ROBUST SCALE ESTIMATION FOR MONOCULAR VISUAL ODOMETRY USING STRUCTURE FROM MOTION AND VANISHING POINTS

Johannes Gräter[1], Tobias Schwarze[1], Martin Lauer[1]*

August 20, 2015

Abstract

While monocular visual odometry has been widely investigated, one of its key issues restrains its broad appliance: the scale drift. To tackle it, we leverage scene inherent information about the ground plane to estimate the scale for usage on Advanced Driver Assistance Systems. The algorithm is conceived so that it is independent of the unscaled ego-motion estimation, augmenting its adaptability to other frameworks. A ground plane estimation using Structure From Motion techniques is complemented by a vanishing point estimation to render our algorithm robust in urban scenarios. The method is evaluated on the KITTI dataset, outperforming state of the art algorithms in areas where urban scenery is dominant.

1 INTRODUCTION AND RELATED WORK

In the digital era camera systems are omnipresent. Especially monocular camera systems are used in robotics and in particular in Advanced Driver Assistance Systems (ADAS), since they are cheap and mechanically robust. A common goal is the estimation of the ego-motion of an agent on which the camera is mounted. Therefore, monocular visual odometry (MonoVO) and monocular Simultaneous Localization and Mapping (SLAM) have been in the focus of image processing during a long period and are still subject of current research. Even though a grand variety of algorithms has been developed for many different applications, one crucial factor is omitted frequently: the scale. Monocular systems underlie the restriction that they cannot observe the absolute depth of a scenery in contrast to stereo camera systems. Nonetheless, the ego-motion and the surroundings can be perceived by motion stereo up to one degree of freedom which is called the scale of the scene. Knowing the scale, the trajectory of the camera-movement and the surroundings are fully recovered, using only a single, inexpensive

*[1] with the Institute of Measurement and Control, Karlsruhe Institute of Technology, Karlsruhe, Germany; johannes.graeter@kit.edu; tobias.schwarze@kit.edu; martin.lauer@kit.edu

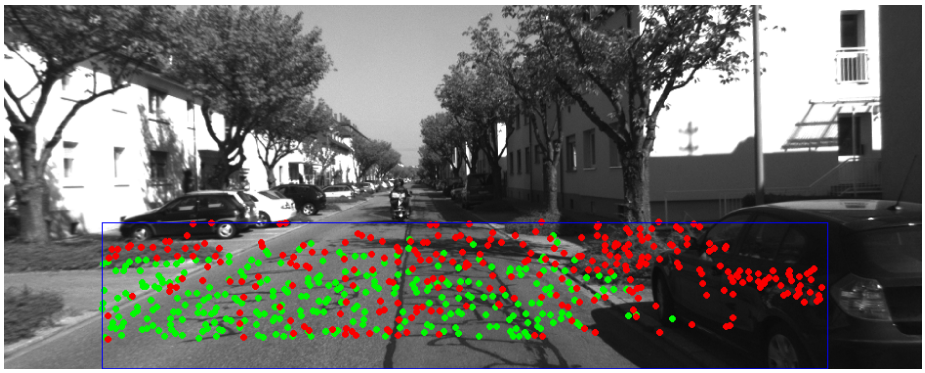


Figure 1: Reconstructed points from two consecutive frames after outlier rejection. Inliers belonging to the ground plane are marked in green, outliers in red.

sensor. The perception of the path of the vehicle and its surroundings, allows ADAS to help the driver to avoid collisions. Therefore, MonoVO is a highly interesting technology for ADAS, since the trajectory can be estimated without the need of stereoscopic systems or laser scanners.

MonoVO and SLAM were first introduced in the field of virtual reality [11, 3]. Here a simple yet effective way of determining the scale is to command the user to indicate it. Such is done in one of the first real-time applications of MonoVO on affordable sensor systems, established by Klein et al. [10], where the user is asked to move the camera by a certain distance on system startup. Thereupon, the scale is tracked to minimize the drift. More recent algorithms catch up this principle and develop it, as for example the work of Engel et al. [1]. Using sparse optical flow to calculate the trajectory of a moving camera, they reconstruct a detailed map of its surroundings. Even on a smart-phone, complex virtual reality applications were realized [12]. In this method, sophisticated tracking and bundling techniques are used to reduce the scale drift and to establish a coherent trajectory. For paths with frequent loop closures, this leads to very accurate trajectories and therefore impressive reconstructions of the surroundings of the camera. However, this method is not appropriate for the use on vehicles for which trajectories are typically loop-free. Moreover, the setup of an ADAS does not allow the input of the scale by the user.

Another trend in the robotic community are small Unmanned Aerial Vehicles (UAVs). In this area, monocular visual odometry is particularly advantageous since cameras are small, cheap and robust. Forster et al. [2] use also optical flow to estimate the trajectory of a UAV. Their algorithm leads to remarkable results for sceneries in which points of interest are seen during a long period, resulting in very low scale drift. The scale is initiated by the starting position of the UAV and then tracked. However, they highly rely on long tracks of characteristic scene points. This renders this method inadequate for the usage on automobiles, where large scene flow dominates the sceneries.

In order to not only reduce the scale drift but eliminate it completely, we aim to calculate the scale frame by frame. This principle was introduced by Geiger et al. [7] tailored for the use in autonomous driving. Using a custom image feature extraction

and matching, combined with an eight-point-algorithm they get astonishingly precise results for the unscaled ego-motion estimation, without any feature tracking. In order to extract the scale, their method fits a plane to reconstructed points in order to obtain the scale by the aid of an a-priori known height over ground of the camera. They use the additional constraint, that the angle between the camera and the ground plane is constant and calibrated. Song et al. recently proposed an extension [13], [14], using additional scene inherent information, i.e. the size of cars detected in the camera image. Moreover, they train a weighting methodology by the aid of the KITTI dataset [5], [6], [4]. While their approach is very accurate on the KITTI dataset, the presence of cars in the image can in general not be guaranteed. Furthermore the constraint of a fixed pitch- and roll-angle of the camera to the ground plane, used by Geiger et al. and Song et al., restricts the general use of these methods. Being a valuable assumption for the KITTI dataset, which was established in flat environments, on a vehicle with low pitch rates, this constraint is violated for vehicles that are less rigid, such as the driver cabin of a truck or motorcycles.

Therefore, we propose a method for estimating the scale of a camera, that is mounted on a vehicle. The scaling algorithm shall be independent of the unscaled ego-motion estimation, which allows us to benefit from the grand variety of ego-motion estimation algorithms available. Moreover, we preserve generality for our method by releasing the constraint of a fixed angle between the camera and the ground plane. Thereto we combine the estimation of vanishing directions with Structure From Motion techniques in order to estimate the ground plane. As a result, we outperform state of the art algorithms in urban areas.

2 CONVENTIONS

Throughout this paper we use the following constraints. A plane in 3d space is expressed by its normal $n \in \mathbb{R}^3$ and a scalar $d \in \mathbb{R}$ which describes the distance of a point $x \in \mathbb{R}^3$ to the plane by

$$n^T x = d, \quad ||n||_2 = 1. \tag{1}$$

Throughout this paper the camera coordinate system x, y, z is chosen so that z points out of the image plane, y points from the camera to the ground and x points to the right-hand-side of the camera.

Introducing the camera coordinates u, v for the columns and rows of a camera respectively, x coincides with u, and y coincides with v. The origin of the coordinate system lies in the left, upper corner of the image.

3 SCALE ESTIMATION

In this section, we develop a novel algorithm for scale estimation in urban environments. Thereto, we choose an existing, efficient algorithm for the unscaled ego-motion estimation as a basis, i.e. "StereoScan" [7]. In this method the estimation of the trajectory up to scale is executed by the eight-point algorithm [9] and a Random Sample

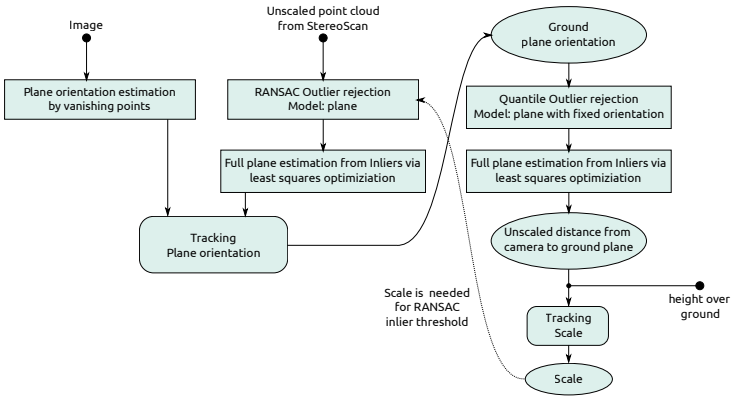


Figure 2: Flowchart of the plane estimation. Estimation algorithms are boxed by a rectangle, estimated variables are shown in ellipses. The tracking blocks are shown in boxes with round corners. A detailed description is given in section 3.1.

Consensus (RANSAC)-based outlier rejection scheme [8]. Using savvy image features, this method leads to a precise, unscaled frame to frame ego-motion.

3.1 Overview

Our algorithm is illustrated in Fig. 2. Given the motion between two frames and the corresponding two dimensional feature matches we can reconstruct points in 3d space by Structure From Motion algorithms [8]. To this point cloud, we apply a RANSAC-based outlier rejection followed by a Least Squares optimization on the inliers, which results in a plane orientation estimate. In parallel, the orientation of the ground plane is estimated by using vanishing points. The two orientations are then fused and tracked by a Kalman Filter [15]. This procedure is explained in detail in section 3.2.

The tracked orientation is used to refine the plane with a quantile-based outlier rejection scheme followed by another Least Squares plane fit. Note that the scale has to be fed back to the RANSAC-based outlier rejection in order to determine its inlier threshold.

Finally, we estimate the scale of the scenery. Since we have a-priori knowledge about the height H over ground of the camera, the scale s is computed by

$$s = \frac{H}{d}. \tag{2}$$

An additional tracking step is inserted to track the scale, as exemplified in section 3.3.

Since sceneries with short occlusion of the ground plane are bridged by the tracking, the decoupling of the plane orientation estimation and the plane distance estimation is particularly advantageous.

Throughout this work we assume that H is constant. Hence we neglect height changes caused by the vehicle's suspension, which is valid given that the camera is

mounted at a height of more than $1m$ and therefore the derivation from the installation height while driving is smaller than 5% on conventional vehicles.

3.2 Plane Orientation Extraction

We combine two methods, in order to obtain the orientation of the ground plane:

1. Fitting a plane to the reconstructed point cloud.
2. Deducing the plane normal from vanishing points in the image of a calibrated camera.

These methods are complementary. In rural areas, where the vanishing directions are challenging to be calculated correctly, the plane computation from reconstructed points leads to an accurate plane estimate. In contrast, the vanishing point estimation performs best in urban areas, where the scene structure is more dense and the ground plane is more likely to be occluded.

A RANSAC-based outlier rejection scheme is followed by a Least Squares fit, to estimate the plane parameters from the reconstructed points. Hereby we have to use the scale of the previous estimation step for determining it, since the inlier threshold of the outlier rejection depends on the current scale. Since the scale of the scene is only dependent on the velocity of the vehicle, we can assume that it changes smoothly. Consequently we can use a previous scale estimation for determining the inlier threshold.

In order to estimate the vanishing directions of the scene, we extract edges in the camera image using canny edge detection with a subsequent line fitting procedure. In an iterative process we can first determine the set of lines which supports a vanishing direction. Secondly, we evaluate the error in the orientation of the line with respect to the vanishing point and optimize the vanishing direction. The method is iterated and converges after few iterations. To initialize the vanishing direction, we use the ground plane estimate in the very first image. In subsequent frames we predict the vanishing direction using the estimated ego-motion. Note, that since only the rotational part of the motion is required, the unknown scale has no influence here. An example with three vanishing directions is shown in Fig. 3. In this work we only use the vertical vanishing direction colored in red, since we are interested in the ground plane orientation.

To fuse the result of both methods, a Kalman Filter is used to track the Euler angles α, and β of the plane. Since the plane orientation changes smoothly, we assume α and β to be constant with small uncertainties. The angles of the plane are dependent on the scene structure and are therefore uncorrelated. Additionally a gating of the angles removes erroneous measurements, so that only valid measurements are updated.

3.3 Plane Refinement

The tracked orientation is used to fit a plane to the reconstructed points, now having only one degree of freedom, i.e. the distance d of the ground plane to the origin. A problem of the RANSAC-based outlier rejection in this application is that the inlier threshold in metric space is not constant in scale space. As a result the scale space

Figure 3: Camera image used for calculating vanishing points. Green lines correspond to the green vanishing point at the horizon. Red lines correspond to the lower vanishing points and blue lines to the left vanishing point, which lie outside of the image. In our approach only the lower vanishing point, which is defined by the red lines, comes to use. The normal of the ground plane is calculated robustly, knowing the lower vanishing points and the calibration of the camera.

is stretched or compressed, dependent on the velocity of the vehicle. Therefore, we propose to use a trimmed Least Squares approach:

1. Given the plane normal, we can compute the distances d_i of the planes described by the reconstructed points P_i.
2. d_i are sorted and the points nearer than the 40% quantile, as well as the points further than the 90% quantile are removed.
3. A plane is fitted with a Least Squares approach to the inliers, revealing the distance of the ground plane to the camera.

A reconstructed point cloud after outlier rejection is shown as a re-projection on the corresponding image in Fig. 1.

At last, the distance of the ground plane and its drift are tracked by a Kalman Filter with the state equation

$$x_{k+1} = g(x) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} x_k. \tag{3}$$

The state x consists of the scale and the scale drift respectively. We assume a constant scale drift from frame to frame and model its variations by the state uncertainties. The measurement equation is formulated as

$$y_k = h(x) = \begin{pmatrix} 1 & 0 \end{pmatrix} x_k. \tag{4}$$

The height H over ground is known a-priori and assumed to be constant. Hence, the scale is obtained by equation (2).

The measurement noise is estimated by applying the plane fitting algorithm on a long sequence of the KITTI dataset [5] and observing the standard deviation of all distances until convergence.

The uncertainties and the initial values of the state are determined by assuming a typical vehicle movement in cities. As initial value we choose a vehicle velocity of $v = 10\frac{m}{s}$. Given a camera frame rate of $f = 10Hz$ and that the length of the unscaled translation is set to 1.0, the initial value of the scale is

$$s_{start} = \frac{v}{f} = 1.0m. \tag{5}$$

Furthermore, the scale has a standard deviation of

$$\sigma_s = \frac{\sigma_a}{f^2}. \tag{6}$$

The standard deviation of the acceleration of the vehicle is approximated by $\sigma_a = 3\frac{m}{s^2}$.

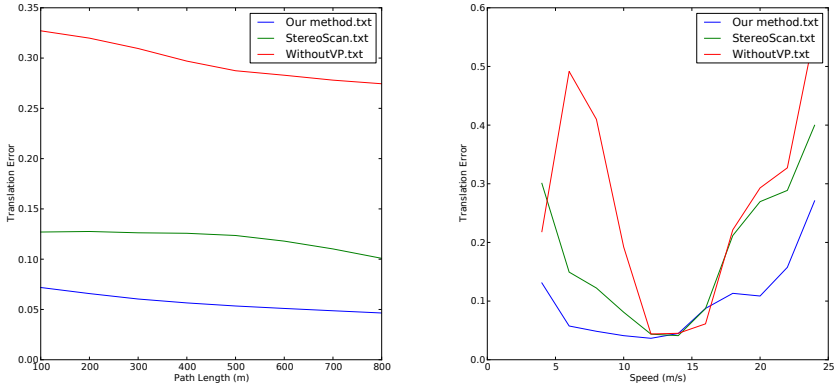
3.4 Gating

Humans permanently apply plausibility checks to understand the current scenery and act accordingly. In that way false conclusions are rejected directly. In a similar manner we apply several gating mechanisms to augment the robustness and the precision of our algorithm. First we mask a region of interest, i.e. the street, as shown in Fig. 1. We only reconstruct points if their location in the image lies within the region of interest, which improves the performance of the plane refinement. Moreover, we reject false values during the tracking of the plane parameters. In that way, plane normals that differ more than 30° from the filtered normal will not be updated. In that case only the prediction step is put into action to obtain values for the normal and the distance. Hereby, an advantage of the encapsulation of the normal estimation and the distance estimation becomes clear; a rejection of the normal of the ground plane due to errors or occlusions does not obstruct the distance estimation and therefore the computation of the scale.

4 RESULTS

To demonstrate the performance of our method, we evaluated it on the KITTI dataset [5] on sequences 00 to 10, which provide sets of images in rural and urban areas and ground-truth with a maximum error of $0.1m$, established by a high-precision GPS. The dataset includes urban, suburban and rural scenes with vehicle velocities from $3\frac{m}{s}$ to $25\frac{m}{s}$.

There are no methods published on the KITTI benchmark which relax the constraint of a fixed angle between the camera and the ground plane. Therefore, we compare the algorithm presented in this paper with the same procedure but without the use of vanishing directions. The results are shown in Fig. 4 and Fig. 5. The usage of vanishing points reduces the mean translation error drastically to errors between 6% and 4%, for velocities from $6\frac{m}{s}$ to $14\frac{m}{s}$. Compared to the baseline algorithm *StereoScan* we can reduce the translation error on average by 30%. In environments with rich structure for vanishing points, we are capable to reduce the translation error by more than 50%, to values between 3.5% and 2% in our velocity range of interest, which is



(a) (b)

Figure 4: 4a: Average translation error over length. Our method converges to a value of approximately 5%. 4b: Average translation error over speed. The mean translation error of our method is at 5% for velocities between $6\frac{m}{s}$ and $14\frac{m}{s}$. At smaller velocity, the reconstruction is inaccurate since the baseline between image features is small and hence, the plane reconstruction is more uncertain.

shown in Fig. 5. Consequently, we conclude that the usage of vanishing points is of high importance to the correct estimation of the scale in urban environments.

For velocities larger than $14\frac{m}{s}$, the scale diverges. This is the result of tuning for the usage in urban scenes, for which the presented scaling method is conceived. The source of the divergence lies in the fact that elevated motion blur is present in proximity to the camera caused by increasing vehicle velocity. Consequently an insufficient number of points with acceptable accuracy is reconstructed in the region of interest. Thus we trade off low accuracy at higher velocity against high accuracy at medium speed, by defining the region of interest in the gating step.

Although the advantages of our approach are already visible in the KITTI dataset, the improvement to previous approaches becomes even more apparent in situations, in which the camera platform shows significant roll- and pitch-movements. Thence our method clearly outperforms existing algorithms in urban scenarios by being more general and precise.

5 CONCLUSIONS AND OUTLOOK

The accurate estimation of a vehicle trajectory is a key feature for various applications in robotics and in particular for Advanced Driver Assistance Systems. Monocular visual odometry is a cheap and handy methodology for this subject, but it fails in practice because the scale cannot be computed correctly. While successful approaches for scale estimation have been proposed, existing methods either lack applicability on vehicles

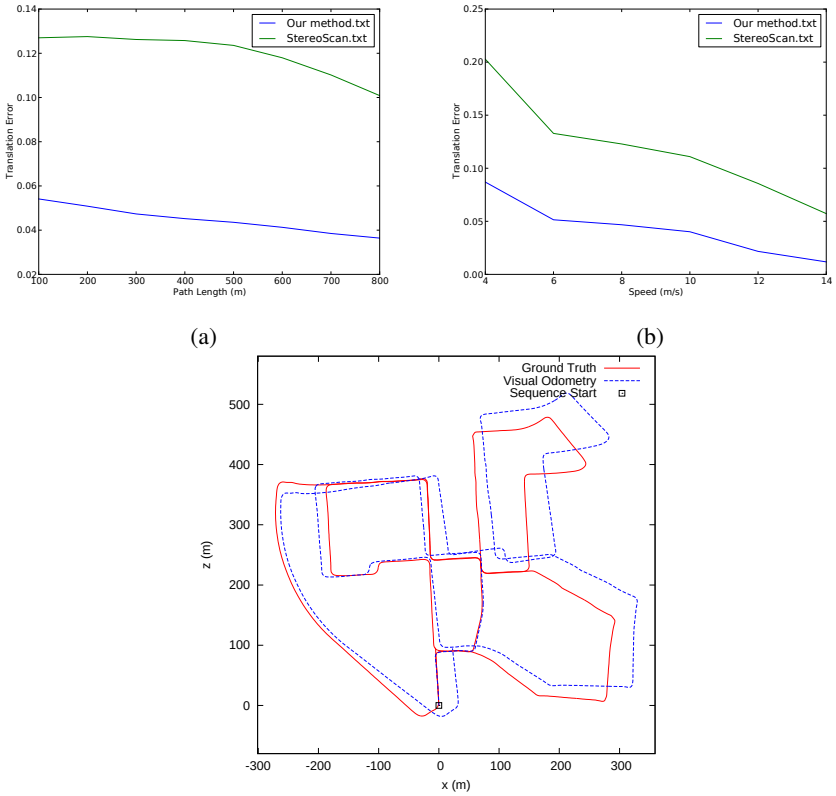


Figure 5: Translation error over path length, translation error over speed and trajectory for sequence 00 of the KITTI dataset, where urban scenery is dominant. 5a: Translation error over path length. Our method converges to a translation error of 4%. 5b: Translation error over speed. The vanishing point estimation renders our method accurate with errors between 1% and 5% at velocities between $6\frac{m}{s}$ and $14\frac{m}{s}$. 5c: Trajectory produced by our method compared with the ground-truth.

or assume severe constraints, especially regarding pitch- and roll-movements. In this paper we proposed a novel pitch- and roll-tolerant scaling algorithm for monocular visual odometry, tailored for urban environments. We used two complementary methods for ground plane normal estimation, which are a fit to reconstructed points and vanishing points. Finally a trimmed Least Squares optimization refined the plane and revealed the scale. The evaluation on the KITTI dataset showed that the method outperforms state of the art scaling algorithms in urban environments with a scaling error smaller than 4%.

In a future extension we will handle the fact that the vanishing direction extraction does not have a benefit in rural areas. Thereto, we will include a quality criterion to quantify whether a measured vanishing point is beneficial or not by employing its stability.

References

[1] Jakob Engel, Thomas Schöps, and Daniel Cremers, *Lsd-slam: Large-scale direct monocular slam*, Computer Vision–ECCV 2014, Springer, 2014, pp. 834–849.

[2] Christian Forster, Matia Pizzoli, and Davide Scaramuzza, *Svo: Fast semi-direct monocular visual odometry*, Proc. IEEE Intl. Conf. on Robotics and Automation, 2014.

[3] Friedrich Fraundorfer and Davide Scaramuzza, *Visual odometry: Part ii: Matching, robustness, optimization, and applications*, Robotics & Automation Magazine, IEEE **19** (2012), no. 2, 78–90.

[4] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger, *A new performance measure and evaluation benchmark for road detection algorithms*, International Conference on Intelligent Transportation Systems (ITSC), 2013.

[5] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, *Vision meets robotics: The kitti dataset*, International Journal of Robotics Research (IJRR) (2013).

[6] Andreas Geiger, Philip Lenz, and Raquel Urtasun, *Are we ready for autonomous driving? the kitti vision benchmark suite*, Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[7] Andreas Geiger, Julius Ziegler, and Christoph Stiller, *Stereoscan: Dense 3d reconstruction in real-time*, IEEE Intelligent Vehicles Symposium (Baden-Baden, Germany), June 2011.

[8] R. Hartley and A. Zisserman, *Multiple view geometry*, 7th ed., Cambridge University Press, 2010.

[9] Richard I Hartley, *In defense of the eight-point algorithm*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **19** (1997), no. 6, 580–593.

[10] Georg Klein and David Murray, *Parallel tracking and mapping for small ar workspaces*, Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on, IEEE, 2007, pp. 225–234.

[11] Davide Scaramuzza and Friedrich Fraundorfer, *Visual odometry [tutorial]*, Robotics & Automation Magazine, IEEE **18** (2011), no. 4, 80–92.

[12] T. Schöps, J. Engel, and D. Cremers, *Semi-dense visual odometry for AR on a smartphone*, September 2014.

[13] Shiyu Song and Manmohan Chandraker, *Robust scale estimation in real-time monocular sfm for autonomous driving*.

[14] Shiyu Song, Manmohan Chandraker, and Clark C Guest, *Parallel, real-time monocular visual odometry*, Robotics and Automation (ICRA), 2013 IEEE International Conference on, IEEE, 2013, pp. 4698–4705.

[15] Greg Welch and Gary Bishop, *An introduction to the kalman filter*, 1995.