Cost-effective camera based ground truth for indoor localization

Daniel Becker¹, Fabian Thiele¹, Oliver Sawade², Ilja Radusch²

Abstract—One of the key requirements for the evaluation of indoor localization systems is an accurate and reliable ground truth. Existing ground truth systems are often expensive due to high hardware cost and complex deployment. In this work, we present a simple yet highly accurate approach for a cost-effective ground truth system based on off-the-shelf infrastructure cameras and printable markers. We developed a marker detection algorithm and systematic 3-layer projection approach between multiple coordinate systems which achieves a median accuracy of 0.48cm, 0.05 degrees and a minimum accuracy of 0.75cm, 0.27 degrees for 2D position and orientation.

I. INTRODUCTION

There is a myriad of applications for vehicles and wheeled robots operating in indoor environments. From industrial robots and automated warehouses to intelligent vehicles, all these applications rely on highly accurate localization. Especially for applications which are completely automated, i.e., which operate without human supervision, localization requirements are particularly high. Depending on the specific application, localization systems need to determine position and orientation of targets (e.g., robot or vehicle) in a 2D plane or 3D space. Hence, in the 2D plane three parameters (x/y position and yaw angle) and in 3D space six parameters (x/y/z position and yaw/pitch/roll angle) need to be determined. Another point that highlights the importance of ground truth systems is the ability to benchmark various localization systems and algorithms under comparable conditions [1] [2].

To evaluate the performance of any localization system, a ground truth system is required which ideally is at least an order of magnitude more accurate than the evaluated system. At first glance, both the ground truth and productive system perform the same task: They estimate position and orientation in 2D plane or 3D space. The difference is that the productive system typically has more cost constraints due to the need for economic operation in real-world applications, whereas the ground truth can be more expensive as it is only used for a limited time in a limited area to evaluate the productive system.

Outdoors, GNSS (Global Navigation Satellite System) based systems are often used as ground truth, especially DGPS (Differential Global Positioning System) where the positioning accuracy is significantly improved using additional transponders at fixed known locations on the ground. Under typical outdoor conditions, DGPS achieves 1m localization accuracy [1]. However, in indoor spaces or when the line-of-sight to the navigation satellites is blocked, the DGPS accuracy drops significantly or ceases completely [1].

In this work, we present a cost-effective ground truth system based on off-the-shelf infrastructure cameras and printable optical markers. Towards the goal of optimizing the system’s accuracy, we introduce a systematic 3-layer projection approach between image, reference grid and global coordinates. Our approach is suitable for vehicles and wheeled robots where the marker can be mounted at a fixed height. A detailed evaluation using manual high-precision measurements as reference shows that the 2D position and orientation can be measured at a median accuracy of 0.48cm, 0.05° and at a minimum accuracy of 0.75cm and 0.27°.

This work is organized as follows: In section II, related work is examined and classified according to several performance metrics. In section III, we explain the individual parts of our proposed methodology forming the reference positioning system. Section IV shows a detailed evaluation of our proposed approach and provides results about position accuracy and real-time performance. Conclusions are drawn in Section VI where we also give an outlook on future work.

II. RELATED WORK

The field of indoor localization has been an active research focus for decades, as an accurate and reliable localization is a key building block for many applications from automatic warehouses to autonomous vehicles. Any localization approach has to be validated by a ground truth system to produce scientifically valid statements about accuracy. For instance, for the task of automated driving in indoor car parks, an accuracy of at least 10cm and 1° for position and orientation is often quoted as sufficient [3], [4]. The ground truth needs to be more accurate to validate productive systems - ideally by an order of magnitude (i.e. 1cm and

![Fig. 1: Trade-off between cost, coverage and effort for localization systems.](image-url)
To achieve this high accuracy, trade-offs need to be made as shown in Fig. 1. At a given accuracy, it is very difficult to get a favorable combination of the three factors: cost, coverage, and effort. For instance, localization accuracy can be improved by utilizing a more expensive technology compared to the productive system, reducing the coverage by increasing the measurement density or by employing more human effort (e.g., manual measurements).

In localization systems, the position is either determined within the target (internally) or in the infrastructure (externally). A common approach based on internal sensors consists of relative positioning technologies, i.e., dead reckoning. To achieve necessary precision, high-precision Inertial Measurement Units (IMU) are typically used in the automotive industry and autonomous robots [5]. However, the errors of IMUs are not bounded, i.e., the errors are constantly rising over time and the estimated position is diverging from the true position. Hence, IMUs are not well-suited without additional sensors. In [6] a system based on multiple cameras mounted to a robot is introduced. The estimated trajectories from each camera are compared to each other but not to an external ground truth. Another alternative localization technology is based on Ultra Wide Band (UWB) where a position can be estimated in relation to the transmitted signals of infrastructure beacons. Under ideal conditions, decimeter accuracy can be reached although it is common to observe high variations and large outliers over time and space [7].

In terms of external localization based on infrastructure sensors, [4] have shown a prototype for automated indoor driving using infrastructure laser scanners. They have measured an accuracy of about 10 cm using manual ground truth. Examples of external localization with cameras are presented in [8] and [9], but neither system is able to achieve a sufficient accuracy to qualify as ground truth. The RAWSEEDS project has presented two ground-truth approaches based on cameras and lidar scanners [10]. The vision-based system achieved an average error in the 10 cm range, while a fusion with the lidar-based approach offered an improvement to 5 cm and 0.5° angle. A comprehensive survey of optical indoor localization systems is provided in [11]. Therein, a comparison of about 20 systems highlights the trade-off between accuracy and cost: All systems surpassing centimeter accuracy require expensive cameras in the price range of 1000$ [11]. In contrast, our system aims to bridge this gap surpassing centimeter accuracy at a significantly lower cost.

### III. Methodology

In the following, we present a 3-layer projection approach towards the goal of obtaining a position and orientation estimate of a vehicle with the best possible accuracy. The first step is the detection of marker positions in the image plane of the camera, \( p_{\text{img}} \) (pixels). Secondly, the position is projected onto a reference grid plane, resulting in position \( p_{\text{ref}} \) (meters). The last step is the determination of the global position \( p_{\text{utm}} \) (UTM projection [12] coordinate in meters). The coordinates in each step are defined as follows:

\[
\begin{align*}
  p_{\text{img}} &= \left[ \begin{array}{c} x_{\text{img}} \, \text{px} \\ y_{\text{img}} \, \text{px} \end{array} \right], \\
  p_{\text{ref}} &= \left[ \begin{array}{c} x_{\text{ref},m} \\ y_{\text{ref},m} \end{array} \right], \\
  p_{\text{utm}} &= \left[ \begin{array}{c} x_{\text{utm},m} \\ y_{\text{utm},m} \end{array} \right] 
\end{align*}
\]

The relations between these three coordinate systems are illustrated in Fig. 2, displaying the placement of a chessboard reference grid in our underground carpark testbed [13] and the coverage of the camera onto the grid (indicated in blue). The grid consists of alternating black/white square shapes with a side length of 16.4 cm. The camera is at a mounting height \( H \) (cf. Fig. 4) of 2.75 m and the camera view covers a total area of 2.1 m x 3.4 m. Moreover, 4 map objects are shown in Fig. 2 (1 wall, 1 square pillar and 2 round pillars). These map objects are represented in a given indoor map with geographic coordinates (UTM [12]).

The first step is the detection of markers in the image, which are projected into reference plane coordinates and finally global UTM coordinates:

\[
p_{\text{img}} \rightarrow p_{\text{ref}} \rightarrow p_{\text{utm}}
\]

#### A. Detection of Image Plane Position

In order to achieve the required accuracy and meet tight cost constraints we have developed printable colored markers in conjunction with a custom detection algorithm. The detection algorithm is divided in multiple steps, shown in Fig. 3 and explained in the following.

1) **Image Preprocessing, Fig. 3a:** Images captured by the camera are converted to grayscale and blurred by convolving the grayscale image with a 3x3 averaging kernel. Grayscale conversion is needed to perform the edge detection in the next step. Blurring the image reduces noise and improves the edge detection result.

2) **Contour Detection, Fig. 3b:** An edge detection algorithm is used to find the image contours. Found contours are approximated with polygons using the Ramer-Douglas-Peucker [14] algorithm which is provided by OpenCV. The contours are approximated with polygons using the Ramer-Douglas-Peucker [14] algorithm which is provided by OpenCV.
approximated polygons are filtered for potential marker outlines by removing all polygons with an edge count unequal to four or an area out of the expected bounds. Approximated polygons are drawn in Fig. 3b: Polygons drawn in magenta have an incorrect vertex count, polygons drawn in cyan have an incorrect size. The green polygon is the one approximating the marker outline.

3) Edge Refinement, Fig. 3c: The vertex coordinates of approximated polygons only offer integer accuracy. Moreover, due to image noise, vertex positions can vary by multiple pixels from frame to frame resulting in an inaccurate, noisy marker position. In order to achieve sub-pixel accuracy the polygon edges are refined by analyzing the two-dimensional gradient image of the grayscale image. The marker outline is a black border on white background resulting in large gradients pointing in a direction orthogonal to the marker edge and away from the center. Hence, the gradient image is searched along each polygon edge for local maxima pointing away from the center of the polygon. In a second step the edge is matched to the found local maxima minimizing the mean distance. Fig. 3c visualizes the gradient image: Red pixels indicate gradient vectors pointing in an incorrect direction, whereas green pixels indicate gradient vectors pointing in the right direction. The brightness of each pixel is proportional to the gradient vector magnitude. The matched polygon is drawn in white.

4) Marker ID sampling, Fig. 3d: Finally, based on the detected marker outline, the color sampling points are computed. The sampled colors define the marker direction and ID. The white corner is defined as the front-left corner. The remaining three corners define the marker ID: Each corner might be red, green or blue resulting in 27 unique combinations. Fig. 3d shows the original input image together with the detected marker outline, marker direction and color sampling points.

5) Height projection, Fig. 4: The markers are not residing on the floor plane but have a height offset (cf. Fig. 4). Hence, there is a projection error $\Delta d$ which depends on relation between the marker height $h$ and the camera mounting height $H$ as well as the angle $\xi$ or offset $\Delta x$ between camera perpendicular and marker.

The projection error $\Delta d$ can be calculated as follows:

$$\Delta d = \tan(\xi) h$$

The angle $\xi$ can be estimated as:

$$\xi = \tan^{-1} \left( \frac{\Delta x}{H - h} \right)$$

$\Delta x$ refers to a distance in the reference coordinate plane and cannot be measured directly. However, it can be estimated using the homography projection between camera and reference plane, as explained in the following chapter. $\Delta d$ needs to be calculated for both dimensions in the reference plane (x/y). Once calculated, it can be subtracted from the detected marker position thus reducing the positioning error.

B. Determination of Reference Plane Position

For a given set of positions $P_{\text{img}} = \{p_{\text{img},1}, \ldots, p_{\text{img},n}\}$ in the image plane and a set of corresponding reference plane positions $P_{\text{ref}} = \{p_{\text{ref},1}, \ldots, p_{\text{ref},n}\}$ (position $p_{\text{img},i}$ corresponds to $p_{\text{ref},i}$), we obtain a 3x3 homography matrix $H_A$ using the OpenCV function cvFindHomography() [15]. $H_A$ describes 2 rotations and 1 translation. A minimum of 4 points is needed to determine $H_A$ but more points improve the robustness against noise in the input data [15]. As a result, the calculation of the coordinates is done as follows:

$$p_{\text{ref}} = H_A p_{\text{img}}, \quad p_{\text{img}} = H_A^{-1} p_{\text{ref}}$$

The sets of image and reference plane positions are determined from the chessboard shapes in the reference plane (see
Fig. 2). The OpenCV function `cvFindChessboardCorners()` calculates the pixel location of the corner points in the chessboard pattern in pixel accuracy and `cvFindCornerSubPix()` is used to refine the result to sub-pixel accuracy [15]. The first step eliminates the need for manual image annotation and the second step drastically improves the accuracy of the homography projection and thus the overall accuracy.

C. Determination of Global Position

Analogous to the calculation of the homography matrix $H_A$, we use a set of corresponding reference plane positions $P_{ref} = \{p_{ref,1}, \ldots, p_{ref,n}\}$ and global UTM positions $P_{utm} = \{p_{utm,1}, \ldots, p_{utm,n}\}$ to obtain homography matrix $H_B$. This enables the conversion of arbitrary reference plane positions into UTM positions and vice versa:

$$p_{utm} = H_B P_{ref}, \quad P_{ref} = H_B^{-1} P_{utm} \quad (6)$$

To determine the sets of UTM positions, a list of characteristic points in the vicinity of the reference grid (e.g. pillars, corners, etc.) are extracted from a given map that contains global coordinates for each map element (cf. Fig. 2). Next, the reference grid coordinates of these points are measured manually, e.g. by using an accurate laser range measurement device placed on the grid. Also, two markers are placed next to the device in order to determine the origin and angle of this measurement, representing a Polar coordinate that allows a straightforward conversion into the Cartesian form.

D. Other Aspects

Additional aspects for the implementation and evaluation of the proposed ground truth are discussed as follows.

1) Camera Distortion: Camera images often incur distortions caused by the camera optics [15]. For the proposed approach, we assume there is no camera distortion which would negatively affect the homographic mapping between image and reference plane. Thus, a camera with integrated undistortion is advisable. Alternatively, a calibration can be performed to estimate the intrinsic camera parameters in order to remove existing distortion [15].

2) Manual ground truth: The proposed ground truth system aims at achieving a high position accuracy of at least 1cm. In order to evaluate this approach, a more accurate measurement method is required. To this end, we equipped the vehicle with two laser pointers projecting onto the chessboard reference grid (cf. Fig. 4 and Fig. 6), yielding two measurements $p_{ref,L1}$ and $p_{ref,L2}$. We assume that the manual measurements are collected at an accuracy of at least 1mm. Further, we define the frontal laser pointer as vehicle reference point, hence $p_{ref,V} = p_{ref,L1}$. Moreover, the vehicle’s orientation $\theta_{ref,L}$ in reference grid coordinates can be calculated from $p_{ref,L1}$ and $p_{ref,L1}$.

3) Calculation of 2D Vehicle Position and Orientation: The marker detection yields the 2D central point as well as identifier for each detected colored marker (cf. Fig. 3). Our test vehicle is equipped with 3 markers as shown in Fig. 4. Thus we obtain 3 position detections $p_{ref,M1}, p_{ref,M2}$ and $p_{ref,M3}$ that are used to calculate two orientation angles $\theta_{ref,M1-M2}$ and $\theta_{ref,M2-M3}$. The mean angle is referred to as $\theta_{ref,M}$ and the centroid marker position as $p_{ref,M}$.

4) Multiple Camera Views: Multiple cameras can be installed viewing the same reference chessboard grid (cf. Fig. 2). Also, multiple grids can be deployed. In this case, each grid has its own coordinate system, hence additional mapping between the grids is required. Assuming two reference grids $\alpha$ and $\beta$, two sets of corresponding coordinates are measured, i.e. $P_{refA} = \{p_{refA,1}, \ldots, p_{refA,n}\}$ and $P_{refB} = \{p_{refB,1}, \ldots, p_{refB,n}\}$ ($p_{refA,i}$ corresponds to $p_{refB,i}$). Using the same methodology as described earlier, we obtain a homography projection matrix $H_{AB}$, which is used to transform arbitrary coordinates between the two reference planes.

5) Map Quality: For indoor environments, such as carparks, there usually are maps available containing the location of structural map elements (e.g. walls, pillars, lanes, etc.) with respect to a global coordinate frame, cf. Fig. 2. However, the quality of the indoor map needs to be taken into account as potential source of localization error. For instance, creating a map from an outdated construction plan potentially introduces errors due to deviations between initial planning and execution of the construction (e.g. displaced pillars, thicker walls, etc.). Thus, we do not perform a direct projection between image and global coordinates, i.e. $p_{img} \rightarrow p_{utm}$. Instead, we added the reference grid spanning its own coordinate frame (cf. Fig. 2) as intermediate step, i.e. $p_{img} \rightarrow p_{ref} \rightarrow p_{utm}$. So the first projection step $p_{img} \rightarrow p_{ref}$ is independent of the external environment manifesting in transparency and robustness. The second projection step $p_{ref} \rightarrow p_{utm}$ can also be referred to as anchoring, as a relation to known real-world map elements is established.

IV. Evaluation

The system’s testing environment, localization errors and real-time performance are investigated in the following.

A. Test Environment

A detailed description about our underground carpark test site is provided in our previous work [13]. We use an AXIS Q1604 (FW: 5.40.3.1) network camera at a mounting height of 2.75m that provides images at a resolution of 1280x720px via Gigabit Ethernet encoded as MJPEG or h264 at 24fps. A Smart Fortwo is used as test vehicle that is equipped with markers at a mounting height of 0.23m above the floor. The software is implemented in C++ utilizing the library OpenCV 2.4.9 and runs on a computer with an Intel(R) Core(TM) i7-4700MQ and 16GB of RAM on Ubuntu 12.04 LTS (64 bit) operating system.

B. Reprojection Error

As described previously, homography projection matrix $H_A$ (between image and reference plane) is generated from
Fig. 5: Evaluation results, Cumulative Density Functions (CDF) of A) position error $E_{\text{pos}}$ in cm, B) orientation error $E_{\theta}$ in $^\circ$ and C) detection time $\Delta T_{\text{det}}$ and total time $\Delta T_{\text{total}}$ in ms.

![Image](image_url)

Fig. 6: Laser pointer mounted to the car for measuring the position on the reference grid.

<table>
<thead>
<tr>
<th>$x$ [cm]</th>
<th>$y$ [cm]</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 percentile</td>
<td>0.01</td>
</tr>
<tr>
<td>70 percentile</td>
<td>0.05</td>
</tr>
<tr>
<td>95 percentile</td>
<td>0.1</td>
</tr>
<tr>
<td>97 percentile</td>
<td>0.2</td>
</tr>
<tr>
<td>100 percentile</td>
<td>0.3</td>
</tr>
</tbody>
</table>

TABLE I: Reprojection error $E_r$ for $x$ and $y$ dimensions.

two sets of correlated positions $P_{\text{img}} = \{p_{\text{img},1}, \ldots, p_{\text{img},n}\}$ and $P_{\text{ref}} = \{p_{\text{ref},1}, \ldots, p_{\text{ref},n}\}$ (position $p_{\text{img},i}$ corresponds to $p_{\text{ref},i}$). For any arbitrary image coordinate $p_{\text{img}}$ and $H_A$, the corresponding reference grid coordinate $p_{\text{ref}}$ can be calculated, cf. (5). So we define the reprojection error $E_r$ as:

$$E_r = H_A p_{\text{img},i} - p_{\text{ref},i} \quad \forall i \in \{1, \ldots, n\}$$

(7)

The following table displays the $x$ and $y$ reprojection error $E_r$ for different percentiles:

In summary, the reprojection error is relatively small and there are no large outliers which underlines the good quality of the homography projection matrix $H_A$. Interestingly, $E_r$ is larger for the $y$ dimension than $x$ dimension of $p_{\text{ref},i}$. This can be explained due to the setup in Fig. 2: The $y$ axis of the chessboard grid pattern used for calibration covers a wider angle range $\xi$ than the $x$ axis in the camera view.

C. Overall Error

As explained previously and shown in Fig. 4 and Fig. 6, we mounted two laser pointers to the car projecting a point onto the grid which can be manually measured. These positions represent the manual ground truth which is used to evaluate our approach. Also, the frontal laser pointer is used as vehicle reference point $p_{\text{ref},V}$.

As the mounting of markers and laser pointers is rigid, the displacement is constant over time. To be able to compare marker position $p_{\text{ref},M}$ and manually measured position $p_{\text{ref},L1}$, we perform a shift to the vehicle reference point:

$$p_{\text{ref},V} = p_{\text{ref},L1} = p_{\text{ref},M}^{\circ}$$

(8)

Consequently, we define the error in terms of position as:

$$E_{\text{pos}} = |p_{\text{ref},V} - p_{\text{ref},M}^{\circ}|$$

(9)

Also, the error for the orientation angle is defined as:

$$E_{\theta} = |\theta_{\text{ref},L} - \theta_{\text{ref},M}|$$

(10)

To measure the overall positioning error, we placed the vehicle at 15 different positions in the camera view, resulting in different variations of $p_{\text{ref},V}$, $\theta_{\text{ref},L}$ and also $\xi$.

Fig. 5 A) displays a Cumulative Density Function (CDF) of the position error $E_{\text{pos}}$. The 50 percentile, 90 percentile and maximum measured errors are at about 0.48cm, 0.65cm and 0.75cm. Fig. 5 B) shows a CDF for the orientation error $E_{\theta}$, where the 50 percentile, 90 percentile and maximum errors are about $0.05^\circ$, $0.23^\circ$ and $0.27^\circ$.

We found that the error is not constant throughout the image. Instead, the measured position error tends to be larger towards the edges of the image, i.e. for larger angles $\xi$. This can be for two reasons: First, the reprojection error towards the edges of the camera view is larger as our used camera was not perfectly calibrated, i.e. the camera view is not perfectly planar. Second, the height projection correction is not able to entirely compensate the height projection error as the estimation of angle $\xi$ is not perfectly accurate.

D. Performance

Fig. 5 C) shows the measured timings of the marker detection algorithm $\Delta T_{\text{det}}$ and the total detection time $\Delta T_{\text{total}}$ that includes $\Delta T_{\text{det}}$ time as well as all remaining processing (e.g. loading an image from the camera, performing the homography projection, etc). The 90 percentile time of $\Delta T_{\text{det}}$ is 30ms and the maximum time is about 45ms. The 90
percentile and maximum time for $\Delta T_{total}$ is 64ms and 80ms resp. Consequently, even in the worst case, the processing rate would be about 12Hz. In 90 percent of all cases, 15Hz can be achieved. We conclude that the proposed system is able to measure the marker positions in real-time at a rate which is suitable for fast moving objects such as vehicles.

E. Detection Rate

Another important point influencing the usability of the ground truth system is the false positive $R_{fp}$ and false negative rate $R_{fn}$. Let $i$ be the number of false detections, $j$ be the number of missed detections and $n$ be the total number of detections. Thus we define:

$$R_{fp} = \frac{i}{n}$$  \hspace{1cm} (11)

$$R_{fn} = \frac{j}{n}$$  \hspace{1cm} (12)

In the conducted evaluation, there have been incorrectly detected markers, especially in the chessboard pattern of the reference grid (cf. Fig. 2). However, the identifiers of these misdetections are not within the valid range of the detected colored markers, hence they can be filtered. Consequently, both $R_{fp}$ and $R_{fn}$ turned out to be zero, i.e. there every marker has been captured and there were no mismatches.

Moreover, carpark environments often have poor lighting conditions [13], which can negatively affect the detection in three ways: First, the detection could fail completely if the black/white contrast of the marker border becomes too low. Second, the color resolution decreases, thus reducing the reliability of obtaining correct identifiers. Third, the detection of moving objects can fail due to motion blur.

As a result, to be able to guarantee a stable detection operation, a good illumination below the camera should be ensured. In this process, diffused light sources are advantageous over focused sources, as inhomogeneous illumination of the marker surface can also affect the detection.

V. Conclusion and Outlook

In this work, we have proposed a highly accurate ground truth system based on marker detections and a systematic projection approach. The system is very cost-effective as it can be operated with off-the-shelf network cameras which are installed at fixed locations in the building infrastructure. Additionally, a uniform chessboard grid is used as reference plane and placed under the cameras. Also, the manual effort is reduced as calibration of the system can be automated by finding the chessboard corner points in the camera image.

The proposed system is suitable for vehicles or wheeled robots, where a marker can be mounted at a fixed height. In a detailed evaluation, we equipped a test vehicle with laser pointers projecting down onto the grid. We used these manually measured positions in order to evaluate the overall positioning accuracy. Thus, we determined a median and maximum position and orientation error of 0.48cm, 0.05° and 0.75cm, 0.27° resp.

In terms of future work we plan to add additional cameras to cover wide areas of our underground carpark test environment. The extended coverage of the ground truth system will facilitate an accurate and systematic evaluation of a wide variety of different vehicular positioning systems indoors. In the long run, this can be used to produce a benchmarking of different technologies and systems under comparable conditions.

REFERENCES


