

# Vision-based Localization of a Wheeled Mobile Robot with a Stereo Camera on a Pan-tilt Unit

A. Zdešar<sup>a</sup>, G. Klančar<sup>b</sup> and I. Škrjanc<sup>c</sup>

University of Ljubljana, Faculty of Electrical Engineering, Tržaška 25, 1000 Ljubljana, Slovenia

**Keywords:** Mobile Robots, Robot Vision, Optimal Filtering, Estimation Algorithms, Localization, System Observability.

**Abstract:** This paper is about a vision-based localization of a wheeled mobile robot (WMR) in an environment that contains multiple artificial landmarks, which are sparsely scattered and at known locations. The WMR is equipped with an on-board stereo camera that can detect the positions and IDs of the landmarks in the stereo image pair. The stereo camera is mounted on a pan-tilt unit that enables rotation of the camera with respect to the mobile robot. The paper presents an approach for calibration of the stereo camera on a pan-tilt unit based on observation of the scene from different views. Calibrated model of the system and the noise model are then used in the extended Kalman filter that estimates the mobile robot pose based on wheel odometry and stereo camera measurements of the landmarks. We assume that the mobile robot drives on a flat surface. In order to enforce this constraint, we transform the localization problem to a two-dimensional space. A short analysis of system observability based on indistinguishable states is also given. The presented models and algorithms were verified and validated in simulation environment.


## 1 INTRODUCTION


Autonomous mobile robots are one of the emerging fields of technology that is not only expected to become an inevitable part of smart factories of tomorrow but will play an essential role in our cities and homes of the future. Advances in the development of intelligent and autonomous systems are paving the way to a new breed of robotic systems that will be able to work alongside humans in a non-intrusive, harmless and cooperative way. Nowadays, the self-driving vehicles are being tested on roads in various traffic conditions daily and some self-driving technologies are already implemented in the most modern consumer vehicles.


Autonomous mobile systems perceive the environment through sensors. Commonly used sensors in mobile robotics are proximity and distance sensors (e.g. ultrasonic distance sensors, laser range scanners, etc.) that enable detection of obstacles, map building and localization. The emergence of new sensor technologies and contemporary computational capabilities broaden the range of available sensors that

are appropriate for environment perception. Cameras are one of the most promising sensors in the field of mobile robotics and they are seldom used to solve the mobile robot localization and mapping problems (Se et al., 2001; Agrawal and Konolige, 2006; Du and Tan, 2016; Fischer et al., 2016; Piasco et al., 2016; Fuentes-Pacheco et al., 2015; Konolige and Agrawal, 2008; Mei et al., 2011). The vision-based environment mapping and localization problems are commonly solved in the framework of Kalman filter (Chen, 2012) or particle filter (Kim et al., 2017; Dellaert et al., 1999).

In this paper a stereo camera is used as a sensor for localization of a wheeled mobile robot (WMR). The environment is sparsely scattered with artificial landmarks at known locations that can be robustly detected in each image of the stereo camera whenever they are visible in the camera field of view. In our case we use square-based markers that can be detected efficiently with the ArUco image processing approach (Garrido-Jurado et al., 2016). In order to extend the tracking range of landmarks, even when the mobile robot moves around the environment, the stereo camera is mounted on a Pan-Tilt Unit (PTU) that enables rotation of the camera around its vertical and

<sup>a</sup>  <https://orcid.org/0000-0002-2254-6069>

<sup>b</sup>  <https://orcid.org/0000-0002-1461-3321>

<sup>c</sup>  <https://orcid.org/0000-0002-0502-5376>

horizontal axis. This paper presents the models and algorithm for localization of the WMR that drives on a flat ground and observes the landmarks with a stereo camera. This is a reasonable assumption for many WMRs in indoor environment, and since we take this assumption implicitly into the model a better performance of the pose estimation can be expected. The presented approach is therefore not suitable for legged robots and robots that operate on uneven terrain.

The rest of the paper is structured as follows. In Section 2 a detailed description of the system is given, along with mathematical model of the system and calibration procedure for estimation of the parameters that describe the static transformation of the stereo camera on the PTU. Section 3 is about mobile robot localization, and it presents the stochastic modelling of the system for the purpose of localization, analysis of the system observability and localization approach. In Section 4 some conclusions are drawn.

## 2 SYSTEM DESCRIPTION

The system is shown in Fig. 1. On the WMR *Pioneer 3-AT* a PTU with a stereo camera is mounted. The mobile system is also equipped with other proximity and distance sensors, but these are not considered in this work.



Figure 1: WMR Pioneer 3-AT with a stereo camera on a PTU.

Let us introduce the coordinate frames that we use: world frame  $W$ ; robot frame  $R$  (origin is in the intersection of the robot vertical rotation axis and the ground plane); base frame  $S$ , pan frame  $U$  and tilt frame  $V$  of the PTU; left (right) camera frame  $C_1$  ( $C_2$ ); stereo camera frame  $C$  (between the left and right camera frame); left (right) image frame  $P_1$  ( $P_2$ ). The transformations between the frames are represented graphically in Fig. 2.

Static transformation that describes the pose of the

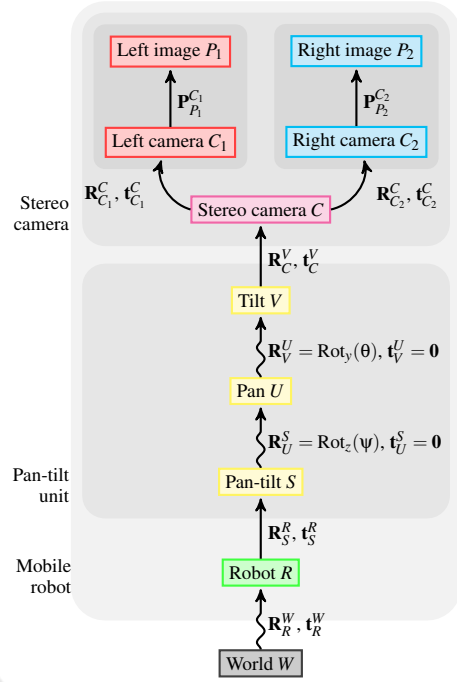


Figure 2: Diagram of the static (solid arrows) and dynamic (curly arrows) transformations between the frames.

PTU on the mobile robot is:

$$\mathbf{R}_S^R = \mathbf{I}, \quad \mathbf{t}_S^R = [0.154 \quad 0.023 \quad 0.563]^T, \quad (1)$$

where  $\mathbf{I}$  is an identity matrix.

The frames  $S$ ,  $U$  and  $V$  have a common origin (in the intersection of the pan and tilt rotation axes):

$$\mathbf{t}_U^S = \mathbf{t}_V^U = \mathbf{0},$$

where  $\mathbf{0}$  is the vector of zeros. Orientations of the frames  $U$  and  $V$ , which are dependent on the pan and tilt angles,  $\psi$  and  $\theta$ , are given in (2) and (3).

$$\mathbf{R}_U^S = \text{Rot}_z(\psi) = \begin{bmatrix} \cos \psi & -\sin \psi & 0 \\ \sin \psi & \cos \psi & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$\mathbf{R}_V^U = \text{Rot}_y(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix} \quad (3)$$

Notation  $\text{Rot}_a(\alpha)$  represents a rotation around the axis  $a$  for the angle  $\alpha$ .

In (4) the static transformations of the cameras with respect to the tip of the PTU (tilt frame  $V$ ) are given.

$$\begin{aligned} \mathbf{R}_{C_1}^V &= \mathbf{R}_{C_2}^V = \text{Rot}_z(-\frac{\pi}{2})\text{Rot}_x(-\frac{\pi}{2}) \\ \mathbf{t}_{C_1}^V &= [0.002 \quad 0.032 \quad 0.057]^T \\ \mathbf{t}_{C_2}^V &= [0.002 \quad -0.032 \quad 0.057]^T \end{aligned} \quad (4)$$

The values in (1) and (4) have been obtained using the calibration procedure described in Section 2.1.

## 2.1 System Calibration

The intrinsic parameters ( $\mathbf{P}_{C_1}^{P_1}$  and  $\mathbf{P}_{C_2}^{P_2}$ ) and the relative pose between the cameras ( $\mathbf{R}_{C_1}^{C_2}$  and  $\mathbf{t}_{C_1}^{C_2}$ ) in the stereo-camera setup can be determined using the stereo camera calibration procedure (Bouguet, 2004). Several images of the chessboard-like pattern from different poses need to be captured. The pattern need to be visible in both images. From the corresponding points in the stereo image pairs and given the known size of the pattern square, the stereo-camera parameters can be obtained. If required, the calibration procedure can be extended in a way that lens distortions are taken into account.

In our case we also need to determine the relative pose of the stereo camera with respect to the tip of the PTU, i.e.  $\mathbf{R}_C^V$  and  $\mathbf{t}_C^V$ . The relative pose of the stereo camera with respect to the tip of the PTU could be measured, but since the location of the camera frame origin and also the location of the tip of the PTU are not directly accessible, this is hard to do without an error. Therefore we would like to estimate this pose with an appropriate calibration procedure that is described next.

Given a calibrated stereo camera we observe a set of points on a static object in the environment (with respect to the base frame of the PTU) from different configurations (views) of the PTU. In the  $i$ -th configuration the pose of the tip with respect to the base of the PTU is represented with rotation  $\mathbf{R}_{V,i}^S$  (the pan angle is  $\psi_i$  and the tilt angle is  $\theta_i$ ). The triangulation approach can be used to estimate the 3D positions (in the stereo camera frame)  $\mathbf{p}_{C,i,l}$ ,  $l = 1, 2, \dots, n$ , of the observed  $n$  image points. Observing the same set of points from  $m$  different views (configurations of the PTU), the transformations  $\mathbf{R}_{C,j}^{C,0}$ ,  $\mathbf{t}_{C,j}^{C,0}$ ,  $j = 1, 2, \dots, m$ , can be obtained from the set of points in views  $i = 0$  and  $i = j$  in the following way. First the centres of the points  $\bar{\mathbf{p}}_{C,i} = \frac{1}{n} \sum_{k=1}^n \mathbf{p}_{C,i,k}$  are evaluated for each view  $i$ , and each set of the points is given with respect to its centre:

$$\tilde{\mathbf{p}}_{C,i,k} = \mathbf{p}_{C,i,k} - \bar{\mathbf{p}}_{C,i}, \quad (5)$$

for every point  $k$  in every view  $i$ . The rigid transformation  $\mathbf{p}_{C,0,k} = \mathbf{R}_{C,j}^{C,0} \mathbf{p}_{C,j,k} + \mathbf{t}_{C,j}^{C,0}$  can be determined with the minimization of the least squares error cost function  $J$ :

$$J = \sum_{k=1}^n \left\| \mathbf{p}_{C,0,k} - \mathbf{R}_{C,j}^{C,0} \mathbf{p}_{C,j,k} - \mathbf{t}_{C,j}^{C,0} \right\|. \quad (6)$$

Taking into account the centred set of points (5), the criterion (6) can be written as:

$$J = \sum_{k=1}^n \left\| \tilde{\mathbf{p}}_{C,0,k} - \mathbf{R}_{C,j}^{C,0} \tilde{\mathbf{p}}_{C,j,k} \right\|, \quad (7)$$

since  $\tilde{\mathbf{p}}_{C,0} - \mathbf{R}_{C,j}^{C,0} \tilde{\mathbf{p}}_{C,j} - \mathbf{t}_{C,j}^{C,0} = \mathbf{0}$ . The cost function (7) has minimum where the trace( $\mathbf{R}_{C,j}^{C,0} \mathbf{H}_j$ ) has maximum (Eggert et al., 1997), and  $\mathbf{H}_j$  is defined as:

$$\mathbf{H}_j = \sum_{k=1}^n \tilde{\mathbf{p}}_{C,j,k} \tilde{\mathbf{p}}_{C,0,k}^T.$$

Rotation  $\mathbf{R}_{C,j}^{C,0}$  that maximizes the aforementioned trace can be obtained from the Singular Value Decomposition (SVD) of  $\mathbf{H}_j = \mathbf{U}_j \mathbf{S}_j \mathbf{V}_j^T$ :

$$\mathbf{R}_{C,j}^{C,0} = \mathbf{V}_j \mathbf{U}_j^T.$$

If the determinant  $\det \mathbf{R}_{C,j}^{C,0}$  is  $-1$  instead of  $+1$ , the transformation represents a reflection rather than rotation. In such a case the rotation can be determined from  $\mathbf{R}_{C,j}^{C,0} = [\mathbf{v}_{j,1}, \mathbf{v}_{j,2}, -\mathbf{v}_{j,3}] \mathbf{U}_j^T$ , where the columns are obtained from the matrix  $\mathbf{V}_j = [\mathbf{v}_{j,1}, \mathbf{v}_{j,2}, \mathbf{v}_{j,3}]$ . Once the rotation is known, the translation vector  $\mathbf{t}_{C,j}^{C,0}$  can be estimated from:

$$\mathbf{t}_{C,j}^{C,0} = \tilde{\mathbf{p}}_{C,0} - \mathbf{R}_{C,j}^{C,0} \tilde{\mathbf{p}}_{C,j}.$$

More closed-form approaches that can be used to solve the presented rigid-motion estimation problem can be found in (Eggert et al., 1997).

The orientation of the PTU tip in the view  $j = 1, 2, \dots, m$  with respect to the view  $i = 0$  is  $\mathbf{R}_{V,j}^{V,0} = \mathbf{R}_S^{V,0} \mathbf{R}_{V,j}^S$ . The static orientation  $\mathbf{R}_C^V$  that is view invariant can be obtained from the set of relations ( $j = 1, 2, \dots, m$ ):

$$\mathbf{R}_{V,j}^{V,0} \mathbf{R}_C^V = \mathbf{R}_C^V \mathbf{R}_{C,j}^{C,0}. \quad (8)$$

At least two sets of (8) are required to estimate  $\mathbf{R}_C^V$ , therefore the pattern of points need to be observed from at least three different configurations of the PTU. The system of equations (8) can be written as:

$$\left( \mathbf{R}_{V,j}^{V,0} \otimes \mathbf{R}_{C,j}^{C,0} - \mathbf{I} \right) \text{vec} \left( \mathbf{R}_C^V \right) = \mathbf{0}, \quad (9)$$

where operator  $\otimes$  represents the Kronecker matrix product and  $\text{vec}(\mathbf{X})$  is a vector with stacked columns of the matrix  $\mathbf{X}$ . Several different pairs of matrices  $\mathbf{R}_{V,j}^{V,0}$  and  $\mathbf{R}_{C,j}^{C,0}$  can therefore be stacked together into the linear form (9) that can be solved using the SVD algorithm. Once the estimate of the rotation  $\mathbf{R}_C^V$  is obtained, the translation vector  $\mathbf{t}_C^V$  can be determined from the linear system of equations:

$$\left( \mathbf{R}_{V,j}^{V,0} - \mathbf{I} \right) \mathbf{t}_C^V = \mathbf{R}_C^V \tilde{\mathbf{p}}_{C,0} - \mathbf{R}_{V,j}^{V,0} \mathbf{R}_C^V \tilde{\mathbf{p}}_{C,j}. \quad (10)$$

At least two systems like (10) need to be stacked together in order to solve for the translation vector  $\mathbf{t}_C^V$ .

With an appropriate modification, the presented calibration procedure can also be used to determine the static rotation  $\mathbf{R}_S^R$  and translation  $\mathbf{t}_S^R$  if the pose of the mobile robot in the world coordinate frame can be measured.

### 3 LOCALIZATION

The problem of localization is about estimation of the transformation between the world and robot frame ( $\mathbf{R}_R^W$  and  $\mathbf{t}_R^W$  in Fig. 2). In our case we would like to achieve this based on fusion of wheel odometry and measurement of multiple landmarks at known locations with a stereo camera. If the WMR motion is constrained to the ground plane, the pose of the robot with respect to the world can be given with the generalized coordinates  $\mathbf{q}^T(k) = [x(k) \ y(k) \ \varphi(k)]$ :

$$\mathbf{R}_R^W = \begin{bmatrix} \cos \varphi(k) & -\sin \varphi(k) & 0 \\ \sin \varphi(k) & \cos \varphi(k) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{t}_R^W = \begin{bmatrix} x(k) \\ y(k) \\ 0 \end{bmatrix},$$

where the coordinates  $x(k)$  and  $y(k)$  represent position and the coordinate  $\varphi(k)$  is orientation of the robot. This is a reasonable assumption for many indoor spaces since we assume that the mobile system drives on a flat surface and that it does not tilt significantly. We also assume that the poses of all the static and also moving sensors on-board of the mobile system are known at all times. This means that both states of the PTU are measurable. All the static transformations can be determined with the procedure presented in the previous section. In the following subsection the modelling of the system with uncertainties is presented that takes into account different simplifications. Since the mobile robot only moves along the ground plane, the estimation problem is also converted to a pure 2D estimation problem. In this paper we also assume that the PTU is not moving during the localization, i. e. the PTU is in its home position all the time.

#### 3.1 System Modelling

##### 3.1.1 Wheeled Mobile Robot

The WMR has a differential drive (the wheels on each side are driven jointly). The kinematic model of the differential drive in the discrete form with the sampling time  $T$  can be written as:

$$\mathbf{q}(k+1) = \mathbf{q}(k) + \begin{bmatrix} Tv(k) \cos \varphi(k) \\ Tv(k) \sin \varphi(k) \\ T\omega(k) \end{bmatrix},$$

where  $v(k)$  and  $\omega(k)$  are the WMR linear and angular velocity, respectively.

Wheels on each side of the mobile robot are equipped with incremental encoders, therefore the odometry can be implemented. Given the true relative encoder readings between one sample time for the left and right wheel  $\mathbf{u}^T(k) = [\Delta\lambda_L(k) \ \Delta\lambda_R(k)]$ , we assume

the following odometry model in the discrete form:

$$\mathbf{q}(k+1) = \mathbf{q}(k) + g_\lambda \begin{bmatrix} \frac{\Delta\lambda_R(k)+w_R(k)+\Delta\lambda_L(k)+w_L(k)}{2} \cos \varphi(k) \\ \frac{\Delta\lambda_R(k)+w_R(k)+\Delta\lambda_L(k)+w_L(k)}{2} \sin \varphi(k) \\ \frac{\Delta\lambda_R(k)+w_R(k)-\Delta\lambda_L(k)-w_L(k)}{L} \end{bmatrix}, \quad (11)$$

where  $g_\lambda$  is the constant that converts the encoder readings into relative distances and  $L$  is the distance between the wheels — both parameters are normally determined with calibration. The encoder readings include a normally distributed white noise  $\mathbf{w}(\mathbf{k}) = [w_L(k) \ w_R(k)]$  with zero mean. The covariance of the noise  $\mathbf{w}(\mathbf{k})$  is assumed to be increasing with the magnitude of the encoder reading, which coincides with the speed of the wheels, according to the model (Teslić et al., 2011):

$$\mathbf{Q}_w(k) = \begin{bmatrix} \Delta\lambda_L^2(k)\sigma_{wL}^2 & 0 \\ 0 & \Delta\lambda_R^2(k)\sigma_{wR}^2 \end{bmatrix}.$$

##### 3.1.2 Stereo Camera

We assume that we have a calibrated stereo camera system in canonical configuration with the baseline distance  $B$  and focal length of each camera  $f$ . The origin of each camera image plane coincides with the camera optical center and the axis  $x$  of the left image frame is colinear with the axis  $x$  in the right image frame. The projections of a single point in the 3-D space to the left and right camera image plane are  $(x_l, y_l)$  and  $(x_r, y_r)$ , respectively. Since the positions of landmarks are measured in each image with a marker detection algorithm (Garrido-Jurado et al., 2016), let us assume that the aforementioned variables have normal distribution. They can be gathered in a vector  $\mathbf{s}$ :

$$\mathbf{s} = [x_l \ y_l \ x_r \ y_r]^T \in \mathcal{N}(\bar{\mathbf{s}}, \mathbf{Q}_s), \quad (12)$$

where  $\bar{\mathbf{s}}$  is the expected value of the measurement vector  $\mathbf{s}$  and  $\mathbf{Q}_s$  is the measurement noise covariance matrix. The shape of the covariance matrix  $\mathbf{Q}_s$  is assumed to be block diagonal, since the measurement noises of the left and right camera are independent of each other (shaking of the stereo camera rig is not considered here):

$$\mathbf{Q}_s = \begin{bmatrix} \sigma_{x_l}^2 & \sigma_{x_l, y_l} & 0 & 0 \\ \sigma_{x_l, y_l} & \sigma_{y_l}^2 & 0 & 0 \\ 0 & 0 & \sigma_{x_r}^2 & \sigma_{x_r, y_r} \\ 0 & 0 & \sigma_{x_r, y_r} & \sigma_{y_r}^2 \end{bmatrix}. \quad (13)$$

In (13) the diagonal elements represent the variances and the off-diagonal elements are the covariances.

Let us represent the position of a point in the 3D camera frame. The origin of the stereo camera frame

$C$  is in the middle of the line that connects the camera focal points,  $z_C$ -axis is perpendicular to the image plane and it is pointing outwards to the scene,  $y_C$ -axis is parallel to the vertical image axis and  $x_C$ -axis is parallel to the horizontal image axis, pointing in the direction that makes the camera frame right handed. Here we introduce a special coordinate vector  $\mathbf{r}$  for presentation of the point in the stereo camera frame:

$$\mathbf{r}^T = [\tan \alpha \quad \tan \beta \quad z_C] = \begin{bmatrix} x_C \\ y_C \\ z_C \end{bmatrix}. \quad (14)$$

This representation can uniquely represent only the half-space in front of the camera ( $z_C > 0$ ) and it is therefore suitable only for the cameras that have the field-of-view smaller than  $\pi$ .

The position of the point  $\mathbf{r}$  in the camera frame can be determined from the projection vector (12) using triangulation:

$$\mathbf{r}^T = \begin{bmatrix} x_l+x_r & y_l+y_r & Bf \\ 2f & 2f & x_l-x_r \end{bmatrix}.$$

The variance of this estimate is therefore:

$$\mathbf{Q}_r = \begin{bmatrix} \frac{\sigma_{x_l}^2 + \sigma_{x_r}^2}{4f^2} & \frac{\sigma_{x_l y_l} + \sigma_{x_r y_r}}{4f^2} & -z_C \frac{\sigma_{x_l}^2 - \sigma_{x_r}^2}{2Bf^2} \\ \frac{\sigma_{x_l y_l} + \sigma_{x_r y_r}}{4f^2} & \frac{\sigma_{y_l}^2 + \sigma_{y_r}^2}{4f^2} & -z_C \frac{\sigma_{x_l y_l} - \sigma_{x_r y_r}}{2Bf^2} \\ -z_C \frac{\sigma_{x_l}^2 - \sigma_{x_r}^2}{2Bf^2} & -z_C \frac{\sigma_{x_l y_l} - \sigma_{x_r y_r}}{2Bf^2} & z_C \frac{\sigma_{x_l}^2 + \sigma_{x_r}^2}{B^2 f^2} \end{bmatrix}.$$

If we assume that the measurement noises of the left and right camera have the same properties, the following equivalence of variances and covariances holds:  $\sigma_{x_l}^2 = \sigma_{x_r}^2$  and  $\sigma_{x_l y_l} = \sigma_{x_r y_r}$ ; and the covariance matrix  $\mathbf{Q}_r$  becomes block diagonal. If we further assume that the covariances are zero ( $\sigma_{x_l y_l} = \sigma_{x_r y_r} = 0$ ), the covariance matrix  $\mathbf{Q}_r$  becomes diagonal, where the diagonal elements are:

$$\begin{bmatrix} \sigma_{\tan \alpha}^2 & \sigma_{\tan \beta}^2 & \sigma_{z_C}^2 \end{bmatrix} = \begin{bmatrix} \frac{\sigma_{x_l}^2}{2f^2} & \frac{\sigma_{y_l}^2}{2f^2} & z_C \frac{2\sigma_{x_l}^2}{B^2 f^2} \end{bmatrix}. \quad (15)$$

From (14) the position of the point in the stereo camera frame  $\mathbf{p}_C$  can be obtained:

$$\mathbf{p}_C^T = [x_C \quad y_C \quad z_C] = [z_C \tan \alpha \quad z_C \tan \beta \quad z_C].$$

Under the assumption (15), the measurement covariance matrix  $\mathbf{Q}_C$  of the vector  $\mathbf{p}_C$  in the stereo camera frame is:

$$\mathbf{Q}_C = \begin{bmatrix} z_C^2 \sigma_{\tan \alpha}^2 + (\tan \alpha)^2 \sigma_{z_C}^2 & \tan \alpha \tan \beta \sigma_{z_C}^2 & \tan \alpha \sigma_{z_C}^2 \\ \tan \alpha \tan \beta \sigma_{z_C}^2 & z_C^2 \sigma_{\tan \beta}^2 + (\tan \beta)^2 \sigma_{z_C}^2 & \tan \beta \sigma_{z_C}^2 \\ \tan \alpha \sigma_{z_C}^2 & \tan \beta \sigma_{z_C}^2 & \sigma_{z_C}^2 \end{bmatrix}. \quad (16)$$

The transformation of the point  $\mathbf{p}_C$  in the stereo camera frame to the point  $\mathbf{p}_R$  in the mobile robot base frame is  $\mathbf{p}_R = \mathbf{R}_C^R \mathbf{p}_C + \mathbf{t}_C^R$  as defined in Section 2. The covariance (16) can also be transformed to the mobile robot base frame:

$$\mathbf{Q}_R = \mathbf{R}_C^R \mathbf{Q}_C (\mathbf{R}_C^R)^T.$$

Since we are dealing with estimation of the mobile robot that can only move in the ground plane, we can make the projection of the 3D points to the ground plane ( $z_R = 0$ ), i.e. the point in the 2D ground plane is therefore  $\mathbf{p}_G = \mathbf{P}_R^G \mathbf{p}_R$ , where  $\mathbf{P}_R^G$  is the parallel projection matrix:

$$\mathbf{P}_R^G = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The covariance matrix  $\mathbf{Q}_G$  that is the projection of the covariance matrix  $\mathbf{Q}_R$  can be given as:

$$\mathbf{Q}_G = \mathbf{P}_R^G \mathbf{Q}_R (\mathbf{P}_R^G)^T. \quad (17)$$

In the configuration when the stereo camera is in its home position, in which case the camera axis  $z_C$  is aligned with the robot axis  $x_R$  and the camera axis  $x_C$  is in the opposite direction of the robot axis  $x_R$ , the  $\mathbf{p}_G$  simplifies to:

$$\mathbf{p}_G^T = \begin{bmatrix} Bf \\ x_l - x_r \end{bmatrix} + x_C^R \quad -\frac{B}{2} \frac{x_l + x_r}{x_l - x_r} + y_C^R, \quad (18)$$

where  $x_C^R$  and  $y_C^R$  are the first and the second element of the translation vector  $\mathbf{t}_C^R$ , respectively. The covariance matrix (17) in this particular case is:

$$\mathbf{Q}_G = \begin{bmatrix} \sigma_{z_C}^2 & -\tan \alpha \sigma_{z_C}^2 \\ -\tan \alpha \sigma_{z_C}^2 & z_C^2 \sigma_{\tan \alpha}^2 + (\tan \alpha)^2 \sigma_{z_C}^2 \end{bmatrix}. \quad (19)$$

The covariance matrix (19) is strongly dependent on the distance to the measured point due to the fact that  $\sigma_{z_C} \propto z_C^2$ . Top row in Figure 3 shows covariances (19) as ellipses for different positions of the landmarks overlaid over simulated noisy data. The proposed noise model was verified with real measurements (bottom row in Figure 3) by temporal observation of landmarks at various positions from a static WMR.

The (18) represents the measurement of the point in the ground plane. This measurement can also be given in polar coordinates  $\mathbf{z}^T = [d \ \theta]$ :

$$\mathbf{z}^T = \left[ \sqrt{x_G^2 + y_G^2} \quad \arctan \frac{y_G}{x_G} + j\pi \right], \quad j \in \{0, 1\}.$$

The stereo camera sensor can therefore be considered as the sensor that measures the distance and angle to the landmark. In our case we considered that the vision system can also determine the ID of the measured landmark, therefore we can distinguish between different landmarks. Let us express the measurement of the landmark with the system state vector  $\mathbf{q}(k)$ :

$$\mathbf{z}(k) = \begin{bmatrix} d(k) \\ \theta(k) \end{bmatrix} = \begin{bmatrix} \sqrt{(x_m - x(k))^2 + (y_m - y(k))^2} \\ \arctan \frac{y_m - y(k)}{x_m - x(k)} - \varphi(k) + j(k)\pi \end{bmatrix}, \quad (20)$$

where  $x_m$  and  $y_m$  are the coordinates of the landmark in the world frame.

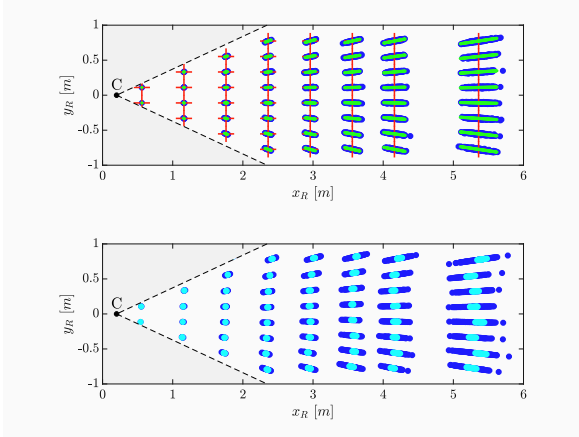


Figure 3: Top, simulation of the stereo camera measurement noise for multiple landmarks that are in the stereo camera  $C$  field of view (red cross – real landmark position, blue – simulated noise, green – noise covariance ellipse). Bottom, real measurements (cyan – static camera, blue – small shaking of the camera).

### 3.2 System Observability

In order to estimate system states they need to be observable. The observability of the non-linear system can be evaluated based on the analysis of indistinguishable states. Briefly, two states are indistinguishable if for every input on a finite time interval, identical outputs are obtained. A system is observable if the set of all the indistinguishable states of a state  $\mathbf{x}$  contains only the state  $\mathbf{x}$  for every state  $\mathbf{x}$  in the domain of definition. For more detailed definition see (Hermann and Krener, 1977).

Let us evaluate the observability of the system in the case of only one landmark graphically. In Figure 4 we can see three robots in different poses that all measured the same distance and angle to the landmark. Even if the robots move along any admissible trajectory (e.g. along the circle) there is no way to distinguish between different robot poses based on measured outputs. The pose of the robot  $\mathbf{q}$  is clearly not observable if only a single landmark is used. In a similar way it is not hard to observe that the system is observable if two or more landmarks are used, since we have assumed that the IDs of the landmarks are also known. If the later would not hold, three or more landmarks would be required.

Now we introduce a new state vector  $\mathbf{q}_p^T(k) = [d(k) \ \theta(k)]$  that has all the states measurable directly. The kinematic model of this system is:

$$\mathbf{q}_p(k+1) = \mathbf{q}_p(k) + \begin{bmatrix} -T v(k) \cos \theta(k) \\ T \omega(k) + T v(k) \frac{\sin \theta(k)}{d(k)} \end{bmatrix}. \quad (21)$$

The system (21) gives only a partial information about

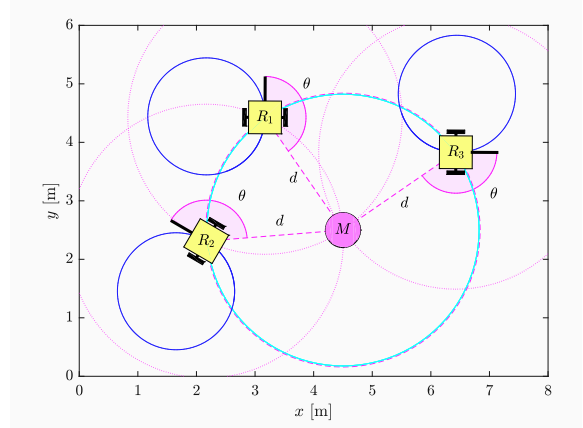


Figure 4: Three poses of the robots that are indistinguishable from the outputs in the case of a single landmark  $M$ .

the system pose (a subspace of possible robot poses), but the system is completely observable already in the case of a single landmark.

### 3.3 Extended Kalman Filter

We use Extended Kalman Filter (EKF) to solve the localization problem, which consists of a prediction and a correction step. The pose of the mobile robot  $\mathbf{q}(k)$  could be estimated directly, if the model (11) would be used in the prediction step of the EKF and the correction step would be made based on the measurements (20) to all the visible landmarks.

In this paper we used a different approach, using multiple partial estimators that estimate the states  $\mathbf{q}_{p,m}(k)$  for each visible landmark  $m$  independently. Each partial estimator uses the model (21) in the prediction step of the EKF:

$$\begin{aligned} \hat{\mathbf{q}}_{p,m}(k+1) &= \mathbf{f}_{p,m}(\mathbf{q}_{p,m}(k), \mathbf{u}(k)), \\ \hat{\mathbf{P}}_{p,m}(k+1) &= \mathbf{A}_{p,m}(k) \mathbf{P}_{p,m}(k) \mathbf{A}_{p,m}^T(k) + \\ &\quad + \mathbf{F}_{p,m}(k) \mathbf{Q}_{p,m}(k) \mathbf{F}_{p,m}^T(k), \end{aligned}$$

where  $\mathbf{A}_{p,m}(k) = \frac{\partial \mathbf{f}_{p,m}}{\partial \mathbf{q}_{p,m}} \Big|_k$  and  $\mathbf{F}_{p,m}(k) = \frac{\partial \mathbf{f}_{p,m}}{\partial \mathbf{w}} \Big|_k$ . The measurement (20) of the associated landmark is then used in the correction of each partial estimator:

$$\begin{aligned} \mathbf{L}_{p,m}(k) &= \mathbf{C}_{p,m}(k) \hat{\mathbf{P}}_{p,m}(k) \mathbf{C}_{p,m}^T(k) + \mathbf{R}_{p,m}(k) \\ \mathbf{K}_{p,m}(k) &= \hat{\mathbf{P}}_{p,m}(k) \mathbf{C}_{p,m}^T(k) \mathbf{L}_{p,m}^{-1}(k) \\ \mathbf{q}_{p,m}(k) &= \hat{\mathbf{q}}_{p,m}(k) + \mathbf{K}_{p,m}(k) (\mathbf{z}_{p,m}(k) - \hat{\mathbf{z}}_{p,m}(k)), \\ \mathbf{P}_{p,m}(k) &= \hat{\mathbf{P}}_{p,m}(k) - \mathbf{K}_{p,m}(k) \mathbf{C}_{p,m}(k) \hat{\mathbf{P}}_{p,m}(k), \end{aligned}$$

where  $\mathbf{C}_{p,m}(k) = \frac{\partial \mathbf{h}_{p,m}}{\partial \mathbf{q}_{p,m}} \Big|_k$ . The  $\mathbf{Q}_{p,m}(k)$  and  $\mathbf{R}_{p,m}(k)$  in the EKF are defined according to the covariance model developed in Section 3.1. In this way we do not estimate the pose of the mobile robot, but only the

distances and headings of each landmark with respect to the WMR. The estimates  $\mathbf{q}_{p,m}(k)$ ,  $m = 1, 2, \dots, M$  are then merged into the estimate of the state  $\mathbf{q}(k)$ . This can be achieved using a triangulation (or trilateration) approach, e.g. using the approach presented in (Betke and Gurvits, 1997).

In the simulation we have used three landmarks and the limited field of view of the camera was not taken into account. Figures 5 to 7 present the results of all three partial estimators. It can be seen that the distance measurement noise is largely dependant on the distance to the landmark, but the output of each partial estimator converges to the true value. Based on the outputs of these estimators, the pose of the mobile robot is calculated as an intersection of all three solutions (circles) in the least squares sense (Figure 8). The pose estimate is shown in Figure 9.

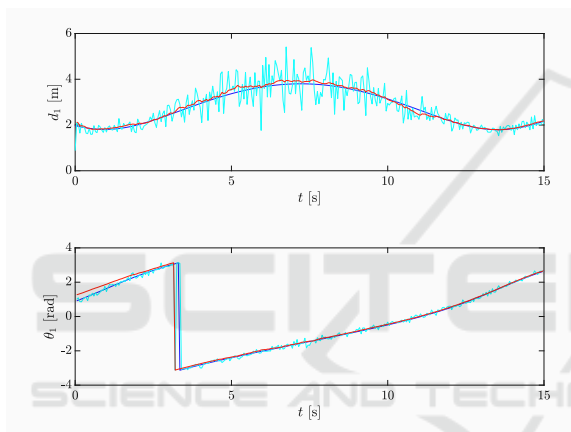


Figure 5: Estimation of the landmark 1 state vector  $\mathbf{q}_p$  (blue – true, red – estimate, cyan – measurement).

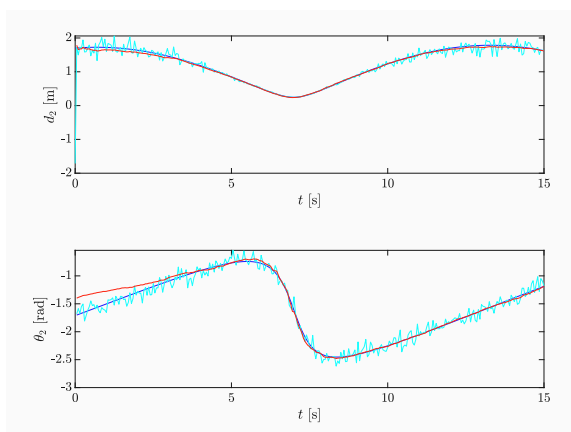


Figure 6: Estimation of the landmark 2 state vector  $\mathbf{q}_p$  (blue – true, red – estimate, cyan – measurement).

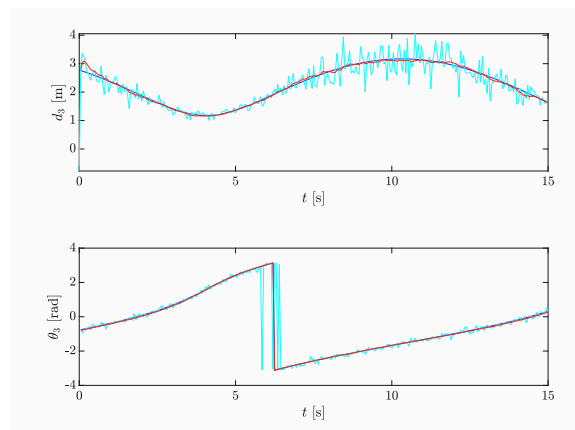


Figure 7: Estimation of the landmark 3 state vector  $\mathbf{q}_p$  (blue – true, red – estimate, cyan – measurement).

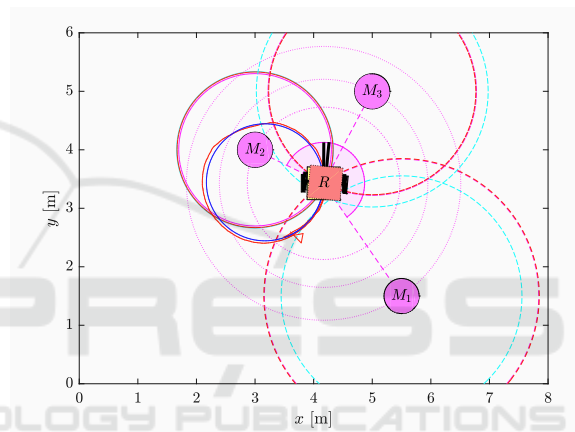


Figure 8: Estimation of the robot pose  $\mathbf{q}$  at the end of time (solid blue – true pose, solid red – estimated pose, dashed cyan – measurements, dashed red – estimated measurements).

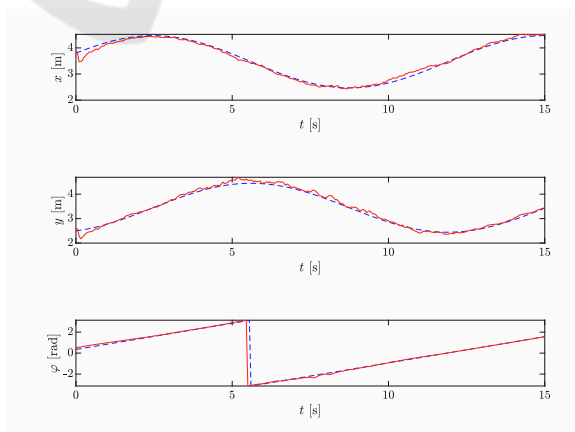


Figure 9: Estimation of the robot state vector  $\mathbf{q}$  (blue – true, red – estimate).

## 4 CONCLUSION

Calibration of the system parameters is essential for optimal performance of the localization algorithms. We have showed a calibration procedure that can be used to determine static transformation between the camera and PTU coordinate frame. The procedure requires only observation of the same set of points from three different configurations of the PTU. This calibration procedure is simple to deploy on a real setting, without any special preparation of the environment solely for the calibration purposes.

In the development of the localization algorithm system uncertainties were taken into account. Introducing the flat ground surface constraint, the considered localization problem can be solved in the two dimensional plane where the camera measures distances and angles to the visible landmarks. Assuming normally distributed noise in image measurement of landmark points, the standard deviation increases predominantly in the direction of the image ray, proportionally to the squared distance from the camera.

We have shown an approach that uses multiple partial Kalman filters, where each filter estimates only the distance and heading to the particular landmark, and the outputs of these estimators are later used to calculate the robot pose. Results demonstrate that this is a feasible approach that converges to the true pose of the mobile robot. The benefit of this approach is computational efficiency, since the covariance matrices are low dimensional. To reduce memory consumption during large-area localization, the landmarks that have not been observed for a long time can be made forgotten. The presented models are also valid when the camera is moving. The proposed system can be augmented with a control for tracking of the nearest visible landmarks to reduce the time when no landmarks are in the stereo camera field of view.

## ACKNOWLEDGEMENTS

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0219).

## REFERENCES

- Agrawal, M. and Konolige, K. (2006). Real-time localization in outdoor environments using stereo vision and inexpensive GPS. In *18th Int. Conf. on Pattern Recognition*, volume 3, pages 1063–1068.
- Betke, M. and Gurrts, L. (1997). Mobile robot localization using landmarks. *IEEE transactions on robotics and automation*, 13(2):251–263.
- Bouguet, J.-Y. (2004). Camera calibration toolbox for Matlab. [online] Available at: [http://www.vision.caltech.edu/bouguetj/calib\\_doc](http://www.vision.caltech.edu/bouguetj/calib_doc) [Accessed April 2019].
- Chen, S. Y. (2012). Kalman filter for robot vision: a survey. *IEEE Trans. on Industrial Electronics*, 59(11):4409–4420.
- Dellaert, F., Burgard, W., Fox, D., and Thrun, S. (1999). Using the CONDENSATION algorithm for robust, vision-based mobile robot localization. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 588–594.
- Du, X. and Tan, K. K. (2016). Comprehensive and practical vision system for self-driving vehicle lane-level localization. *IEEE Trans. on Image Processing*, 25(5):2075–2088.
- Eggert, D. W., Lorusso, A., and Fisher, R. B. (1997). Estimating 3-D rigid body transformations: a comparison of four major algorithms. *Machine vision and applications*, 9(5–6):272–290.
- Fischer, T., Pire, T., Ěřžek, P., Cristóforis, P. D., and Faigl, J. (2016). Stereo vision-based localization for hexapod walking robots operating in rough terrains. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 2492–2497.
- Fuentes-Pacheco, J., Ruiz-Ascencio, J., and Rendón-Mancha, J. M. (2015). Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81.
- Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., and Medina-Carnicer, R. (2016). Generation of fiducial marker dictionaries using mixed integer linear programming. *Pattern Recognition*, 51:481–491.
- Hermann, R. and Krener, A. (1977). Nonlinear controllability and observability. *IEEE Trans. on Automatic Control*, 22(5):728–740.
- Kim, H., Liu, B., Goh, C. Y., Lee, S., and Myung, H. (2017). Robust vehicle localization using entropy-weighted particle filter-based data fusion of vertical and road intensity information for a large scale urban area. *IEEE Robotics and Automation Letters*, 2(3):1518–1524.
- Konolige, K. and Agrawal, M. (2008). FrameSLAM: From bundle adjustment to real-time visual mapping. *IEEE Trans. on Robotics*, 24(5):1066–1077.
- Mei, C., Sibley, G., Cummins, M., Newman, P., and Reid, I. (2011). RSLAM: A system for large-scale mapping in constant-time using stereo. *International journal of computer vision*, 94(2):198–214.
- Piasco, N., Marzat, J., and Sanfourche, M. (2016). Collaborative localization and formation flying using distributed stereo-vision. In *IEEE Int. Conf. on Robotics and Automation*, pages 1202–1207.
- Se, S., Lowe, D., and Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In *IEEE Int. Conf. on Robotics and Automation*, volume 2, pages 2051–2058.
- Teslić, L., Škrjanc, I., and Klančar, G. (2011). EKF-based localization of a wheeled mobile robot in structured environments. *Journal of Intelligent and Robotic Systems*, 62:187–203.