CaRaCTO: Robust Camera-Radar Extrinsic Calibration with Triple Constraint Optimization

Mahdi Chamseddine[1,2][a], Jason Rambach[1][b] and Didier Stricker[1,2][c]
[1]*German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany*
[2]*RPTU Kaiserslautern, Kaiserslautern, Germany*
{firstname.lastname}@dfki.de

Keywords: Calibration, Camera, Radar, Robotics.

Abstract: The use of cameras and radar sensors is well established in various automation and surveillance tasks. The multimodal nature of the data captured by those two sensors allows for a myriad of applications where one covers for the shortcomings of the other. While cameras can capture high resolution color data, radar can capture the depth and velocity of targets. Calibration is a necessary step before applying fusion algorithms to the data. In this work, a robust extrinsic calibration algorithm is developed for camera-radar setups. The standard geometric constraints used in calibration are extended with elevation constraints to improve the optimization. Furthermore, the method does not rely on any external measurements beyond the camera and radar data, and does not require complex targets unlike existing work. The calibration is done in 3D thus allowing for the estimation of the elevation information that is lost when using 2D radar. The results are evaluated against a sub-millimeter ground truth system and show superior results to existing more complex algorithms. https://github.com/mahdichamseddine/CaRaCTO.

1 INTRODUCTION

Environment sensing is an integral task in many modern applications. Whether it is for robotics, surveillance (Roy et al., 2009; Roy et al., 2011), autonomous or assistive driving (Cho et al., 2014; Chavez-Garcia and Aycard, 2015), sensors such as camera, radar, and lidar are used to detect and classify objects and obstacles in the respective environments. The sensors used have different characteristics which make them complementary rather than redundant. Cameras provide high resolution color, texture, as well as context information whereas lidar and radar provide depth and dimensions. While lidar data is of a higher spatial density than radar data, the latter is more robust to weather and lighting conditions and can measure velocities.

Data from the different sensors is usually fused together to get a better understanding of the state of the environment. Using the fused data, it is possible to detect the different objects and obstacles using multimodal features like dimensions and position, velocity and orientation, etc. (Sugimoto et al., 2004; Wang et al., 2011; Wang et al., 2014; Kim and Jeon, 2014). The tasks of sensor fusion however are preceded by a necessary calibration that aligns the data from all sensors in a common reference frame so that data association is done correctly. This makes calibration an essential step in any data processing problem.

Lidar sensors still suffer from high retail prices, and as such have not seen as much commercial adoption as cameras and radar sensors that have been around for a much longer time. And even though high resolution 3D radar sensors are starting to gain popularity (Stateczny et al., 2019; Wise et al., 2021), the 2D radar sensors are still the most widely used type of radar in commercial applications. Therefore, the calibration approach presented in this work is targeting a 2D radar and camera setup due to low price and wide adoption.

In this work, an extrinsic calibration algorithm for camera-radar systems is presented. Unlike other approaches which project the radar data to 2D, in this approach, the elevation is not disregarded but rather estimated with the help of the camera to realize 3D reconstruction of targets. The work also aims to stabilize the optimization problem against bad initialization and simplify the calibration setup to make it more

[a] https://orcid.org/0000-0003-4119-457X
[b] https://orcid.org/0000-0001-8122-6789
[c] https://orcid.org/0000-0002-5708-6023

Chamseddine, M., Rambach, J. and Stricker, D.
CaRaCTO: Robust Camera-Radar Extrinsic Calibration with Triple Constraint Optimization.
DOI: 10.5220/0012369700003654

In *Proceedings of the 13th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2024)*, pages 534-545
ISBN: 978-989-758-684-2; ISSN: 2184-4313

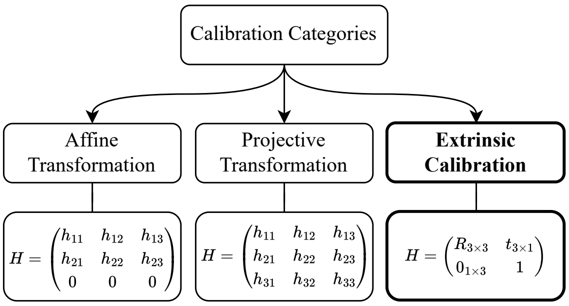

Figure 1: The different calibration categories as presented by Oh et al. (Oh et al., 2018). Our approach belongs to the third category.

accessible while maintaining the same quality or improving upon existing algorithms.

The main contributions can be summarized as:

- A camera-radar 3D calibration approach that does not require external sensing.
- Improved optimization formulation with additional elevation constraints for improved stability.
- Evaluation against state-of-the-art approaches using high accuracy optical measurements as ground truth and showing significant improvement.

The rest of the paper is structured as follows: Section 2 discusses the related work and previous contributions to the field. In Section 3 the problem is defined, our used notation is explained and the system model is described. The proposed method is presented in Section 4, and then evaluated in Section 5. Finally, concluding remarks are given in Section 6.

2 RELATED WORK

Several works on camera-radar calibration have been published. A comparative study by Oh et al. (Oh et al., 2018) differentiates between three different categories of camera-radar calibration (see Figure 1): affine transformations, projective transformations, and extrinsic calibration which is the target of our work.

In the work by Wang et al. (Wang et al., 2011) and Kim et al. (Kim and Jeon, 2014), an affine transformation is calculated between the 2D radar points and their corresponding pixel locations in the image. Pseudo inverse is used to solve a least squares setup of the two dimensional affine transformation. The quality of the transformation is measured using the image distance of the transformed radar points relative to their corresponding image points.

Whereas the 2D affine transformation calibration estimates six out of nine transformation parameters, the 2D projective transformation method estimates the complete 3×3 homography between the radar and camera planes. Sugimoto et al. (Sugimoto et al., 2004) and then Wang et al. (Wang et al., 2014) use the projective transformation method for camera-radar calibration. Even though this method can provide more accurate results for the calibration than affine transforms, the calibration disregards the 3D representation of the data and only provides point correspondences from the radar plane to the camera plane.

The third calibration category is extrinsic calibration, and the literature shows two different types: multi-sensor extrinsic calibration combining camera, lidar, and radar, and camera-radar only extrinsic calibration.

Peršiś et al. (Peršić et al., 2019) developed a radar-lidar-camera calibration method where the 3D lidar information is used to transform the radar frame in 2D and neglecting the elevation. The optimization is then performed in the 2D radar plane to estimate a transformation (rotation and translation) between the different sensors. Domhof et al. (Domhof et al., 2019) treat the radar data similarly and euclidean error in 2D is used for solving the optimization and calculating the extrinsic parameters. They additionally designed a complex joint target for camera, lidar, and radar.

Using a 3D radar sensor, Wise et al. (Wise et al., 2021) develop a continuous extrinsic calibration method that makes use of the extra dimension measurement as well the radar velocity measurement to develop a calibration algorithm that does not require radar retroreflectors. While their method properly takes into consideration the 3D nature of the problem and the complexity of using specialized targets, it is limited in application to less widely used 3D radars.

Unlike other camera-radar calibration methods, El Natour et al. (El Natour et al., 2015a) formulates the problem with the underlying assumption that the 2D representations in the image and radar data correspond to targets in 3D, thus the optimization is done using the 3D form and using the distance between multiple targets to recover the full 3D representation from 2D sensors. This allows for 3D reconstruction of targets after the system is calibrated. However, to achieve this result, multiple targets need to be present and the distance between them measured accurately. The authors try to overcome this limitation in (El Natour et al., 2015b) by moving the sensor system while keeping the targets fixed and adding the sensor trajectory estimation.

In our work, an extrinsic calibration algorithm is presented to estimate the rotation and translation between the camera and radar sensors and using the 3D representation of the targets. In contrast to prior work, only a single retroreflector is used without the need for a complex target design. Furthermore, the algorithm is made stable even with sub-optimal initialization through additional elevation constraints and the results are verified against other works using high quality ground truth system for the first time in such evaluations.

3 PROBLEM DEFINITION

Extrinsic calibration is the task of calculating the transformation (rotation and translation), between the coordinate systems of different sensors. The transformation can then be used to project a point from one system to the other and also reconstruct the 3D position.

For radar calibration, a specific target is used to collect and reflect the received radar signal. Such target is called a retroreflector characterized by its ability to reflect radiation back at its source (i.e. radar) with minimal scattering. In this work, a corner reflector is used which has a pyramidal shape made of 3 right isosceles triangles joined at their vertex angle (see Figures 3 and 4).

3.1 Notation

Since the calibration algorithm estimating the extrinsic transformation between two different sensors, it is first necessary to define the notation used in the rest of the sections and for this the notation defined in (Rambach et al., 2021) is used.

Let $\boldsymbol{m}_a = [x_a, y_a, z_a]^\top$ be the point $\boldsymbol{m}$ in a coordinate system A. The rotation and translation to convert $\boldsymbol{m}$ from system A to system B are then defined as $\boldsymbol{R}_{ba}$ and $\boldsymbol{b}_a$ respectively such that $\boldsymbol{R}_{ba}$ represents the rotation from coordinate system A to B and $\boldsymbol{b}_a$ is the origin of system B represented in system A. The transformation and its inverse can then be written as

$$\begin{aligned} \boldsymbol{m}_b &= \boldsymbol{R}_{ba}(\boldsymbol{m}_a - \boldsymbol{b}_a), \\ \boldsymbol{m}_a &= \boldsymbol{R}_{ab}(\boldsymbol{m}_b - \boldsymbol{a}_b), \end{aligned} \tag{1}$$

where $\boldsymbol{R}_{ab} = \boldsymbol{R}_{ba}^{-1} = \boldsymbol{R}_{ba}^\top$ and $\boldsymbol{a}_b = -\boldsymbol{R}_{ba}\boldsymbol{b}_a$. Thus, $\boldsymbol{m}_b$ can be expressed as

$$\boldsymbol{m}_b = \boldsymbol{R}_{ba}\boldsymbol{m}_a + \boldsymbol{a}_b. \tag{2}$$

Finally, the homogeneous transformation $\boldsymbol{H}_{ba}$ from coordinate system A to B can then be expressed as

$$\boldsymbol{H}_{ba} = \begin{bmatrix} \boldsymbol{R}_{ba} & \boldsymbol{a}_b \\ \boldsymbol{0} & 1 \end{bmatrix}. \tag{3}$$

3.2 System Model

The sensor setup is a radar-camera system connected rigidly and separated by a short baseline much smaller than the distance to the measured target. To avoid confusion in the terminology, the coordinate systems will be referred to as C for camera and S for radar (sensor). It then follows that $\boldsymbol{R}_{cs}$ and $\boldsymbol{R}_{sc}$ are the rotations from the radar to the camera system and its inverse respectively. Similarly, $\boldsymbol{c}_s$ is the origin of the camera in the radar coordinate system $\boldsymbol{s}_c$ is the origin of the radar in the camera coordinate system.

The camera model used is the pinhole model shown in Figure 2a, and it is used to project a point $\boldsymbol{m}_c = [x_c, y_c, z_c]^\top$ in the camera coordinate system to $\boldsymbol{p} = [u, v, 1]^\top$ on the image plane.

$$\begin{aligned} z_c\boldsymbol{p} &= \boldsymbol{K}\boldsymbol{m}_c, \\ z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} &= \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix}, \end{aligned} \tag{4}$$

where $\boldsymbol{K}$ is the intrinsic parameters matrix of the camera and is be calculated using the standard method in (Zhang, 2000) and u and v are the pixel coordinates of the point in an image. The radar is a frequency-modulated continuous-wave (FMCW) radar that returns the range, azimuth, doppler velocity, as well as radar cross section (reflection amplitude). Only the range and azimuth (ρ, θ) are used for the calibration. It can be noticed that the radar does not provide any elevation information ϕ, with ϕ representing the angle with positive *z-axis*. The general representation of the point $\boldsymbol{m}_s = [x_s, y_s, z_s]^\top$ in the radar coordinate system is shown in Figure 2b and defined as

$$\begin{aligned} x_s &= \rho \sin\phi \cos\theta, \\ y_s &= \rho \sin\phi \sin\theta, \\ z_s &= \rho \cos\phi, \end{aligned} \tag{5}$$

since ϕ is usually unknown, output can only be interpreted in 2D in other approaches (Sugimoto et al., 2004; Wang et al., 2011; Wang et al., 2014; Kim and Jeon, 2014; Peršić et al., 2019; Domhof et al., 2019) and is assumed that $\phi = \pi/2$.

Overall, a radar point in the radar coordinate system can be represented in the camera coordinate system using

$$\begin{aligned} \boldsymbol{m}_c &= \boldsymbol{R}_{cs}\boldsymbol{m}_s + \boldsymbol{s}_c, \\ \text{or} \begin{bmatrix} \boldsymbol{m}_c \\ 1 \end{bmatrix} &= \boldsymbol{H}_{cs} \begin{bmatrix} \boldsymbol{m}_s \\ 1 \end{bmatrix}. \end{aligned} \tag{6}$$

Figure 2c shows the same target represeted in both camera and radar coordinate systems. Combining Equation (4) with Equation (6), a relationship de-

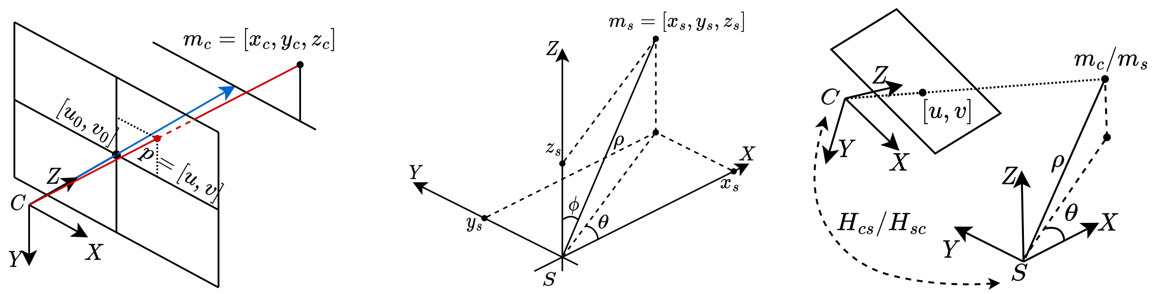

Figure 2: (a) and (b) show the measurement setup of both the camera and radar respectively. The pinhole model in (a) shows how an object in 3D can be represented in the camera coordinate system, and the pixel representation on the image plane. The radar data in (b) is measured in spherical coordinates, the diagram shows how they can be visualized as cartesian coordinates. (c) shows how an object visible in both the camera and radar frames can be represented in either frames and the transformation can be used to go from one representation to the other.

scribing a transformation between the radar and image data is then defined as follows

$$z_c \boldsymbol{p} = [\boldsymbol{K}|\boldsymbol{0}]\boldsymbol{H}_{cs} \begin{bmatrix} \boldsymbol{m}_s \\ 1 \end{bmatrix}, \tag{7}$$

where the $\boldsymbol{K}$ matrix is extended by a zero column to match the dimensions of the $\boldsymbol{H}$ matrix.

4 PROPOSED APPROACH

Our goal is to setup a system of equations to compute the residuals for the optimization. The residuals are minimized by finding the parameters of the extrinsic calibration. We first define the geometric relations that describe the measurements, we then formulate the optimization, and finally we reconstruct the 3D point cloud using the estimated calibration parameters.

4.1 Geometric Constraints

Using the measurement principles of the sensors used, different constraints and relationships can be observed. Given that the distance of a target to the radar is measured, the locus of the target can be restricted to a sphere of radius ρ centered at the radar

$$x_s^2 + y_s^2 + z_s^2 = \rho^2. \tag{8}$$

Knowing the azimuth angle of the target with respect to the radar positive *x-axis*, target also belongs to a plane passing through the radar center and perpendicular to the *xy-plane*. The normal vector to the plane is simply defined at the angle $(\theta + \pi/2)$. Thus the unit normal vector to the plane passing through the target point and the radar center is

$$\begin{aligned} \vec{\boldsymbol{n}} &= (\cos(\theta+\pi/2), \sin(\theta+\pi/2), 0) \\ &= (-\sin\theta, \cos\theta, 0), \\ \text{or } \vec{\boldsymbol{n}} &= (\sin\theta, -\cos\theta, 0). \end{aligned} \tag{9}$$

The locus of the target in the radar coordinate system can then be restricted to the intersection between the sphere defined in Equation (8) and the plane

$$x_s \sin\theta - y_s \cos\theta = 0, x_s > 0, \tag{10}$$

the condition that $x_s > 0$ means that the target should belong to the positive semi-circle in front of the radar.

The target also belongs to the line passing through the camera center and (u, v), the target's projection on the image plane. This line intersects the semi-circle from Equations (8) and (10) at one point representing the position of the target in 3D.

4.2 Optimization Formulation

Based on the constraints defined earlier, an optimization system is setup based on Equations (8) and (10) as follows

$$\begin{aligned} x_s^2 + y_s^2 + z_s^2 - \rho^2 &= \varepsilon_1, \\ x_s \sin\theta - y_s \cos\theta &= \varepsilon_2, \end{aligned} \tag{11}$$

where ε_1 and ε_2 are the residuals to be minimized as to ensure the 3D radar points satisfy the constraints. The position of the target in the image is then used to derive the representation of $m_s = [x_s, y_s, z_s]^\top$ in terms of (u, v) and $\boldsymbol{H}_{sc}$. From Equation (7) the following can be derived

$$\begin{aligned} \begin{bmatrix} \boldsymbol{m}_s \\ 1 \end{bmatrix} &= \boldsymbol{H}_{cs}^{-1} \begin{bmatrix} z_c \boldsymbol{K}^{-1} \boldsymbol{p} \\ 1 \end{bmatrix} \\ &= \boldsymbol{H}_{sc} \begin{bmatrix} z_c \boldsymbol{K}^{-1} \boldsymbol{p} \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{R}_{sc} & \boldsymbol{c}_s \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} z_c \boldsymbol{K}^{-1} \boldsymbol{p} \\ 1 \end{bmatrix}, \end{aligned} \tag{12}$$

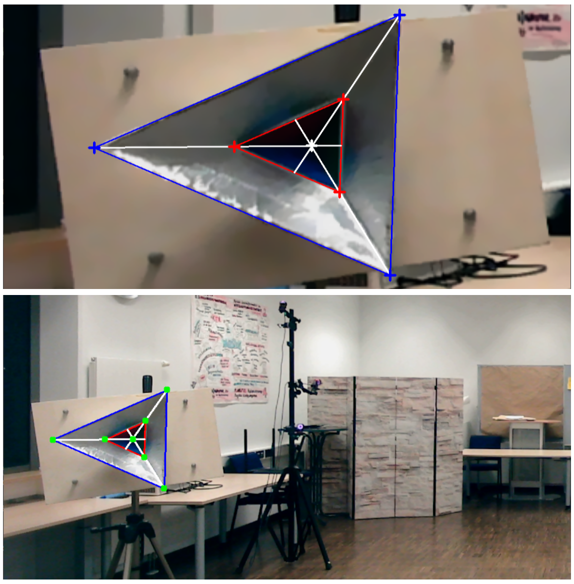

Figure 3: Top: Detection of corners of calibration target, a minimum of 4 points is needed to solve the PnP problem. Bottom: the reprojected solution of the PnP problem (green).

where $\boldsymbol{R}_{sc} = \boldsymbol{R}_\gamma \boldsymbol{R}_\beta \boldsymbol{R}_\alpha$ and α, β, and γ are the rotation angles around x, y, and z respectively. So the parameters to be estimated are the three rotation angles and the three translations represented by $\boldsymbol{c}_s = [x_{c_s}, y_{c_s}, z_{c_s}]^\top$.

The last unknown to be estimated in Equation (12) is z_c. Existing work (El Natour et al., 2015a) tackles the problem by using a minimum of six fixed targets and accurate measurement of the distances between them to estimate z_c, whereas (El Natour et al., 2015b) requires the ability to move the whole radar camera system to estimate z_c as part of the optimization. However, in this work we introduce two methods to estimate z_c with a single target measured at different positions thus simplifying the setup significantly.

4.2.1 Method 1: Using Radar Range as an Estimate for z_c

Since z_c is a measure of depth of a target with respect to the camera, and since the radar can directly detect a target's depth, it is logical to benefit from the multi-modal measuring capabilities of the sensor setup and use $z_c = \rho$. This assumption is limited to the case when the baseline between the camera and the radar is much smaller than the measured distance and when the camera and radar are close to each other.

4.2.2 Method 2: Using Camera Correspondences to Calculate z_c

This method overcomes the limitations of the first method and removes the short baseline requirement. Using the known dimensions of the radar retroreflector and intrinsic calibration matrix $\boldsymbol{K}$, it is possible to solve the perspective-n-point (PnP) problem to obtain the 6 *DoF* pose in the camera coordinate system (Lepetit et al., 2009). The euclidean distance to the center of the retroreflector is then used as z_c.

The target can be detected and matched to a labeled template to align the corners using the GMS Feature Matcher (Bian et al., 2017), the PnP problem is then solved on the aligned corners. It is worth noting that restricting the search area results in more reliable matching. Figure 3 shows the reprojection of the reflector corners. A small reprojection error indicates the correctness of the pose estimation.

4.3 Elevation Constraint

In addition to the residuals defined in Equation (11), an extra residual is added as a stabilizing term to the optimization and limit the pitch angle *beta* from deviating and speed up converging. Radar sensors are characterized with a relatively narrow vertical field of view ($\pm 15°$) and thus the data is distributed around the *xy-plane*. Based on those characteristics the stabilizing residual is defined

$$|z_s| = \varepsilon_3, \tag{13}$$

The system of optimization equations is solved using the Levenberg-Marquardt (LM) non-linear least squares optimization (Moré, 1978). The desired outcome is to find the set of parameters $[\alpha, \beta, \gamma, x_{c_s}, y_{c_s}, z_{c_s}]$ (rotation angles and translations) that minimizes the sum of squared residuals from Equations (11) and (13), $(\varepsilon_1)_i^2 + (\varepsilon_2)_i^2 + (\varepsilon_3)_i^2$, for each measured target i.

Overall we can formulate the objective function as

$$\underset{\alpha,\beta,\gamma,x_{c_s},y_{c_s},z_{c_s}}{\arg\min} \sum_{\forall i} (\varepsilon_1)_i^2 + (\varepsilon_2)_i^2 + (\varepsilon_3)_i^2. \tag{14}$$

4.4 Point Cloud Reconstruction

The computed extrinsic calibration is used to *fuse* the radar and camera measurement and retrieve the 3D coordinates of targets similar to (El Natour et al., 2015a). Using the pinhole model in Equation (4), a point $\boldsymbol{m}_c$ can be represented in terms of the image coordinates and the intrinsic calibration matrix K

$$\boldsymbol{m}_c = z_c \boldsymbol{K}^{-1} \boldsymbol{p} = z_c \boldsymbol{q}, \tag{15}$$

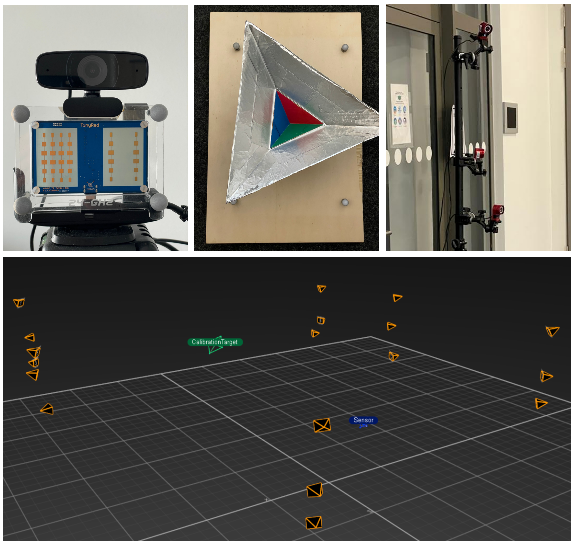

Figure 4: Top: the camera-radar setup with the reflective markers, the calibration target, three OptiTrack cameras. Bottom: OptiTrack cameras detecting the calibration target (green) and the sensor (blue) in 3D.

where $\boldsymbol{q} = [q_1, q_2, q_3]^\top = \boldsymbol{K}^{-1}\boldsymbol{p}$. To compute z_c, the equation of the sphere in Equation (8) is used in the camera coordinate system and replacing the values for $\boldsymbol{m}_c$ as in Equation (15)

$$\begin{aligned} &(x_c - x_{s_c})^2 + (y_c - y_{s_c})^2 + (z_c - z_{s_c})^2 = \rho^2 \\ \Rightarrow &z_c^2(q_1^2 + q_2^2 + q_3^2) - 2z_c(q_1 x_{s_c} + q_2 y_{s_c} + q_3 z_{s_c}) \\ &+ (x_{s_c}^2 + y_{s_c}^2 + z_{s_c}^2 - \rho^2) = 0. \end{aligned} \tag{16}$$

The solution to the quadratic equation in Equation (16) yields two possible solutions. The correct solution is the one that gives a closer results of $\boldsymbol{m}_s = \boldsymbol{H}_{sc}\boldsymbol{m}_c$ to Equation (5) with $\phi = \pi/2$.

5 EVALUATION

The results of the algorithm are evaluated against the work of El Natour et al. (El Natour et al., 2015a). This is the only existing camera-radar 3D calibration method for 2D radar devices for static calibration. The evaluation of both methods is done for the first time against highly accurate ground truth measurements from an optical tracker.

5.1 Ground Truth Acquisition

The evaluation calibration algorithms and 3D reconstruction requires an accurate and precise method for measuring the ground truth values. Therefore, an optical motion capture (OptiTrack) system that can achieve < 1 *mm* localization error is used. This system is only used for a quantitative evaluation of the calibration results, and not part of the calibration algorithm.

Eighteen OptiTrack Flex 13[1] cameras are mounted as to cover the empty space where the calibration target is placed. Reflective markers are placed on both the sensor and the calibration targets to be detected by the cameras. Both the camera-radar setup as well as the calibration target are visible in the cameras' field of view at all times thus allowing the accurate measurement of their relative positions.

Since the motion capture system uses infrared light to detect the reflective markers, the calibration measurements with this type of ground truth are only possible to perform indoors for validation purposes.

5.2 Hardware Setup

The radar in use is an Analog Devices TinyRad[2], this radar is an evaluation module operating at 24 GHz with a range resolution of 0.6 *m* and an azimuth resolution 0.35 *rad*.

The camera used carries a 2 megapixels sensor capable of recording full HD images (1080p) at 30 FPS (frames per second) with a 78° FOV (field-of-view).

The sensors are mounted with the camera on top of the radar with a short baseline ≈ 5 *cm* as seen in Figure 4.

5.3 Initialization

Knowing the difference in coordinate system orientations between the camera and radar, an initial parameter vector $[\alpha_0, \beta_0, \gamma_0, 0, 0, 0]$ is used to roughly align the axes and thus speed up the convergence of the rotation matrix.

While (El Natour et al., 2015a) uses stereo based 3D reconstruction to calculate their ground truth as well as their required *a priori* inter-target distance, this paper reproduces their work using the more accurate OptiTrack data instead. The inter-target distance is used to solve an optimization problem to calculate the z_c values needed to solve the main optimization problem.

Furthermore, it was not possible to reproduce and converge the z_c solution from (El Natour et al., 2015a) using a zero vector for initializing the optimization. Since the original authors provide no instructions to reproduce their work, the radar range ρ is used for

[1]https://optitrack.com/cameras/flex-13/

[2]https://www.analog.com/en/design-center/evaluation-hardware-and-software/evaluation-boards-kits/eval-tinyrad.html

initialization to achieve proper estimates. Another method is to manually measure the distances which would further complicate their approach.

5.4 Calibration Results

Different evaluation criteria are used to evaluate the quality of the calibration both in 3D and 2D. The convergence of the optimization is tested against different initializations and the quality of the calibration is evaluated against the number of measurements needed as well as the level of noise level in the measurement. In addition to that, an ablation study of the elevation constraint is performed to show its importance.

The error in 3D is the distance between the estimated 3D reprojection of the target and the ground truth as measured by the OptiTrack system, and the error in 2D is the distance between the projection of the estimated 3D target and the ground truth on the *xy-plane*.

5.4.1 Evaluation of the Initialization

To evaluate the effect of initialization on the calibration, the optimization is performed with different starting parameters while maintaining the same setup in all experiments. Table 1 show that both approaches presented in this work achieved the same average errors and standard deviations for all initialization conditions as well as significantly outperformed the rival approach. The initialization levels are defined as

$$\begin{aligned} \textbf{Best: } & [\alpha_0, \beta_0, \gamma_0, 0, 0, 0], \\ \textbf{Moderate: } & [\alpha_0, \beta_0, \gamma_0, 0, 0, 0] + \mu_{1\times 6}, \\ \textbf{Bad: } & [\alpha_0, \beta_0, \gamma_0, 0, 0, 0] + \nu_{1\times 6}, \\ \text{where } & \mu_{1-3} \in [-1\ rad, 1\ rad] \ \&\ \mu_{4-6} \in [-0.1, 0.1] \\ \text{and } & \nu_{1-3} \in [-2\ rad, 2\ rad] \ \&\ \nu_{4-6} \in [-0.5, 0.5], \end{aligned} \tag{17}$$

the components of μ and ν are uniformly sampled from the respective ranges and added to the initialization parameters described in Section 5.3. The results also show that the rival method was able to optimize the 2D projection on the $xy-$plane better than 3D space.

Our approaches show consistently better results regardless of the initialization and both in 3D as well as 2D projections both on the radar plane and the image plane. It is also worth noting the lower standard deviation accompanied with the lower error indicates higher confidence as well a better fit to the data.

5.4.2 Evaluation of the Number of Targets

Another experiment was performed to observe the dependency of the calibration algorithms on the number of measurements needed. The LM implementation (Moré, 1978) requires the number of residuals to be greater than or equal to the number of parameters to be estimated. As mentioned in Section 4.3, we are estimating 6 parameters $[\alpha, \beta, \gamma, x_{c_s}, y_{c_s}, z_{c_s}]$, and each measured target position generates 3 residuals, thus a theoretical minimum of 2 target positions are needed. However, our experiments showed that in practice, 5 target positions are needed to converge to a valid solution.

The results in Figure 5 show significantly lower dependency on the number of targets for our approaches and even though some improvement can be seen with more measurements, the calibration yielded with only 5 measurements, an error more than five times lower than (El Natour et al., 2015a) with 36 measurements for 3D reconstruction and two times lower for 2D. This experiment was repeated 250 times for each $n \in [\![5, 36]\!]$ measurements randomly sampled without replacement out of the 36 measurements. The results are the averaged over the runs.

The poor performance of the method of El Natour (El Natour et al., 2015a) on our data does not come as a surprise if we consider that their real-data evaluation published in (El Natour et al., 2015a) shows an error that is a few orders of magnitude worse than the error on simulated data. In their real-data evaluation, (El Natour et al., 2015a) shows a mean error of 0.63 m, on a longer range and in an outdoor setting which is less affected by multi-path interference.

5.4.3 Simulations of Noise Levels

We simulated the effect of different levels of noise on the reconstruction error of our calibration algorithm. We identified three main sources of noise: radar range measurement ρ, radar azimuth measurement θ, and camera pixel error (u, v). In this experiment, the OptiTrack ground truth measurements are used as a baseline ($level-0$) and define our noise levels such that for each target i we have

$$\begin{aligned} \rho_{il} &= \rho_{i0} + \mathcal{N}(0, (0.05 \times l)^2), \\ \theta_{il} &= \theta_{i0} + \mathcal{N}(0, (0.01 \times l)^2), \\ (u_{il}, v_{il}) &= (u_{i0} + \mathcal{N}(0, l^2), v_{i0} + \mathcal{N}(0, l^2)), \end{aligned} \tag{18}$$

where l is the noise level, $\mathcal{N}$ is the normal distribution, and $l \in [\![1, 10]\!]$, and ρ_{i0}, θ_{i0}, and (u_{i0}, v_{i0}) are the $level-0$ measurements.

Table 1: A comparison of the mean error in meters between our method and (El Natour et al., 2015a). Different initialization setups are used to evaluate the sensitivity of the optimizations to their starting point. Where **Best:** $[\alpha_0, \beta_0, \gamma_0, 0, 0, 0]$, **Moderate:** $[\alpha_0, \beta_0, \gamma_0, 0, 0, 0] + \mu_{1\times 6} \ni \{\mu_{1-3} \in [-1\ rad, 1\ rad]\ \&\ \mu_{4-6} \in [-0.1, 0.1]\}$, and **Bad:** $[\alpha_0, \beta_0, \gamma_0, 0, 0, 0] + \nu_{1\times 6} \ni \{\nu_{1-3} \in [-2\ rad, 2\ rad]\ \&\ \nu_{4-6} \in [-0.5, 0.5]\}$.

Method	**Error**	**Initialization**		
		Best	**Moderate**	**Bad**
Using camera correspondences **(ours)**	3D	**0.175 $m \pm$ 0.049**	**0.175 $m \pm$ 0.049**	**0.175 $m \pm$ 0.049**
	2D	**0.129 $m \pm$ 0.065**	**0.129 $m \pm$ 0.065**	**0.129 $m \pm$ 0.065**
El Natour et al. (El Natour et al., 2015a)	3D	1.474 $m \pm$ 0.640	1.474 $m \pm$ 0.640	1.606 $m \pm$ 0.440
	2D	0.272 $m \pm$ 0.166	0.272 $m \pm$ 0.166	0.313 $m \pm$ 0.090

Table 2: An ablation study showing the effect of adding the constraint in Equation (13) to the optimization and the difference between using the radar range and camera correspondences to calculate the distance to the target. The experiments were done using the **Best** initialization parameters $[\alpha_0, \beta_0, \gamma_0, 0, 0, 0]$.

Method	**Error**	**Results**	
		Without Eq. (13) constraint	**With Eq. (13) constraint**
Using radar range **(ours)**	3D	1.369 $m \pm$ 0.620	0.180 $m \pm$ 0.053
	2D	0.250 $m \pm$ 0.152	0.133 $m \pm$ 0.067
Using camera correspondences **(ours)**	3D	1.546 $m \pm$ 0.676	**0.175 $m \pm$ 0.049**
	2D	0.293 $m \pm$ 0.182	**0.129 $m \pm$ 0.065**

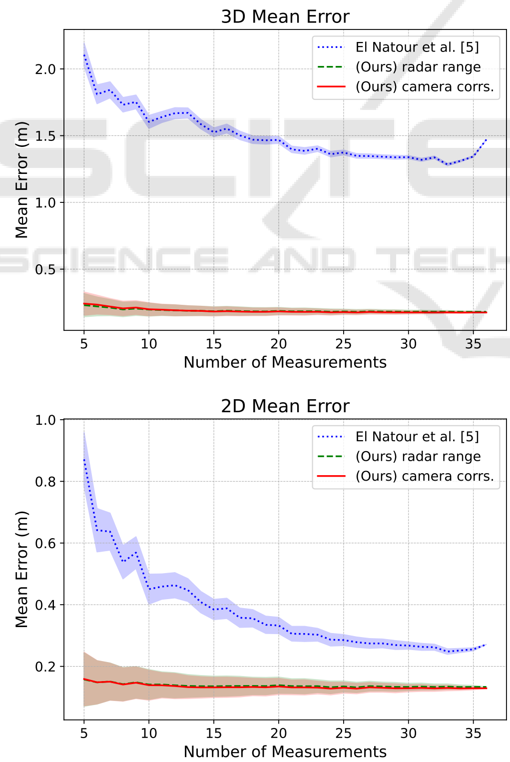

Figure 5: The effect of the number of measured targets on the result of the calibration and subsequently on the quality of the 2D and 3D reconstruction. Our methods (overlapping red and green) show better error even for a low number of measurements.

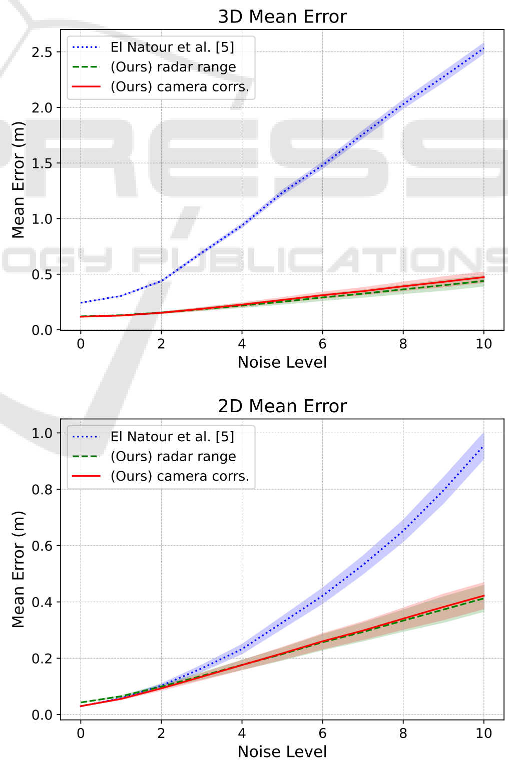

Figure 6: Our methods (overlapping red and green) show resilience to the increasing noise levels for both 2D and 3D reconstruction and achieve a worst case of 0.5 m mean error for 3D reconstruction. Both methods outperform (El Natour et al., 2015a) which performs 5 times worse in 3D reconstruction and 2 times worse in 2D reconstruction at noise level 10.

Figure 6 shows the 2D and 3D target reconstructions errors for the different noise levels. We can see our method outperforming (El Natour et al., 2015a) in for all levels of noise in 3D reconstruction. For 2D reconstruction, all three methods show similar reconstruction error for $level-0$ noise, however, (El Natour et al., 2015a) quickly diverges after $level-4$ noise. This experiment was repeated 250 times for all noise levels, and show our methods' robustness to noise.

5.4.4 Ablation Study of the Elevation Constraint

To highlight the importance of our elevation constraint, defined in Equation (13), we ran an ablation study on both of our range calculation methods. This was done using the **Best** initialization parameters as described in Equation (17) and the results can be seen in Table 2. The mean errors achieved without using the elevation constraint are considerably higher than the results achieved when including it. The errors without the constraint in Equation (13) are also close to the results of (El Natour et al., 2015a) as seen in Table 1. This is expected because the main difference between them is the distance calculation method. We observe that before adding the constraint, our method using the radar range for distance measurement performed better than the method using the camera correspondences, this result was reversed after adding the constraint.

6 CONCLUSION

In this work, we introduced a new method for extrinsic calibration of a camera-radar system. The method was tested against a high-accuracy motion capture system, which served as the ground truth. Our setup is not only simpler, as it operates independently without external sensing, but it also delivers superior results. Even with less accurate initial parameters and fewer measurement points, the additional optimization constraints we introduced allow the calibration to converge effectively. We also utilized the calibration output to reconstruct the 3D targets from the data matched by the camera-radar system. Instead of more complicated target designs, our streamlined setup requires fewer calibration targets and merely uses a single standard retroreflector. While our current approach only focuses on static targets, calibrating on a moving target would likely yield better radar target detection. However, this would come at the cost of complicating the process, including the setup and target detection phases of our work.

ACKNOWLEDGMENT

This work has been partially funded by the German Ministry of Education and Research (BMBF) of the Federal Republic of Germany under the research project RACKET (Grant number 01IW20009). Special thanks to Stephan Krauß, Narek Minaskan, and Alain Pagani for their insightful remarks and discussions.

REFERENCES

Bian, J., Lin, W.-Y., Matsushita, Y., Yeung, S.-K., Nguyen, T.-D., and Cheng, M.-M. (2017). Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4181–4190. IEEE.

Chavez-Garcia, R. O. and Aycard, O. (2015). Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):525–534.

Cho, H., Seo, Y.-W., Kumar, B. V., and Rajkumar, R. R. (2014). A multi-sensor fusion system for moving object detection and tracking in urban driving environments. In *International Conference on Robotics and Automation (ICRA)*, pages 1836–1843. IEEE.

Domhof, J., Kooij, J. F., and Gavrila, D. M. (2019). An extrinsic calibration tool for radar, camera and lidar. In *International Conference on Robotics and Automation (ICRA)*, pages 8107–8113. IEEE.

El Natour, G., Aider, O. A., Rouveure, R., Berry, F., and Faure, P. (2015a). Radar and vision sensors calibration for outdoor 3d reconstruction. In *International Conference on Robotics and Automation (ICRA)*, pages 2084–2089. IEEE.

El Natour, G., Ait-Aider, O., Rouveure, R., Berry, F., and Faure, P. (2015b). Toward 3d reconstruction of outdoor scenes using an mmw radar and a monocular vision sensor. *Sensors*, 15(10):25937–25967.

Kim, D. Y. and Jeon, M. (2014). Data fusion of radar and image measurements for multi-object tracking via kalman filtering. *Information Sciences*, 278:641–652.

Lepetit, V., Moreno-Noguer, F., and Fua, P. (2009). Epnp: An accurate o (n) solution to the pnp problem. *International journal of computer vision*, 81(2):155–166.

Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on Artificial intelligence (IJCAI)*, volume 2, pages 674–679.

Moré, J. J. (1978). The levenberg-marquardt algorithm: implementation and theory. In *Numerical analysis*, pages 105–116. Springer.

Oh, J., Kim, K.-S., Park, M., and Kim, S. (2018). A comparative study on camera-radar calibration methods. In *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 1057–1062. IEEE.

OpenCV (2023a). Camera calibration and 3d reconstruction.

OpenCV (2023b). Object detection.

Peršić, J., Marković, I., and Petrović, I. (2019). Extrinsic 6dof calibration of a radar–lidar–camera system enhanced by radar cross section estimates evaluation. *Robotics and Autonomous Systems*, 114:217–230.

Rambach, J., Pagani, A., and Stricker, D. (2021). Principles of object tracking and mapping. In *Springer Handbook of Augmented Reality*, pages 53–84. Springer.

Roy, A., Gale, N., and Hong, L. (2009). Fusion of doppler radar and video information for automated traffic surveillance. In *12th International Conference on Information Fusion*, pages 1989–1996. IEEE.

Roy, A., Gale, N., and Hong, L. (2011). Automated traffic surveillance using fusion of doppler radar and video information. *Mathematical and Computer Modelling*, 54(1-2):531–543.

Stateczny, A., Kazimierski, W., Gronska-Sledz, D., and Motyl, W. (2019). The empirical application of automotive 3d radar sensor for target detection for an autonomous surface vehicle's navigation. *Remote Sensing*, 11(10):1156.

Sugimoto, S., Tateda, H., Takahashi, H., and Okutomi, M. (2004). Obstacle detection using millimeter-wave radar and its visualization on image sequence. In *International Conference on Pattern Recognition (ICPR)*, volume 3, pages 342–345. IEEE.

Wang, T., Zheng, N., Xin, J., and Ma, Z. (2011). Integrating millimeter wave radar with a monocular vision sensor for on-road obstacle detection applications. *Sensors*, 11(9):8992–9008.

Wang, X., Xu, L., Sun, H., Xin, J., and Zheng, N. (2014). Bionic vision inspired on-road obstacle detection and tracking using radar and visual information. In *International Conference on Intelligent Transportation Systems (ITSC)*, pages 39–44. IEEE.

Wise, E., Peršić, J., Grebe, C., Petrović, I., and Kelly, J. (2021). A continuous-time approach for 3d radar-to-camera extrinsic calibration. In *International Conference on Robotics and Automation (ICRA)*, pages 13164–13170. IEEE.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334.

APPENDIX

Target Matching in Camera Frames

In this section we will explain in detail the process of automatically detecting the calibration target as well its corners and center as required in Section 4.2.2. This step is a decoupled problem from the calibration algorithm and other approaches are available to achieve this task.

Figure 7: A sample input frame where the corner targets must be detected.

Template Matching

Given a sample input frame as shown in Figure 7, it is desired to first find the location of the target in the frame to reduce the search space for feature matching and improve its robustness.

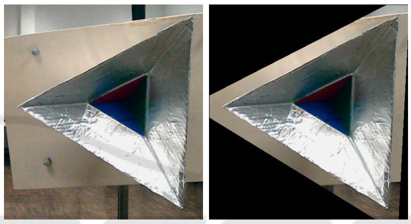

Figure 8: (left) Patch of the calibration target to be used as a template. (right) Masked version of the patch.

A template (Figure 8) is used to find the candidate positions of the target in the input frame. Template matching (OpenCV, 2023b) is applied to the input frame with different scales and rotations of the template. Our experiments show that masking out the clutter around the target in the template as seen in Figure 8 yields more robust detection.

The template matching results in multiple candidate positions that are merged depending on their overlap value with a vote being assigned to the merged patch corresponding to the number of candidates merged at that position. The result of steps (1) and (2) of Figure 9 shows the candidate positions before and after merging as well as the final selected patch in the input frame.

Template Alignment

After finding the desired patch, we use the GMS Feature Matcher (Bian et al., 2017) to detect correspondences between the template and the selected patch. The GMS Matcher allows for more robust feature matching between the two patches. Figure 9 shows the result of the feature matching (3) after resizing the patches to similar dimensions and applying edge enhancing filtering.

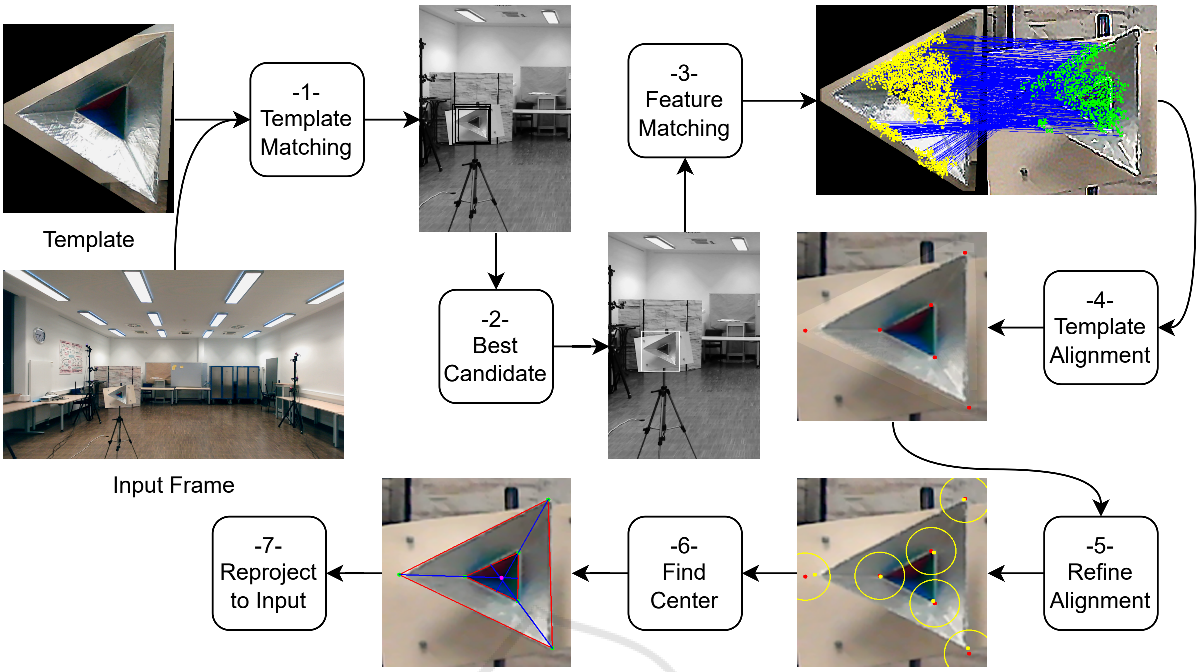

Figure 9: The complete diagram describing the detection of the target. (1) Generate candidate positions of the template matching. (2) Find best patch location. (3) Match features between the template and the chosen patch location. (4) Align the template to the patch, the red dots represent the known corner positions of the template. (5) Refine the corner positions, the yellow dots show the refined position of the corner. (6) Calculate the center point as the intersection of the lines connecting the corners. (7) Reproject the center to the original image.

The matched features are used to compute the homography (OpenCV, 2023a)and align the template to the image patch. Step (4) of Figure 9 presents the aligned template on top of the image patch, the red dots represent the known corner positions in the template. The figure shows that some refinement is needed for proper alignment of the corners.

Corner Refinement

To further refine the position of the corners in the image patch, we apply Lukas-Kanade method (Lucas and Kanade, 1981) for sparse optical flow. We assume small changes in the corner position from one image to the other and apply the method for each corner separately for more robust results. Figure 9 shows the result of the corner refinement in yellow (5), the circle around the original red corner position has a radius of 50 pixels (image is upscaled by 3.5 times).

Center Detection

After detecting the corners on the target, the detection of the center becomes straightforward, by connecting the corners of both triangles and finding the mean intersections between the 3 lines as seen in Figure 9.

Finally, Figure 10 shows the position of the center point obtained in the original image.

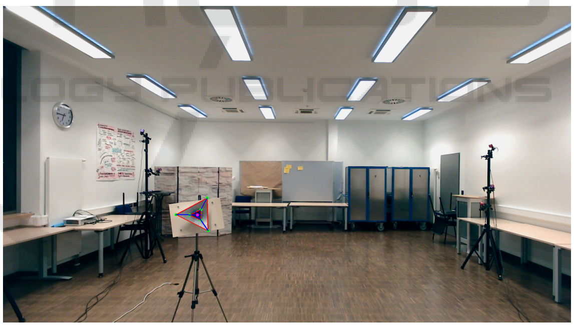

Figure 10: The target center point projected back to the original image.

Decoupled Noise Simulations

In addition to the experiments presented in Section 5.4.3 on the combined noise levels, we performed 3 experiments exploring the decoupled effects of the added noise. Similar to the simulations in Section 5.4.3, we ran the experiments 250 times for each noise level, and averaged over the runs. The noise is defined similar to Section 5.4.3 as

$$\begin{aligned} \rho_{il} &= \rho_{i0} + \mathcal{N}(0,(0.05\times l)^2), \\ \theta_{il} &= \theta_{i0} + \mathcal{N}(0,(0.01\times l)^2), \\ (u_{il},v_{il}) &= (u_{i0}+\mathcal{N}(0,l^2), v_{i0}+\mathcal{N}(0,l^2)), \end{aligned} \tag{19}$$

where l is the noise level, $\mathcal{N}$ is the normal distribution, and $l \in [\![1, 10]\!]$, and ρ_{i0}, θ_{i0}, and (u_{i0}, v_{i0}) are the $level - 0$ measurements. For each of the experiments noise is added to one of the parameters while $level - 0$ noise is used for the other two.

Radar Range Noise

Simulating radar range noise shows a similar increase in 3D mean error for both our methods and (El Natour et al., 2015a). Figure 11 shows that our methods outperform El Natour et al. (El Natour et al., 2015a) over for all standard deviation values with the difference growing slightly the larger the noise standard deviation.

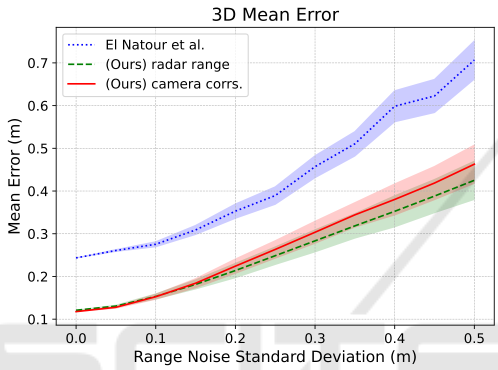

Figure 11: The mean error in 3D reconstruction as a function of the standard deviation of the added noise to the range.

Radar Azimuth Noise

Figure 12 shows that noise in the radar azimuth measurement has the biggest effect on the quality of (El Natour et al., 2015a) while our methods are much more robust to this type of noise. The mean reconstruction error of (El Natour et al., 2015a) shoots to more than 2 m for a noise standard deviation of 0.1 rad while our methods maintain an error lower than 0.25 m for the same range.

Camera Pixel Noise

Introducing noise of up to 10 pixels to all methods had the smallest effect on all methods. While our methods show an average increase in 3D reconstruction error of around 1 cm, El Natour et al. (El Natour et al., 2015a) increase by 2 to 3 cm.

Analysis of Results

The experiments show the robustness of our method to different sources of noise with the radar range noise

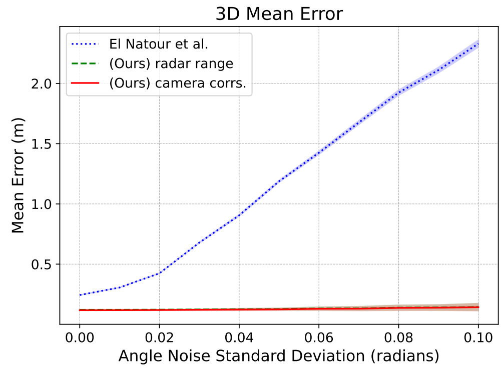

Figure 12: The mean error in 3D reconstruction as a function of the standard deviation of the added noise to the azimuth angle. The method by El Natour et al. (El Natour et al., 2015a) shows very high sensitivity to the azimuth noise.

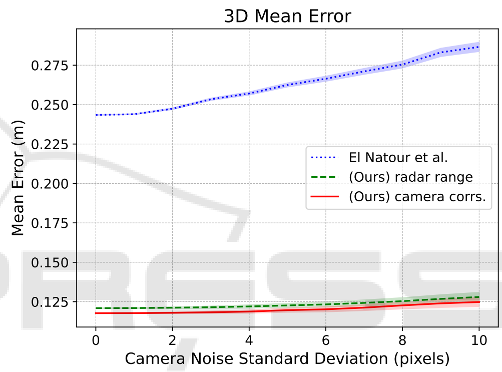

Figure 13: The mean error in 3D reconstruction as a function of the standard deviation of the added noise to the target pixel location.

having the biggest effect on our 3D reconstruction results. On the other hand, the method in (El Natour et al., 2015a) shows a lot of sensitivity to variations in the radar azimuth measurements, and overall lower accuracy of the 3D reconstruction of the measurement targets.