Collective RObotics (MICRO) Lab. He has been a Visiting Scientist for international collaboration with Nanyang Technological University, Singapore, and the University of Manchester, U.K. His current research interests include autonomous systems, biologically inspired robotics, collective intelligence, and learning control.