Evaluation of Fisheye-Camera Based Visual Multi-Session Localization in a Real-World Scenario

Peter Muehlfellner[1]*, Paul Furgale[2], Wojciech Derendarz[1], Roland Philippsen[3]

Abstract

The European V-Charge project seeks to develop fully automated valet parking and charging of electric vehicles using only low-cost sensors. One of the challenges is to implement robust visual localization using only cameras and stock vehicle sensors. We integrated four monocular, wide-angle, fisheye cameras on a consumer car and implemented a mapping and localization pipeline. Visual features and odometry are combined to build and localize against a keyframe-based three dimensional map. We report results for the first stage of the project, based on two months worth of data acquired under varying conditions, with the objective of localizing against a map created offline.

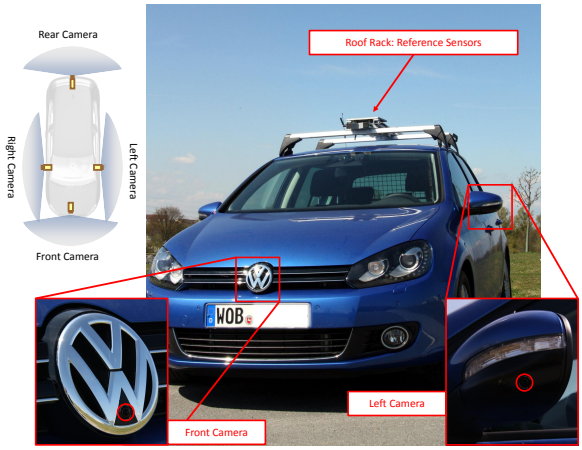

Figure 1: The V-Charge Golf, showing its integrated sensors and the very subtle differences to a regular "consumer car".

1. INTRODUCTION

Pose estimation for mobile robots via visual sensors has come a long way in the last decade. In the context of Visual SLAM (Simultaneous Localization And Mapping) advances have been made that solve the basic SLAM problem robustly, efficiently and over huge distances ([1], [2]). Nonetheless, even though proven possible ([3], [4]), the application of visual approaches for the localization of fully automated cars has not undergone in-depth analysis. We therefore take a closer look at the performance of a state-of-the-art visual localization system, integrated into a research vehicle that performs fully automated driving, in a particularly challenging real-world scenario: parking lots and parking garages.

The requirement for this comes from the EU-funded project Autonomous Valet Parking and Charging ("V-Charge"), which is concerned with driverless cars in parking lots or garages. V-Charge distinguishes itself from previous efforts into automated driving (e.g. DARPA) by using a sensor setup much more closely oriented on what could be found in a consumer car. Cameras are the mainstay of the V-Charge sensor system, due to their low-cost nature combined with the ability to still provide rich information about the environment. Fig. 1 depicts the sensor-setup used for visual localization, which consists of four monocular, wide-angle, fisheye cameras that together give a 360° view of the vehicle-surroundings.

V-Charge takes a map-based approach to automated navigation. This means that a parking lot or garage needs to be mapped by a survey vehicle prior to driverless operation. Such pre-built maps contain information relevant to tasks of localization and navigation: visual landmarks, and the road network. In this paper we want to explore the challenges resulting from the fact that the localization maps need to "live" longer than the comparatively limited time span of a single automated run.

For this paper, we do not consider the full multi-session SLAM problem, as we do not allow updates to

* The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013, Challenge 2, Cognitive Systems, Interaction, Robotics, under grant agreement No 269916, V-Charge.

[1] Volkswagen AG, Group Research, Germany [2] Autonomous Systems Lab, ETH Zürich, Switzerland [3] Intelligent Systems Lab, Halmstad University, Sweden

the map with each new session. Rather, we work with a map created at one point in time, that we use for localization afterwards.

2. RELATED WORK

The SLAM problem is commonly formulated probabilistically as estimating the joint posterior distribution system state and map, conditioned on previous states, control-inputs and observations [5]. Assuming additive Gaussian noise, the maximum likelihood solution for estimating the state can be found using non-linear batch optimization. While the full maximum likelihood solution to the SLAM problem has complexity $O((M+N)^3)$ in the number of poses N and observations M and is thus impractical, various approaches have been developed to either marginalize the full problem in some way (see [6]), or to deal with subsets (e.g. [7]). The sparsity of the problem can be exploited to allow tractable solutions that involve non-marginalized poses, with superior results [8].

Long-term (a.k.a. lifelong or multi-session) SLAM emphasizes deployment over long time scales. Examples of systems that re-use maps are [9][2][10][11][12] for "classical" Vision-based SLAM; [13][14][15] for topological localization; as well as [16] for LASER-based SLAM. In road environments, localization based on homographical maps (or "overhead views") [17][18] has shown success.

A recent survey of Visual SLAM for driverless cars can be found in [19]. The most successfuly applications of such systems are by [4] and [3].

In [4], a map of an urban environment is constructed based on GPS-data and stereo-vision. The environment is represented as a set of sparse 3D points resulting from local image features (e.g. SURF). In a separate localization stage, these 3D points are matched to currently observed images and used as an input for BA.

The work of [3] builds on the Relative Bundle Adjusmtent (RBA) framework described in [1]. The output of stereo-based online Visual SLAM is saved as an "experience" if it is visually distinct from previous traversals of an environment. Experiences are re-used over many sessions and added to as necessary. An evaluation over several months under varying conditions indicates that the number of distinct experiences is bounded in the considered environment.

3. SYSTEM OVERVIEW

The V-Charge sensor setup differs significantly from the systems commonly employed for Visual SLAM. Instead of a stereo-pair or a monocular camera, the setup shown in Fig. 1 consists of four fisheye cameras resulting in increased complexity that is handled by a pipeline (Fig. 2) that combines several state-of-the art techniques.

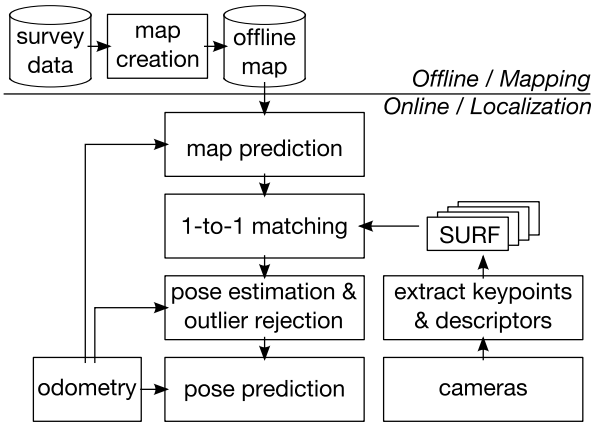

Figure 2: The block structure of our localization pipeline, showing both the offline (map-building)- and online (localization)- phases of our system.

We perform map-building and localization in two separate phases. Both rely solely on (1) the images provided by the four wide-angle monocular fisheye cameras and (2) wheel-odometry. Mapping is carried out offline, on survey data-sets that are recorded whilst manually driving the vehicle. In the online phase, we localize against this pre-built map.

3.1. Map Representation and Map-Building

Requirements for our visual localization system are influenced by other components for automated driving as well as characteristics of the environment and the sensor setup. The following two points serve as a motivation for the chosen map structure (presented further below):

- Path planning and following in a complex environment requires metric information. Due to the inherent 3D-structure of our environment (e.g. ramps, multi-level garages), we represent poses with the six degrees of freedom (6-DOF) for describing spatial rigid body transforms.
- Global positioning systems are unavailable, thus errors will accumulate over large distances. However, while locally accurate pose estimates are needed, global map precision is not important for the V-Charge scenario. Therefore, the map is represented topologically, as a set of relatively defined coordinate frames (also termed map nodes).

We organize our map as a set of nodes V_i annotated with: the set of the four camera images obtained at time i, the vehicle coordinate frame $\underrightarrow{\mathscr{F}}_{V_i}$, and a "best guess" pose estimate relative to some central map reference frame $\underrightarrow{\mathscr{F}}_{M}$. Within each frame $\underrightarrow{\mathscr{F}}_{V_i}$, we express a sparse set of 3D point *landmarks* $\mathbf{l}_{ik}$ arising from local image keypoints (e.g. SURF [20]).

For map creation, camera inputs from the survey dataset are used if they satisfy a minimal baseline to the previous image-set, via wheel-odometry. Then, based on keypoints and descriptors extracted from the individual images, as well as manually defined loop closures, we conduct a series of optimization steps that provide estimates for the sparse 3D-structure of the environments and the poses of the map nodes. Optimization starts out with basic, open-loop visual odometry, followed by a global bundle adjustment of the open loop, manual loop-closure selection, pose graph relaxation, and a final global bundle adjustment on the closed loop.

3.2. Localization Against a Known Map

Localization amounts to the problem of estimating the 6-DOF transformation $\mathbf{T}_{V_i,V_j}$ between a reference frame $\underrightarrow{\mathscr{F}}_{V_i}$ in the map and the current vehicle frame $\underrightarrow{\mathscr{F}}_{V_j}$. Using the keypoints from the current camera images and the landmark keypoints from the map frame, this transformation is the solution of a two-frame bundle adjustment problem.

We perform robust data-association between the current observation and the frame predicted to be closest in the map. A 6-DOF nearest neighbour search is carried out based on a relative pose predicted from odometry and the last localization. The 3D-landmarks associated with this node are then projected onto vehicle camera system at the predicted position, resulting in a set of sparse image-points. Each of these points matched to the currently observed local image features (using a SURF GPU-pipeline [21]) based on distances in image- and descriptor space.

The set of correspondences formed this way, together with the initial vehicle pose estimate provided by odometry, then forms the input for a non-linear optimization problem. This problem is based on the reprojection error for each 2D-3D correspondence, which is a function of the (time-invariant) camera parameters, the feature-landmark correspondence and the transformation $\mathbf{T}_{V_i,V_j}$ between the mapped- and the current vehicle frame. From these error-terms, an objective function, to be minimized, is built and a solution for $\mathbf{T}_{V_i,V_j}$ is obtained using Levenberg-Marquardt iterations in an efficient framework.

4. EVALUATION METHODOLOGY

In order to determine the viability of our visual localization system for path planning and vehicle control, we are interested in the precision, the robustness and the availability of visual localization in the given scenario. In the following we define metrics and experiments to quantify these factors.

We are foremost interested in the Metric localization error. It can be defined as a 6-DOF error transformation $\mathbf{T}^{error}_{V_i,V_j}$ that quantifies the difference between a transformation $\mathbf{T}_{V_i,V_j}$ estimated by our system, and the corresponding ground truth (GT) transformation $\hat{\mathbf{T}}_{V_i,V_j}$. For a GT sensor (such as high precision INS/DGPS) that provides information about the vehicle pose in some global reference frame $\underrightarrow{\mathscr{F}}_{G}$, the relative ground truth is calculated as $\mathbf{T}^{-1}_{G,V_i}\mathbf{T}_{G,V_j}$. Here $\mathbf{T}_{G,V_i}$ is the output of the GT sensor synchronized in time with the map-frame $\underrightarrow{\mathscr{F}}_{V_i}$, and $\mathbf{T}_{G,V_i}$ the GT-output synchronized with the online-localization frame $\underrightarrow{\mathscr{F}}_{V_j}$.

This error-formulation only takes into account the *relative* localization error. Differences between the estimated pose of map nodes and the associated ground-truth poses are discarded. This reflects the fact that both localization and planning take place on a manifold that is defined by the topological structure of our map. Paths on this manifold can be (re-)traversed using only locally accurate pose estimates, as shown in [22].

To quantify the error values present in $\mathbf{T}^{e}_{V_i,V_j}$, we split this transformation into its translation and rotation components $\mathbf{t}_e$ and $\mathbf{R}_e$. We take the absolute length $\|\mathbf{t}_e\|$ of the translation ($\mathbf{t}_e = (x,y,z)^T$). For calculating the rotational error, we transfer $\mathbf{R}_e$ into an axis-angle representation and use the magnitude of the rotation angle, $|\phi_e|$ as the error metric. This provides us with an uniform sampling of the error space.

The series of experiments that are performed in order to obtain these numbers share a common setup. A map of the path that we want to navigate later on (see section 3) is built based on a survey-dataset taken at some fixed point in time. Evaluation-datasets are collected at later time in a similar process, with time-differences ranging from minutes to months (in the long run, we expect to collect data over several months and even years). When collecting evaluation data we neither stray far from the recorded paths, nor take special care to drive exactly the same paths. We perform this first evaluation in an outdoor area, where we have a INS/DGPS-system for providing ground truth with an error standard deviation of around 2 cm.

We want to confirm and quantify the following statements with our experiments:

Table 1: Statistics for each dataset. $\mathbf{t}_e$, $t_{x,y,z}$ denotes the translational error and its components, ϕ_e the angular error.

	Same Day	Week One	
$RMS\|\mathbf{t}_e\|$	0.028	0.0937	(m)
$RMS\|\phi_e\|$	0.134	0.3427	(deg)
$Max\|t_x\|$	0.0756	0.1926	(m)
$Max\|t_y\|$	0.0515	0.2079	(m)
$Max\|t_z\|$	0.0289	0.0663	(m)
$Max\|\phi_e\|$	0.8001	1.3347	(deg)

- Precision: The relative metric precision of the localization system allows automated driving. For determining this, we look at the average values of the metric error values ($\|\mathbf{t}_e\|$ and $|\phi_e|$.).
- Robustness: The relative localization error is bounded, as long as we successfully localize against the map. This means that the distribution of the metric error values should also be bounded.
- Availability: Maps can be re-used over at least moderate time-spans (weeks), as long as environmental conditions remain similar. This means that the number of inlier-matches after optimization is above some threshold (which we arbitrarily set to 20) and that the localization error bounded and low.

5. RESULTS

For evaluation in this paper, we use datasets collected on the parking lot of our research campus, which is also where we tested fully automated navigation with the localization results provided by our system.

Three datasets, including INS/DGPS ground-truth data, are available to us. The dataset named "Same Day" was captured minutes after the mapping-data was acquired, the "Week One" dataset was taken within a week from map-creation, and the "Two Months" dataset was created two months prior to the map we use. Snapshots from all of these datasets are given in Figure 3. It can be seen that the first three images appear very similar, while in the last image, significant differences not only in the occupancy of the parking lot, but also in usually more static elements — such as the foliage — are present. Evaluation is performed by creating a map for the first dataset, and attempting localization using this map with all available data.

Table 1 summarizes the results of the experiment. Here, *RMS* denotes the Root Mean Squared value for

Figure 3: Example snapshots from each dataset.

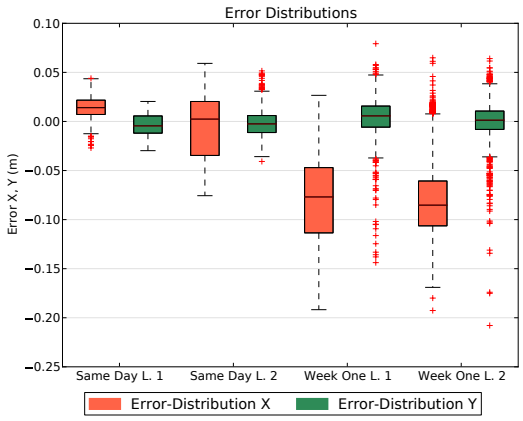

Figure 4: Box-plots for the X-(longitudinal) and Y-(lateral) error values for each loop of the Same Day and Week One datasets.

the relative translational and angular errors, and *Max* the maximum error, for the respective dataset. The map created from the first dataset did not allow successful localization with the Two Months dataset — the number of inlier-matches after optimization was persistently low (< 20). This dataset is therefore not included in the table. On the other hand, localization over the period of one week was successful, achieving precision in the order of centimeters.

An additional overview of the localization results is provided in Fig. 4 for which the datasets were split up into individual loops of the parking lot. The former shows box-plots[1] of the X- and Y- components of the

[1]Each box is structured as follows: the center line shows the median value, the boundaries of the box give the two quartiles and the two "whiskers" have the length of 1.5 times the Inter Quantile Range. Data outside of the whiskers is plotted as individual samples.

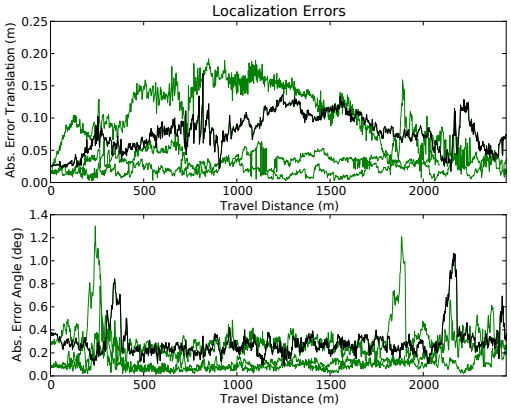

Figure 5: The absolute translational- and angular errors for each separate loop in the evaluated datasets. The error for the last loop in the week-one dataset is highlighted in black.

translational error. Note that here we give the error in a reference frame attached to the the vehicle: the X-axis points in the same direction as the nose of the car. These errors can also be referred to as the longitudinal and lateral deviations. Fig. 5 details the translational- and angular errors over sample-number overlaid each other for all loops.

Fig. 6 shows the poses estimated by our system in a single metric coordinate frame. Both the results of the localization runs using the various datasets from different days, as well as the path-estimate for the map after manual loop-closure and offline-optimization, are shown. Comparing this enlarged image to a similarly zoomed version of plotted reference trajectories shows that the trajectories, locally, are qualitatively similar to those provided by the ground truth sensor.

Timing measurements for the main localization step yield results in the order of seconds for the processing of a single set of four images on a modern PC. As for fully automated navigation a much more frequent update of the vehicle pose is required, we bridge the missing localization measurements by appending wheel-odometry measurements at 100Hz. It turns out, that even with a localisation rate below 1Hz, owing to low speeds and the local precision of wheel-odometry, smooth automated navigation within parking lots is possible.

5.1. Discussion

The results for the average metric errors, summarized in Table 1 are well within the limits required for localizing within a single lane of a parking lot. This

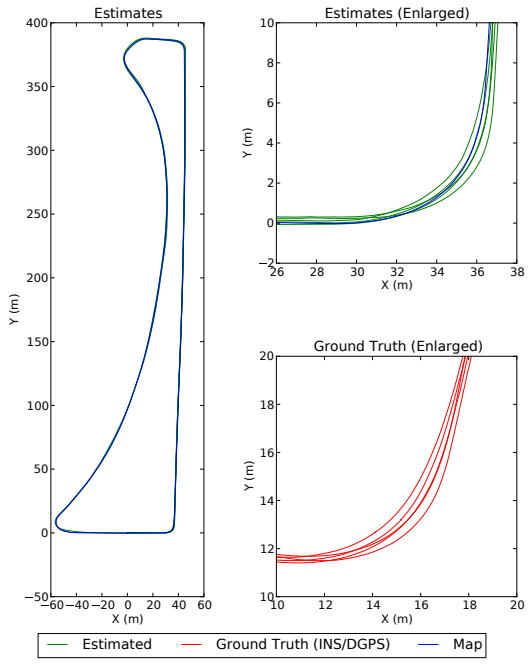

Figure 6: Overhead view of the mapping and localization results. On the right, estimated and Ground Truh trajectories are shown for an enlarged tile of the map, based on the lower right corner of the trajectory.

performance already allowed us to perform a number of successful fully automated trips in the mapped area. Furthermore, as can be seen from the distribution of the error values the overall system is also quite robust as long as successful localization against the map is possible.

Over longer time-frames and under drastically changed environmental conditions, the localization might fail to match to the map. This is illustrated by the complete failure to localize against the Two Months dataset.

One way to deal with this, would be to complement the map by including more and more data over time, as changes occur. While the convergence of such a process is shown experimentally in [3], this nonetheless means that the amount data stored for each map would increase. In this case the need for exploration of compression techniques — in order to save bandwidth — becomes evident.

Another open question not addressed here, is the portability of maps between different vehicles and camera systems. Similar to changeable environments, creating and storing a complete map for each single car seems infeasible for a widespread deployment of fully automated vehicles. The performance when re-using

maps between vehicles with roughly similar camera systems, but different calibration parameters, needs to be determined in the future. A map-representation that allows transferal of data between vehicles with different camera systems (stereo, N-cameras) seems even more desirable.

6. CONCLUSION

We have presented a state-of-the art visual localization system for the use in a driverless car, employing a close-to-market sensor setup consisting of four wide-angle mono cameras. Based on pre-recorded map, our solution combines the images provided by these four cameras with odometry in order to achieve real-time capabable localization. The pose estimates for this system were evaluated on datasets separated from the survey data by timespans varying from several minutes to 2 months. The results for a single week show success, and we were also able to employ the thusly generated poses for fully automated navigation of the mapped parking lot. For the visually very different two months old dataset, localization failed, since not enough valid matches could be found. This shows the importance of dealing with long-term changes in the structure or appearance of the environment.

References

[1] G. Sibley, C. Mei, I. Reid, and P. Newman, "Vast-scale outdoor navigation using adaptive relative bundle adjustment," *The International Journal of Robotics Research*, vol. 29, no. 8, pp. 958–980, 2010.

[2] P. Furgale and T. Barfoot, "Visual teach and repeat for long-range rover autonomy," *Journal of Field Robotics*, vol. 27, no. 5, pp. 534–560, 2010.

[3] W. Churchill and P. Newman, "Practice makes perfect? managing and leveraging visual experiences for lifelong navigation," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4525–4532.

[4] H. Lategahn and C. Stiller, "City gps using stereo vision," in *Vehicular Electronics and Safety (ICVES), 2012 IEEE International Conference on*, july 2012, pp. 1 –6.

[5] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *Robotics & Automation Magazine, IEEE*, vol. 13, no. 2, pp. 99–110, 2006.

[6] S. Thrun, W. Burgard, D. Fox *et al.*, *Probabilistic robotics*. MIT press Cambridge, MA, 2005, vol. 1.

[7] A. Davison, I. Reid, N. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 1052–1067, 2007.

[8] H. Strasdat, J. Montiel, and A. Davison, "Real-time monocular slam: Why filter?" in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, may 2010, pp. 2657 –2664.

[9] K. Konolige and J. Bowman, "Towards lifelong visual maps," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2009, pp. 1156–1163.

[10] E. Royer, M. Lhuillier, M. Dhome, and J. Lavest, "Monocular vision for mobile robot localization and autonomous navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 237–260, 2007.

[11] H. Johannsson, M. Kaess, M. Fallon, and J. Leonard, "Temporally scalable visual slam using a reduced pose graph," 2012.

[12] J. McDonald, M. Kaess, C. Cadena, J. Neira, and J. Leonard, "6-dof multi-session visual slam using anchor nodes," in *Proc. of European Conference on Mobile Robots, ECMR*, 2011, pp. 69–76.

[13] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 219–236, 2007.

[14] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *Intelligent Vehicles Symposium (IV), 2011 IEEE*. IEEE, 2011, pp. 794–799.

[15] C. Valgren and A. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features," in *Proceedings of the European conference on mobile robots (ECMR)*, 2007, pp. 253–258.

[16] P. Biber and T. Duckett, "Experimental analysis of sample-based maps for long-term slam," *The International Journal of Robotics Research*, vol. 28, no. 1, pp. 20–33, 2009.

[17] A. Napier and P. Newman, "Generation and exploitation of local orthographic imagery for road vehicle localisation," in *Intelligent Vehicles Symposium (IV), 2012 IEEE*. IEEE, 2012, pp. 590–596.

[18] O. Pink, "Visual map matching and localization using a global feature map," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE, 2008, pp. 1–7.

[19] G. Ros, A. Sappa, D. Ponsa, and A. Lopez, "Visual slam for driverless cars: A brief survey," in *Intelligent Vehicles Symposium (IV) Workshops, 2012 IEEE*. IEEE, 2012.

[20] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Computer Vision–ECCV 2006*, pp. 404–417, 2006.

[21] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[22] A. Howard, "Multi-robot mapping using manifold representations," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 4198–4203.