Probabilistic Map Localization through Visual Odometry

Marcus A. Brubaker
TTI Chicago
mbrubake@cs.toronto.edu

Andreas Geiger
KIT & MPI Tübingen
geiger@kit.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

1. Introduction

Self-localization is key for building autonomous systems that are able to help humans in everyday tasks. In this paper we are interested in building *affordable* and *robust* solutions to self-localization for the autonomous driving scenario. Currently, the leading technology in this setting is GPS. While being a fantastic aid for human driving, it has important limitations in the context of autonomous systems. Notably, the GPS signal is not always available, and its localization can become imprecise (*e.g.*, in the presence of skyscrapers, tunnels or jammed signals). While this might still be viable for human driving, consequences can be catastrophic for self-driving cars.

To provide alternatives to GPS localization, place recognition approaches have been developed which assume that image or depth features which identify the relevant locations are stored in a database, and cast the localization problem as a retrieval task. In combination with GPS, impressive results have been demonstrated (*e.g.*, the Google self-driving car) but it remains unclear if maintaining an up-to-date world representation will be feasible given the computation, memory and communication requirements. Furthermore, these solutions require that all locations to be localized have been visited before.

In contrast to the above mentioned approaches, we tackle the problem of self-localization in places that we have not been seen before. We take inspiration from humans, which perform this task while having access to only a rough cartographic description of the environment, *i.e.*, a map. We propose to exploit the ready availability of community-developed maps from the OpenStreetMap (OSM) project, for the task of vision-based localization. The OSM maps are detailed and freely available, making this an inexpensive solution. Towards this goal, we derive a probabilistic map localization approach that uses visual odometry estimates and OSM data as the only inputs. We demonstrate the effectiveness of our approach on a variety of challenging scenarios making use of the recently released KITTI visual odometry benchmark [4]. Our experiments show that we are able to localize after only a few seconds of driving with an accuracy of 3 meters on a 18km^2 map containing $2,150$km of drivable roads.

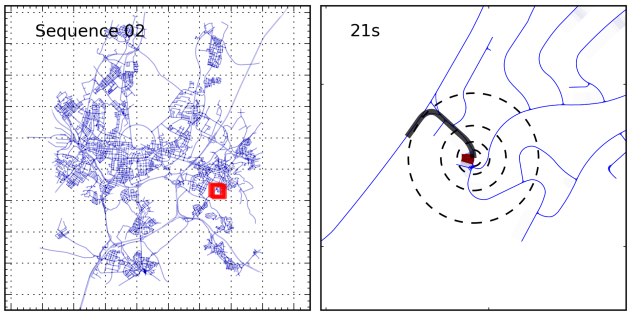

Figure 1. **Visual Self-Localization:** We demonstrate localizing a vehicle with an average accuracy of 3.1m within a map of $\sim 2,150$km of road using only visual odometry measurements and freely available maps. In this case, localization took less than 21 seconds. Grid lines are every 2km.

Early approaches for *map localization* [2, 3, 6, 8] made use of Monte Carlo methods and the Markov assumption to maintain a sample-based posterior representation of the agent's pose. However, they are generally restricted to small-scale environments and low-noise laser-scan observations which provided strong location cues. In contrast, our method operates on large map areas (up to 18km square) using only noisy measurements of egomotion, computed using weak location cues of visual odometry [5, 7].

2. Visual Localization

We propose to use roof-mounted cameras to self-localize a driving vehicle. The only other information we have is a map of the environment in which the vehicle is driving which contains streets as connected line segments. We exploit visual odometry in order to obtain the trajectory of the vehicle. As this trajectory is often too noisy or ambiguous for direct shape matching, we propose a probabilistic approach to self-localization that employs visual odometry measurements in order to determine the instantaneous position and orientation of the vehicle in a given map.

The map data is represented by a directed graph where nodes represent street segments and edges define the connectivity of the roads. We define the position and orientation of a vehicle in the map in terms of the street segment u that the vehicle is on, the distance from the origin of that

		00	01	02	03	04	05	06	07	08	09	10	Average
Position Error	M	15.6m	*	8.1m	18.8m	*	5.6m	*	15.5m	45.2m	5.4m	*	18.4m
	S	2.1m	3.8m	4.1m	4.8m	*	2.6m	*	1.8m	2.4m	4.2m	3.9m	3.1m
Heading Error	M	2.0°	*	1.5°	2.4°	*	2.0°	*	1.3°	10.3°	1.6°	*	3.6°
	S	1.2°	2.7°	1.3°	1.6°	*	1.4°	*	1.9°	1.2°	1.3°	1.3°	1.3°

Table 1. **Sequence Errors:** Average position and heading errors for 11 training sequences. "M" and "S" indicate monocular and stereo odometry. All averages are computed over localized frames and "*" indicates sequences which did not localize.

street segment d and the offset of the local street heading θ.

We define the state of the model at time t to be $\mathbf{x}_t = (u_t, \mathbf{s}_t)$ where u_t is the identity of the current street and $\mathbf{s}_t$ is the position and orientation of the vehicle on that street at the current and previous frames. The motion dynamics use a simple second order linear model with additive Gaussian noise. Visual odometry observations at time t, $\mathbf{y}_t$, measure the linear and angular displacement from time $t-1$ to time t. The observations are modelled using a linear model with additive Gaussian noise. Street transitions occur probabilistically, where the probability of transitioning from the current street to a connected one is a a sigmoid-like function of the relative distance to the start of the next street.

Given this model we wish to compute the filtering distribution, $p(\mathbf{x}_t|\mathbf{y}_{1:t}) = p(\mathbf{s}_t|u_t, \mathbf{y}_{1:t})p(u_t|\mathbf{y}_{1:t})$, where $p(u_t|\mathbf{y}_{1:t})$ is a discrete distribution over streets and $p(\mathbf{s}_t|u_t, \mathbf{y}_{1:t})$ is a continuous distribution over the position and orientation on a given street which we represent using a Mixture of Gaussians. Inference exploits the Gauss-Linear nature of dynamics when possible (*i.e.*, in the absence of street transitions) and exploit efficient approximations otherwise. This is coupled with a simplification procedure, which reduces the number of mixture components, without loosing significant details. Our inference algorithm is easily parallelized and can run at frame rate on average for moderate sized maps. Full details are available in [1].

3. Results

To evaluate our approach in realistic situations, we performed experiments on the recently released KITTI benchmark for visual odometry [4]. We utilize the 11 training sequences for quantitative evaluation (where ground truth GPS data is available). The visual odometry input to our system is computed using LIBVISO2 [5], a freely available library for monocular and stereo visual odometry. For illustration purposes, here we extracted mid-size regions of OpenStreetMap data which included the true trajectory and the surrounding region. On average, they cover an area of 2km^2 and contain 47km of drivable roads. It is important to note that our method also localizes successfully on much larger maps, see Fig. 1 for example, which covers 18km^2 and contains 2,150km of drivable roads. Quantitative results can be found in Table 1. Here, "M" and "S" indicate results using monocular and stereo visual odometry respectively. In addition, we computed odometry measurements from the GPS trajectories (entry "G" in the table) and ran our algorithm using the parameters for the stereo data.

The accuracy of position and heading estimates is not well defined until the posterior has converged to a single mode. Thus, we compute accuracy once a sequence has been *localized*. We define a sequence to be localized when for at least five seconds there is a single mode in the posterior and its distance to the ground truth position is less than 20 meters. Once the criteria for localization is met, all subsequent frames are considered localized. Errors in global position and heading of the MAP state for localized frames were computed using the GPS data as ground truth.

Overall, we are able to estimate the position and heading to 3.1m and $1.3°$ using stereo visual odometry. Notice that simply projecting the GPS data onto the nearest road segment in the map produces an error of 1.44m! These results also outperform typical consumer grade navigation systems which offer accuracies of around 10m. Using monocular odometry as input performs worse, but is still accurate to 18.4m and $3.6°$, once it is localized. However, due to its stronger drift, it fails to localize in some cases as in sequence 01 which contains highway driving only, where high speeds and sparse visual features results in an accumulated odometry error of more than 500m.

References

[1] M. A. Brubaker, A. Geiger, and R. Urtasun. Lost! Leveraging the Crowd for Probabilistic Visual Self-Localization. In *CVPR*, 2013.

[2] F. Dellaert, W. Burgard, D. Fox, and S. Thrun. Using the condensation algorithm for robust, vision-based mobile robot localization. *CVPR*, 1999.

[3] D. Fox, W. Burgard, F. Dellaert, and S. Thrun. Monte carlo localization: Efficient position estimation for mobile robots. In *AAAI*, 1999.

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, 2012.

[5] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IV*, 2011.

[6] J.-S. Gutmann, W. Burgard, D. Fox, and K. Konolige. An experimental comparison of localization methods. In *ICIRS*, 1998.

[7] D. Nister, O. Naroditsky, and J. R. Bergen. Visual odometry. In *CVPR*, 2004.

[8] S. M. Oh, S. Tariq, B. N. Walker, and F. Dellaert. Map-based priors for localization. In *ICIRS*, 2004.