
Ihtisham Ali 1,* Olli Suominen 1, Atanas Gotchev 1 and Emilio Ruiz Morales 2

1 Faculty of Information Technology and Communication, Tampere University, 33720 Tampere, Finland; olli.j.suominen@tuni.fi (O.S.); atanas.gotchev@tuni.fi (A.G.)
2 Fusion for Energy (F4E), ITER Delivery Department, Remote Handling Project Team, 08019 Barcelona, Spain; Emilio.Ruiz@f4e.europa.eu
* Correspondence: ihtisham.ali@tuni.fi; Tel.: +358-417-268-110

Received: 1 June 2019; Accepted: 21 June 2019; Published: 25 June 2019

Abstract: In this paper, we propose two novel methods for robot-world-hand–eye calibration and provide a comparative analysis against six state-of-the-art methods. We examine the calibration problem from two alternative geometrical interpretations, called ‘hand–eye’ and ‘robot-world-hand–eye’, respectively. The study analyses the effects of specifying the objective function as pose error or reprojection error minimization problem. We provide three real and three simulated datasets with rendered images as part of the study. In addition, we propose a robotic arm error modeling approach to be used along with the simulated datasets for generating a realistic response. The tests on simulated data are performed in both ideal cases and with pseudo-realistic robotic arm pose and visual noise. Our methods show significant improvement and robustness on many metrics in various scenarios compared to state-of-the-art methods.

Keywords: robot-world-hand–eye calibration; hand–eye calibration; optimization

1. Introduction

Hand–eye calibration is an essential component of vision-based robot control also known as visual servoing. Visual servoing effectively uses visual information from the camera as feedback to plan and control action and motion for various applications such as robotic grasping [1] and medical procedures [2]. All such applications require accurate hand–eye calibration primarily to complement the accurate robotic arm pose with the sensor-based measurement of the observed environment into a more complete set of information.

Hand–eye calibration requires accurate estimation of the homogenous transformation between the robot hand/end-effector and the optical frame of the camera affixed to the end effector. The problem can be formulated as \( AX = XB \), where \( A \) and \( B \) are the robotic arm and camera poses between two successive time frames, respectively, and \( X \) is the unknown transform between the robot hand (end effector) and the camera [3,4].

Alternatively, the estimation of a homogeneous transformation from the robot base to the calibration pattern/world coordinate system can be obtained as a byproduct of the problem solution widely known as robot-world-hand–eye (RWHE) calibration, formulated as \( AX = ZB \). In this formulation, we define \( X \) as the transformation from robot base to world/pattern coordinate and \( Z \) is the transformation from the tool center point (TCP) to the camera frame. These two notations might be opposite in some other studies. The transformations \( A \) and \( B \) no longer represent the relative motion poses between different time instants. Instead, they now represent the transformation from TCP to the robot base frame, and the transformation from the camera to the world frame.
A considerable number of studies have been carried out to solve the problem of hand–eye calibration. While the core problem has been well addressed, the need for improved accuracy and robustness has increased with time as the hand–eye calibration problem expands to find its uses in various fields of science.

The earliest approach presented for hand–eye calibration estimated the rotational and translational parts individually. Due to the nature of the approach, the solution is known as separable solution. Shiu and Ahmed [4] presented a closed-form approach to finding the solution for the problem formulation $AX = XB$ by separately estimating the rotation and translation from robot wrist to the camera in that order. The drawback of the approach presented was that the linear system doubles at each new entry of the image frame. Tsai [3] approached the problem from the same perspective, however, they improved the efficiency of the method by keeping the number of unknowns fixed irrespective of the number of images and robot poses. Moreover, the derivation is both simpler and computationally efficient compared to [4]. Zhuang [5] adopted the quaternion representation for solving the rotation transformation from hand to eye and robot base to the world. The translation components are then computed using linear least squares. Liang et al. [6] proposed a closed-form solution by linearly decomposing the relative poses. The implementation is relatively simple; however, the approach is not robust to noise in the measurements and suffers intensely in terms of accuracy.

Hirsh et al. [7] proposed a separable approach that solves for $X$ and $Z$ alternatingly in an iterative process. The approach makes an assumption that one of the unknown is pseudo-known for that time being and estimates the best possible values for the other unknown by distributing the error. In the first case, it assumes that $Z$ is known by the system and estimates $X$ by averaging over the equation $X = ZB_nA^{-1}$ for all $n$ poses of $B$. Similarly, an estimation for $Z$ is obtained by using the previously obtained $X$. This process continues until the system reaches the condition to terminate the iterative estimation. In a recent study, Shah [8] proposed a separable approach that forms its bases on the methods presented by Li et al. [9]. Shah suggests using the Kronecker product to solve the hand–eye calibration problem. The method first computes the rotational matrices for the unknown $X$, followed by computing the translation vectors. Kronecker product is an effective approach to estimate the optimal transformation in this problem. However, the resulting rotational matrices might not follow orthogonality. To compensate for this issue, the best approximations for orthonormal rotational matrices are obtained using Singular Value Decomposition (SVD). The primary difference between the work of [8] and [9] is that Li et al. do not update the positions that were only optimal for the rotational transformation before the orthonormal approximation. This augments to any errors that might already be present in the solution. In contrast, Shah [8] explicitly re-computes the translations based on the new orthonormal approximations of the rotations $R_X$ and $R_Z$. Earlier studies have shown that separable approaches have a core limitation, which results in slightly high position errors. Since the orientations and translation are computed independently and in the mentioned order, the errors from orientations step propagate to the position estimation step. Typically, separate solution based approaches have good orientation accuracy; however, the position accuracy is often compromised.

The second class of solutions is based on simultaneous computation of the orientation and position. Chen [10] argued that rotation and translation are interdependent quantities and, therefore, should not be estimated separately. He proposed a simultaneous approach to the hand–eye problem based on screw theory where both the rotation and translation components are computed altogether. In his work, Chen estimates a rigid transformation to align the camera screw axis to the robot screw axis. Dornaika and Horaud [11], proposed a nonlinear least square based approach to solve the hand–eye calibration problem. The optimization approach solved for an abundant number of parameters that represent rotations in the form of matrices. The cost function constrained the optimization to solve for orthonormal rotation matrices. It was observed that the nonlinear iterative approach yielded better results to linear and closed form solution in term of accuracy. Henceforth, many studies have opted for nonlinear cost minimization approach since they are more tolerant to nonlinearities present in measurements in the form of noise and errors. Shi et al. [12] proposed to replace the rotation
matrices with quaternion representation to facilitate the iterative optimization approach towards a solution. In [13], Wei et al. contributed an approach for an online hand–eye calibration approach that estimate the transformations through active motion. The method discards degenerative cases where no or little rotation cases induce high errors into the system. Strobel and Hirzinger [14], proposed an adaptive technique for hand–eye calibration using nonlinear optimization. The approach adjusts weights that are assigned to the rotation and translation errors during the cost minimization step. In [15], Fassi and Legnai construed a geometrical interpretation of the hand–eye calibration problem for the formulation $AX = XB$. They argued that the general formulation can lead to an infinite solution and therefore a constrained multi-equation based system is always suitable to optimize. Some cases that result in singularity were also discussed. Zhao [16] presents a convex cost function by employing the Kronecker product in both rotational matrix and quaternion form. The study argues that a global solution can be obtained using linear optimization without specifying any initial points. This serves as an advantage over using L2 based optimization. Heller et al. [17] proposed a solution to the hand–eye calibration problem using the branch-and-bound (BnB) method introduced in [18]. The authors minimize the cost function under the epipolar constraints and claim to yield a globally optimum solution with respect to $L_\infty$ norm. Tabb [19] tackled the problem of hand–eye calibration from the iterative optimization based approach and compared the performance of various objective functions. The study focused on $AX = ZB$ formulation and solved for the orientation and translation both separately and simultaneously using the nonlinear optimizer. Moreover, a variety of rotation representations was adopted including Euler, rotation matrix and quaternion in order to study their effect on accuracy. The study explored the possibility of a robust and accurate solution by minimizing pose and reprojection errors using different costs. The authors used the nonlinear optimizer Ceres [20] to solve for a solution using the Levenberg-Marquardt algorithm.

In this study, we present a collection of iterative methods for the hand–eye calibration problem under both $AX = XB$ and $AX = ZB$ formulations. We adopt the iterative cost minimization based approach similar to Tabb [19]. However, the geometrical formulation is reverted to the generic form for better coherence. Moreover, we study the problem from $AX = ZB$ formulation, which is not present in [19]. The prospects of a new cost functions for the non-linear regression step are also studied. Each method is quantified from pose optimization and reprojection error minimization perspective. The main contributions of this study are as follows:

1. We provide a comprehensive analysis and comparison of various cost functions for various problem formulations.
2. We provide a dataset composed of three simulated sequences and three real data sequence, which we believe is handful for testing and validation by the research community. To the best of our knowledge, this is the first simulated data set for hand–eye calibration with synthetic images that are available for public use. Moreover, the real data sequences include chess and ChArUco calibration board of varying sizes. The datasets are available from [21].
3. We provide extensive testing and validation results on a simulated dataset with realistic robot (position and orientation) and camera noise to allow comparisons between the estimated and true solutions more accurately.
4. We provide an open-source code of the implementation of this study along with the surveyed approaches to support reproducible research. The code is available from [21].

The article is organized as follows: In Section 2, we present in detail the problem formulations for robot-world-hand–eye calibration. In Section 3, we discuss the development of real and synthetic dataset for evaluation purpose. Section 4 presents the error metrics used to quantify the performance of the calibration methods. Section 5 summarizes the experimental results using both synthetic and real datasets against the aforementioned error metrics. Finally, Section 6 concludes the article.
2. Methods

For the needs of our study, we introduce notations, as illustrated in Figure 1. Throughout this article, we will represent homogenous transformations by $T$ supported with various sub-indexes. The sub-indexes $b$, $t$, $c$ and $w$ indicate the coordinate frames associated with robot base, robot tool, camera and the calibration pattern, respectively. The sub-indexes $i$ and $j$ are associated with time instants of the state of the system. For the first general formulation $AX = XB$ illustrated in Figure 1a, $b\ T^b$ is the equivalent to $A_i$ and denotes the homogenous transformations from robot base to the tool center point (TCP)/end-effector. $c\ T^w$ is the equivalent of $B_i$ and denotes the homogenous transformation from camera to the world/calibration pattern. The formulation uses the relative transformation $A (t_i T^b_i)$ and $B (t_c T^c_i)$ from their respective previous pose to another pose. The unknown $X$ or $t_T^c$ is the required homogenous transformation from the end effector to the camera.

![Figure 1. Formulations relating geometrical transformation for calibration; (a) hand–eye calibration; (b) robot-world-hand–eye Calibration.](image)

The second general formulation, $AX = ZB$ is illustrated in Figure 1b. The formulation uses absolute transformation $A(t_i T^b_i)$ and $B(t_c T^w_i)$ from their respective coordinate frames. The unknown $X(t_b T^w_i)$ and $Z(t_c T^c_i)$ are the homogenous transformations from robot base to the world frame and the end effector to the camera frame, respectively. The hand–eye transformation is referred to as $Z$ in this formulation for coherence in literature, since many studies opt for such notation.

In this section, we focus on various cost functions for the two general problem formulations with the aim to analyze their performance under real situations. For both cases, we consider solving the problem by minimizing pose error and reprojection error. Some studies including [19] propose to optimize the camera’s intrinsic parameters using the nonlinear solver to yield better results. However, Koide and Menegatti [22] argue that the approach involving camera intrinsic optimization overfits the model on the data for the reprojection error; consequently, the results are poor for other error metrics including reconstruction accuracy. Following the insight from [22], we solve for the transformation by minimizing the reprojection error.

The main information required for hand–eye calibration are the Tool Centre Point (TCP)/end effector poses and the camera poses. The TCP pose of the robotic arm is directly provided by the software of the robotic arm against the base of the arm. The pose is typically quite accurate due to the high accuracy of the encoders in the robotic arm that provide feedback for the angles of the joints. In general, for many robotic arms, the precision for the end effector’s position is around 0.1–0.2 mm. On the other hand, the camera pose against the world frame can be obtained through various methods. The common approach is to use a calibration pattern for simultaneously calculating the calibration parameters of the camera and the pose of the camera against the pattern or in this case world frame. Many researchers favor this approach since the calibration pattern is easy to acquire and its use yields good results. In contrast, some studies [23,24] prefer Structure from Motion (SFM) to acquire the relative camera transformation when the camera is moved from one point to another. The approach is independent of the calibration pattern and can acquire the correspondences from the feature-rich environment. However, SFM based camera calibration and camera pose computation.
are prone to errors. The approach inherits additional errors into the hand–eye calibration process and reduces the overall accuracy of the system. To compensate for these errors, the process must include additional steps to mitigate the effects. The added efforts deviate the focus from the core target, which is accurate hand–eye calibration. In this study, we utilize industrial-grade calibration boards in order to estimate the camera intrinsic parameters and camera extrinsics for the robot-world-hand–eye calibration problem. The camera calibration approach used in this study is based on the widely adopted method by Zhang [25].

2.1. Hand–Eye Formulation

This mathematical problem formulation involves estimating one unknown with the help of two known homogenous transformations in a single equation. Let $bT^i$ be the homogenous transformation from the base of the robot to the robot TCP. The homogenous transformation relating the camera coordinate frame to the world coordinate frame affixed to the calibration pattens is $c_iT^w$. The unknown homogenous transformation from the tool to the camera coordinate frame to be estimated is represented by $t_iT^c$. Then from Figure 1a, we can form the following relationship

$$bT_i^{c-1}bT^i t^c = c_2T^w c_1T^w c_1T_i^{c-1} \left\{ t_iT^c \right\} = bT^c$$

(1)

$$t_2bT^i t^c = c_2T^w w_iT^c$$

(2)

Generalizing Equation (2) gives us Equation (3).

$$t_iT^i t^c = t^c c^c_i$$

(3)

Equation (4) represents the direct geometrical relationship between various coordinate frames involved in the system. In order to attain a solution and achieve dependable results it is required that the data is recorded for at least 3 positions with non-parallel movements of the rotational axis [14]. We can directly minimize the relationship in Equation (4) to estimate the unknown parameters presented in Equations (5) and (6) as Xc1 and Xc2, respectively.

$$\{ q_i(q_{(\ell)}, t^c) \} = \arg\min_{q_i(q_{(\ell)}, t^c)} \sum_{i=1, j=i+1}^{n-1} \left\| t_iT^i \left[ q_i(q_{(\ell)}, t^c) \right]_{HT} - \left[ q_i(q_{(\ell)}, t^c) \right]_{HT} c^c_i T^c \right\|^2_2$$

(5)

In light of recommendation of [19], we can also re-arrange Equation (5) in the following manner.

$$\{ q_i(q_{(\ell)}, t^c) \} = \arg\min_{q_i(q_{(\ell)}, t^c)} \sum_{i=1, j=i+1}^{n-1} \left\| t_iT^i - \left[ q_i(q_{(\ell)}, t^c) \right]_{HT} c^c_i T^c \left[ q_i(q_{(\ell)}, t^c) \right]_{HT} \right\|^2_2$$

(6)

Here, the symbol $\left\| \right\|_{HT}$ denotes the conversion of the parameters to homogenous transformation representation. The solver optimizes the parameters in quaternion representation $q_i(q_{(\ell)})$ of the rotational matrix $R^{c}$ and translation $t^c$. The operation $\bar{n}$ denotes the aggregation of the $4 \times 4$ error matrix into a scalar value by summation of normalized values of quaternion angles and normalized translation vector. The terms $q_i(q_{(\ell)})$ and $t^c$ are the quaternion and translation vector obtained from the inverse of $t_iT^c$. The objective functions minimize the L2-norm of the residual scalar values. The solutions in Equations (5) and (6) belong to the simultaneous solution category of hand–eye calibration because the rotation and translation are solved at the same time. We use the Levenberg–Marquardt algorithm to search for a minimum in the search space. The objective function successfully converges to a solution without any initial estimates for the $q_i(q_{(\ell)})$ and $t^c$. We have observed that the cost function in
Equation (6) enjoys a slight improvement in some cases over Equation (5), which will be discussed in the experimental results and discussion section.

The second approach to seek a solution is based on reprojection-based methods. Reprojection error minimization has shown promising results for pose estimation in various problem cases [26,27]. Tabb [19] examined the reprojection-based method for the AX = XB formulation. We generalize this approach for the case of the AX = XB formulation. Let W be the 3D points in the world frame and \( P^c \) be the same points in the camera frame. In the case of the chessboard pattern, these points are the corners of the chessboard. The following relationship represents the objective function for minimizing the reprojection error of the 3D points from pose \( i \) to pose \( j \). The cost function in Equation (7) is referred to as RX here onwards.

\[
\left\{ q_{(i,c)}, t^c \right\} = \operatorname{argmin}_{\tilde{q}_{(i,c)}, t^c} \sum_{i=1, j=i+1}^{n-1} ||\tilde{P}_j ||^2 - \Pi_{K, \tilde{q}_{(i,c)}, t^c} \left[ \left[ \tilde{q}_{(i,c)}, t^c \right]_{HT} \right] \left[ \left[ \tilde{q}_{(i,c)}, t^c \right]_{HT} \right]^2 \right\}_2^2 (7)
\]

In the equation, \( \Pi \) represents the operation that projects the 3D points from world space to image space using the camera intrinsic \( K \) and the camera extrinsic obtained using the homogenous transformations given in Equation (7), while \( \tilde{P}_j \) are the observed 2D points in the \( j \)-th image.

It is important to note that the reprojection error minimization based approach is not invariant to the choice of initial estimates for the solver. However, if a good initial estimate is provided, the nonlinear optimization of reprojection error can provide a more accurate solution with a fine resolution.

2.2. Robot-World-Hand–Eye Formulation

This mathematical formulation involves the estimation of an additional homogenous transformation that is between the robot base frame and world frame. Therefore, we have two known and two unknown homogenous transformations. Let \( t^b \) be the homogenous transformation from robot TCP to the base of the robot. The homogenous transformation relating the camera coordinate frame to the world coordinate is \( c^w \). The additional unknown homogenous transformation from the robot base frame to the world frame is \( b^w \). Then from Figure 1b, we can form a straightforward geometrical relationship as:

\[
\begin{bmatrix}
    t^b \\
    1
\end{bmatrix} \begin{bmatrix}
    t^R^c & t^b  \\
    0_{1X3} & 1
\end{bmatrix} = \begin{bmatrix}
    b^R^w & b^t^w  \\
    0_{1X3} & 1
\end{bmatrix} \begin{bmatrix}
    t^R^c & t^c \\
    0_{1X3} & 1
\end{bmatrix} \begin{bmatrix}
    c^R^w & c^t^w \\
    0_{1X3} & 1
\end{bmatrix}
\]

Similar to the previous cases, we can directly use the relationship in aforementioned equations to obtain \( t^c \) and \( b^w \) using nonlinear minimization of their respective costs

\[
\left\{ q_{(i,c)}, t^c, q_{(b,w)}, b^w \right\} = \operatorname{argmin}_{q_{(i,c)}, t^c, q_{(b,w)}, b^w} \sum_{i=1}^{n-1} ||\Pi_{K, q_{(b,w)}, b^w} \left[ \left[ q_{(i,c)}, t^c \right]_{HT} \right] - \left[ q_{(i,c)}, t^c \right]_{HT} \left[ q_{(i,c)}, t^c \right]_{HT} \right|_2^2 \right\}_2^2 (10)
\]

We can observe from Equation (10), that we are attempting to solve for two unknown homogenous transformations. The adopted parametrization involves optimizing over 14 parameters, where the two quaternions and translation vectors contribute to 8 and 6 parameters, respectively. While the robot-world-hand–eye calibration involves more unknowns for estimation, nonetheless, it constrains the geometry with more anchor points and helps to converge closer to the global minimum. With the advent of modern nonlinear solvers, the problem of optimizing for a large number of unknowns has become simpler and more efficient. As before, the objective function in Equation (10) can be re-arranged in the form of Equation (11). The cost functions in Equations (10) and (11) are referred to as Zc1 and Zc2, respectively, in Tabb [19].

\[
\left\{ q_{(i,c)}, t^c, q_{(b,w)}, b^w \right\} = \operatorname{argmin}_{q_{(i,c)}, t^c, q_{(b,w)}, b^w} \sum_{i=1}^{n-1} ||\Pi_{K, q_{(b,w)}, b^w} \left[ \left[ q_{(i,c)}, t^c \right]_{HT} - \left[ q_{(i,c)}, t^c \right]_{HT} \right] - \left[ q_{(i,c)}, t^c \right]_{HT} \left[ q_{(i,c)}, t^c \right]_{HT} \right|_2^2 \right\}_2^2 (11)
\]
The initial estimates can be acquired from any fast closed-form method such as Tsai [3] or Shah [8].

This formulation can also be viewed as reprojection error minimization problem. The following equation presents a cost function that minimizes the reprojection of the 3D world points $W$ onto the image space in camera frame, where $\overline{P}_i$ are the observed 2D points in the $i$-th image. The cost functions in Equation (12) is referred to as $rp1$ in [19].

$$\left[ q_{(t,c)}, t^c, q_{(b,w)} , b^w \right] = \arg\min_{q_{(t,c)}, t^c, q_{(b,w)} , b^w} \sum_{i=1}^n \| \Pi \left( K, \left[ q_{(t,c)}, t^c \right] b^w \right) - I_i \|^2_{\Sigma}$$  \hspace{1cm} (12)

In contrast to the reprojection error cost function for problem formulation $AX = XB$ this formulation from [19] has the added advantage that it is not explicitly affected by the errors in pose estimation caused by blurred images or low camera resolution. If the camera intrinsic parameters are accurate enough, then the extrinsic can be indirectly computed as a transformation through $T^c$, $t^b$ and $b^w$ through the minimization of the objective function. On the contrary, the reprojection error cost function presented for problem formulation $AX = XB$ is more robust to robot pose errors given good images.

A marginal improvement in the results can be observed in various cases by using $\log(cosh(x))$ as the loss function. The relative improvement is discussed in detail in Section 5. $\log(cosh(x))$ approximates $\frac{x^2}{2}$ for small value of $x$ and $abs(x) - \log(2)$, for large values. This essentially means that $log(cosh(x))$ imitates the behavior of the mean squared error but is more robust to noise and outliers. Moreover, the function is twice differentiable everywhere and therefore does not deteriorate the convexity of the problem. The modified version is given as followed, where $E(x)$ is the error in terms of difference between the observed points and the reprojected points. The cost function in Equation (13) is referred to as RZ hereafter.

$$\left[ q_{(t,c)}, t^c, q_{(b,w)} , b^w \right] = \arg\min_{q_{(t,c)}, t^c, q_{(b,w)} , b^w} \sum_{i=1}^n \| \log(cosh\left( E(x) \right)) \|^2_{\Sigma}$$  \hspace{1cm} (13)

3. Performance Evaluation Using Datasets

In order to assess the performance of the robot-world-hand–eye calibration methods, we present multiple datasets to test the methods in laboratory and near field settings. These datasets contain data acquired using various combinations of camera, lens, calibration patterns and robot poses. A detailed description of datasets is provided in Table 1. The table also lists the length of each side of square of the calibration patterns, focal length of the lenses, and number of robot poses used to acquire images. The datasets include real data and simulated data with synthetic images. To the best of our knowledge, this study is the first to provide simulated robot-world-hand–eye calibration dataset with synthetic/generated images as open source for public use. A more detailed explanation of the datasets is presented in the following subsections.

<table>
<thead>
<tr>
<th>No.</th>
<th>Dataset</th>
<th>Data Type</th>
<th>Lens Focal Length [mm]</th>
<th>Square Size [mm]</th>
<th>Image Size</th>
<th>Robot</th>
<th>Poses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>kuka_1</td>
<td>Real</td>
<td>12</td>
<td>20</td>
<td>1928 x 1208</td>
<td>KR16L6-2</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>kuka_2</td>
<td>Real</td>
<td>16</td>
<td>15</td>
<td>1920 x 1200</td>
<td>KR16L6-2</td>
<td>28</td>
</tr>
<tr>
<td>3</td>
<td>kuka_3</td>
<td>Real</td>
<td>12</td>
<td>60</td>
<td>1928 x 1208</td>
<td>KR16L6-2</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>CS_synthetic_1</td>
<td>Simulated</td>
<td>18</td>
<td>200</td>
<td>1920 x 1080</td>
<td>N/A</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>CS_synthetic_2</td>
<td>Simulated</td>
<td>18</td>
<td>200</td>
<td>1920 x 1080</td>
<td>N/A</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>CS_synthetic_3</td>
<td>Simulated</td>
<td>18</td>
<td>200</td>
<td>1920 x 1080</td>
<td>N/A</td>
<td>30</td>
</tr>
</tbody>
</table>
3.1. Real Datasets

To acquire real data for this experiment, a KUKA KR16L6-2 serial 6-DOF robot arm was used with Basler acA1920-50gc camera using 12 mm and 16 mm lenses as shown in Figure 2. The primary aim in recording these datasets was to collect real data for various robot-world-hand–eye calibration tests. With this aim, the collection provides three real datasets with varying robot poses and calibration patterns as shown in Figure 3. In this study, we primarily use the chessboard pattern for accurate camera calibration and robot-world-hand–eye calibration. A minor yet significant difference between the datasets [28], used in [19], is that the robot hand/camera orientation changes are quite gentle. This is done to facilitate the OpenCV camera calibration implementation used in [19], therefore the aforementioned implementation is not invariant to significant orientation changes and as a result, it flips the origin of the calibration pattern. For our experiments, we utilized MATLAB’s implementation of [25], which can correctly detect the orientation of the pattern in any given pose. However, this neat trick requires that the calibration pattern is asymmetric in the number of rows and columns and that one of the sides has an even number of squares while the other side has odd. This requirement makes the datasets in [28], which have chessboard patterns with even number of rows and columns, unusable in our tests.

![Figure 2. An example of the setup for acquiring the datasets; (a) robotic arm moving in the workspace; (b) cameras and Lenses for data acquisition.](image)

![Figure 3. Example of captured images from the dataset 1 through 3; (a) checkerboard from dataset 1; (b) checkerboard from dataset 2; (c) ChArUco from dataset 3.](image)

In addition, the calibration board used in the third dataset is a ChArUco pattern with square size of 60 mm, shown in Figure 3c. ChArUco tries to combine the benefit of both chessboard and ArUco markers and tends to facilitate the calibration process by fast, robust and accurate corner detection even in occluded or partial views [29]. The ChArUco dataset is only provided as open source material for future testing and has not been utilized in this study.

3.2. Simulated Dataset with Synthetic Images

The real data has the advantage of encapsulating all the uncertainties of a real system; however, in such cases we do not have any ground truth information. It is not possible to acquire the ground truth TCP-to-camera transformation, since the camera frame lies inside the camera. While various metric
We aim at introducing a realistic amount of noise. The robot position repeatability is generally provided in the datasheets, which ranges from 0.1–0.3 mm for various robots. However, the orientation repeatability is absent since it cannot be measured for real robots at such a fine resolution. Here, we propose a reverse engineering approach to acquire a statistically valid amount of orientation noise. The position and orientation error of the TCP arises from the accumulated errors of the individual joints of the robotic arm due to robot flexibility and backlash. Using inverse kinematics, we can find the joint angles for any position of TCP within its workspace.

Once the joint angles are available, we can introduce noise into the individual joints through trial and error until it produces the end-effector position error comparable to the realistic error. Through forward kinematics, we can then estimate the position and orientation of the end-effector under various arm configurations. Figure 5 shows the operation flow for computing the error range of the new orientations.

For our test, we used the position error of the KUKA KR16L6-2 computed through highly accurate laser sensor. The mean of the errors in X, Y and Z axes were 0.06 mm, –0.05 mm and –0.04 mm, while the standard deviation of the errors were 0.22 mm, 0.18 mm and 0.17 mm. A normally distributed error for each axis is generated based on these values and introduced to the system to estimate the corresponding effects in the orientation of the TCP. The range of realistic valid error for the TCP position is shown in Figure 6a, while the output of the orientation error using the aforementioned framework is shown in Figure 6b.
The second error metric focuses on computing the translation errors. As above, the translation errors emerge from the same Equations (4) and (9).

\[
e_{tr} = \frac{1}{n} \sum_{i=1}^{n-1} \| ( t_j^R b^t^c ) + t_j^R t_i^c - ( t_j^R c_j^t^c ) + t_i^c \|_2^2
\]

\[
e_{tr} = \frac{1}{n} \sum_{i=1}^{n} \| ( t_j^R b^t^w ) + t_j^b - ( t_j^R c_j^t^w ) + t_i^c \|_2^2
\]
The third metric to measure the quality of the results is the reprojection root mean squared error (\( \text{rrmse} \)). The \( \text{rrmse} \) is measured in pixels and is a good metric to measure the quality of the results in the absence of ground truth information. The \( \text{rrmse} \) provides an added advantage that it back-projects the 3D points from the calibration board onto the images by first transforming them through the robotic arm. In case, if the hand eye calibration is not correct, the reprojection errors will be large. The \( \text{rrmse} \) for both the formulations are given in Equations (19) and (20).

\[
\text{rrmse} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left\| \vec{P}_j - \Pi \left( K, \left[ \tilde{q}(t_c), t_c \right]_H, t_i \tilde{T}_i^b \left[ \tilde{q}(t_c), t_c \right]_H, P_c \right) \right\|_2^2} \tag{19}
\]

\[
\text{rrmse} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left\| \vec{P}_i - \Pi \left( K, \left[ \tilde{q}(t_c), t_c \right]_H, t_i^b \left[ \tilde{q}(b,W), t^w \right]_H, W \right) \right\|_2^2} \tag{20}
\]

For the case of simulated data, we have accurate ground truth pose from the robot TCP to the camera. We can effectively utilize that information to acquire the absolute rotation error and absolute translation errors. The absolute rotation error can be obtained using Equation (21), while the absolute translation error is given using Equation (22). Here, \( t_R^{c_{gt}} \) and \( t_t^{c_{gt}} \) are the ground truth values.

\[
e_{\text{aR}} = \left\| \text{angle}(t_R^{c^{-1}} t_R^{c_{gt}}) \right\|_2^2 \tag{21}
\]

\[
e_{\text{at}} = \left\| t_t^{c_{gt}} - t_t^c \right\|_2^2 \tag{22}
\]

5. Experimental Results and Discussion

In this section, we report the experimental results for various cases and discuss the obtained results. We tabulate the results obtained for these cases using our own and six other studies to provide a clear comparison. Tables 1–4 summarize the results using the error metrics described in Section 4, over the datasets presented in Section 3. To elaborate on the naming, Xc1, Xc2, RX, and RZ refer to the optimization of the cost function based on Equations (5)–(7) and (13), respectively. In addition, Figure 7 illustrates the results from simulated data in dataset 5 over varying visual noise in the presence of the pseudo-realistic robotic arm pose noise. Tables 2 and 3 shows the evaluation of the aforementioned methods on datasets 1 and 2, respectively. Both datasets vary in the use of camera lenses and robot arm poses. It can be observed that the method by Shah [8] provides a better distribution of the rotational error and hence has the lowest relative rotation error (\( \varepsilon_{\text{R}} \)) values, while the method by Li et al. [9] yields a comparable result. The lowest relative translation error (\( \varepsilon_{\text{t}} \)) varies for both datasets and is yielded by the proposed method Xc2 and Park and Martin [32]. However, for dataset 2, it seems that Xc2 has not converged properly and has obtained a local minimum. On the other hand, the method by Park and Martin [32], still yields a relatively low \( \varepsilon_{\text{t}} \). Moreover, for both datasets 1 and 2, the method by Horaud and Dornaika [11] provides comparable results to Park and Martin [32].

For the reprojection root mean squared error \( \text{rrmse} \), the best results are obtained using the proposed RX approach for both tests. This is aided by the fact that the recorded datasets do not have large visual errors and as a result, RX performs comparably better. Moreover, since the cost function has only one unknown transformation to minimize for, the optimizer distributes the errors more evenly for the reprojection based cost function. Other reprojection based approaches namely Tabb’s rp1 [19] and RZ achieve quite close results to RX. It is noteworthy, that in spite of being a closed-form approach, Shah [8] obtains quite good \( \text{rrmse} \) that are at a competitive level to the reprojection errors based approaches.
We further study the performance of the methods using our simulated datasets. The primary difference between dataset 4 and 6 is the number and complexity of the unique camera poses for image acquisition. During experimentation, we observed that the resolution of the accuracy slightly improved with the increased number of images acquired from significantly different poses. However, none of the methods suffered significantly from comparably less information in dataset 4, therefore, we consider datasets 5 and 6 for extensive quantitative comparison of the methods. In addition to the previous tabulated results, Tables 4 and 5 provide experimental results on simulated data with synthetic images from dataset 6. The main difference between the two tests is that the first test (Table 4) considers ideal simulated data, while the second test (Table 5) has visual and robot pose noise induced. The robot pose noise is derived from the process explained in Section 3.3, while the visual noise is selected such that the resultant reprojection error amounts to the reprojection errors of real data tests.
Tables 4 and 5, present two absolute errors due to the presence of ground truth information for the simulated cases. It can be observed that Tabb’s rp1 [19] achieves the least $e_{Rr}$, $e_{rt}$, $e_{rmse}$ and $e_{df}$. Xc2 yields minimum Absolute Rotation Error ($e_{dR}$). For this dataset, the method by Park and Martin [32], failed to find a solution as it suffered from singularity. It is important to note for an ensued comparison that the proposed method RZ yields the second best results over most of the error metrics with minor differences from the least errors. This is important in a sense that all the errors are equally distributed and restricted close to their minimum values.

The backend experiments for the results in Table 5 use the same methods, metrics and dataset, as for Table 4. In agreement with the results of real data, Shah [8] yields the least $e_{dR}$ for this dataset as well. In addition to a validation on the performance of Shah [8], this indicates that a realistic amount
of orientation noise is present in the system for the method to emanate similar response. The proposed method \textit{RZ} yields the minimum $\epsilon_{r1}, \epsilon_{\text{rrmse}}$ and, $\epsilon_{at}$ and the second best result for $\epsilon_{aR}$. Tabb Zc1\cite{19} obtains the minimum $\epsilon_{aR}$.

**Table 5.** Comparison of methods using the described error metrics for dataset 6 with robot pose and visual noise.

<table>
<thead>
<tr>
<th>Method</th>
<th>Evaluation Form</th>
<th>Relative Rotation ($\epsilon_{r1}$)</th>
<th>Relative Translation ($\epsilon_{at}$)</th>
<th>Reprojection Error ($\epsilon_{\text{rrmse}}$)</th>
<th>Absolute Rotation Error ($\epsilon_{aR}$)</th>
<th>Absolute Translation Error ($\epsilon_{at}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsai [3]</td>
<td>AXXB</td>
<td>34.925</td>
<td>2476.4</td>
<td>99190</td>
<td>28.04</td>
<td>747.48</td>
</tr>
<tr>
<td>Horaud and Dornaika [11]</td>
<td>AXXB</td>
<td>1.723</td>
<td>199.92</td>
<td>18.764</td>
<td>0.43124</td>
<td>47.913</td>
</tr>
<tr>
<td>Park and Martin [32]</td>
<td>AXXB</td>
<td>1.7208</td>
<td>199.98</td>
<td>18.916</td>
<td>0.43819</td>
<td>47.733</td>
</tr>
<tr>
<td>Li et al. [9]</td>
<td>AXZB</td>
<td>1.177</td>
<td>80.061</td>
<td>7.8757</td>
<td>0.0029485</td>
<td>23.272</td>
</tr>
<tr>
<td>Shah [8]</td>
<td>AXZB</td>
<td>1.1767</td>
<td>58.352</td>
<td>8.5123</td>
<td>0.51765</td>
<td>8.3389</td>
</tr>
<tr>
<td>Xc1</td>
<td>AXXB</td>
<td>1.7932</td>
<td>192.86</td>
<td>17.442</td>
<td>0.12827</td>
<td>37.068</td>
</tr>
<tr>
<td>Xc2</td>
<td>AXXB</td>
<td>1.8026</td>
<td>193.22</td>
<td>19.031</td>
<td>0.20831</td>
<td>40.4</td>
</tr>
<tr>
<td>Tabb Zc1 [19]</td>
<td>AXXB</td>
<td>1.7989</td>
<td>206.01</td>
<td>13.445</td>
<td>0.0042528</td>
<td>11.368</td>
</tr>
<tr>
<td>Tabb Zc2 [19]</td>
<td>AXZB</td>
<td>1.2571</td>
<td>86.444</td>
<td>13.991</td>
<td>0.0501026</td>
<td>27.247</td>
</tr>
<tr>
<td>RX</td>
<td>AXXB</td>
<td>1.8087</td>
<td>204.06</td>
<td>12.534</td>
<td>0.027714</td>
<td>7.0139</td>
</tr>
<tr>
<td>Tabb rp1 [19]</td>
<td>AXZB</td>
<td>1.2093</td>
<td>44.982</td>
<td>1.5463</td>
<td>0.0073401</td>
<td>0.95304</td>
</tr>
<tr>
<td>RX</td>
<td>AXZB</td>
<td>1.2079</td>
<td>44.932</td>
<td>1.546</td>
<td>0.0069577</td>
<td>0.95845</td>
</tr>
</tbody>
</table>

This comparison demonstrates that the proposed \textit{RZ} is more robust to outliers present in the data and performs marginally better compared to Tabb’s rp1 \[19\] in the presence of noise.

Figure 7 shows the evaluation results for dataset 5 composed of simulated data. As before, the dataset is injected pseudo-realistic robotic arm pose noise and tested over varying realistic range of visual noise. The plots represent the averaged results over 1000 iterations in order to achieve a stable response. The 95% confidence interval from all the iterations for each experimentation point is also shown in Figure 7. It can be observed that the confidence intervals are quite narrow with the exception of the response of Tsai \[3\] over reprojection error metric. The narrow range of confidence interval indicates that we are 95% sure that our true mean lies somewhere within that narrow interval. Moreover, this implies that the noise introduced during different iterations is consistent in behavior and emulates a coherent response from the methods. The plot curves for each method pass through the mean values at each experimentation point. The results show that Tabb rp1 \[19\] and the proposed \textit{RZ} are quite robust to the increments in visual noise compared to other methods over all error metrics. Moreover, at high visual noise \textit{RZ} shows a slight improvement over Tabb rp1 \[19\]. It is noteworthy that despite the increase in relative rotation, translation and reprojection error, the absolute rotation and translation errors stay much more the same for Tabb rp1 \[19\] and \textit{RZ}. Tsai \[3\] performs poorly and erratically in the presence of noise in data. In the absence of visual noise Tabb’s Zc1 \[19\], Xc1, RX and Shah \[8\] can achieve lower errors compared to Tabb rp1 \[19\] and \textit{RZ}. Tsai \[3\] performs poorly and erratically in the presence of noise in data. In the absence of visual noise Tabb’s Zc1 \[19\], Xc1, RX and Shah \[8\] can achieve lower errors compared to Tabb rp1 \[19\] and \textit{RZ} for multiple metrics. However, real data always contains some magnitude of visual noise due to various reasons. The presence of visual noise may affect each method differently depending on the approach used. Nonetheless, the nonlinear reprojection based methods of the formulation $AX = ZB$ show good estimation results under visual noise and hand pose noise.

### 6. Conclusions

This study has examined the robot-world-hand–eye calibration problem in its two alternative geometrical interpretations, and has proposed a collection of novel methods. It benefits from non-linear optimizers that iteratively minimize the cost function and determine the transformations. We have conducted a comparative study to quantify the performances of optimizing over pose errors and reprojection errors. The code for the presented methods is provided as open-source for further use. Our collection of methods was evaluated with respect to state-of-the-art methods. The study also contributes new datasets for testing and validation purposes. These include subsets of three real data and three simulated data with synthetic images. Simulated data are beneficial as they provide ground truth. We have proposed a noise modeling approach to generate realistic robot TCP orientation noise to study the robustness of methods under realistic conditions. We showed that our methods perform
well under different testing conditions. RX yields good results with high accuracy under realistic visual noise with respect to reprojection error. In addition, RZ is more robust to visual noise and yields more consistent results for a greater range of visual noise.

**Author Contributions:** Conceptualization, I.A. and O.S.; Methodology, I.A.; Software, I.A.; Experiments and Validation, I.A.; Writing—Original Draft Preparation, I.A.; Writing—Review & Editing, E.R.M. and A.G.; Supervision, O.S. and A.G.; Funding Body Member, E.R.M.

**Funding:** The work presented in this paper was funded by Fusion for Energy (F4E) and Tampere University under the F4E grant contract F4E-GRT-0901. The results are intended to be integrated in advanced camera-based systems attached to manipulator arms in order to achieve complex Remote Handling maintenance operations in a safe and efficient way.

**Conflicts of Interest:** The authors declare no conflict of interest. This article reflects the views of the authors. F4E and TUNI cannot be held responsible for any use which may be made of the information contained herein.

**References**


