Monocular Self Localization in an Urban Environment Using a Prior-Based Soft Optimization Robust Estimation Method

Shay Dekel[1], Dan Levi[1], Michael Slutsky[1], and Ilan Shimshoni[2]

[1]General Motors Advanced Technical Center Israel
[2]Department of Information Systems, University of Haifa

***Abstract*— Autonomous vehicle driving in urban environments is a challenging task that requires localization accuracy exceeding that available from GPS-based inertial guidance systems. For map-based driving, a 3D laser scanner can be utilized to localize the vehicle within a previously recorded 3D map. Such scanners are however not feasible for mass production due to cost considerations. In this paper we present a localization algorithm that creates an off-line predefined map and then localizes with respect to this map. First, the map is constructed by a service vehicle equipped with a calibrated stereo camera rig and a high precision navigation system. Then, the global localization ego-pose can be obtained in any vehicle equipped with a standard GPS and a single forward looking camera for extracting and matching features to relevant map candidates. We use a recently proposed estimation method called SOREPP (Soft Optimization method for Robust Estimation based on Pose Priors) that utilizes relevant priors for achieving high performance, fast and reliable estimation, even with a small fraction of inliers. During the estimation it uses all the matched correspondences without need for random sampling to find the inliers. This method eventually obtains an outlier-free set of landmarks, used to estimate the ego-pose with high accuracy. We evaluate our algorithm on real world data comprised of a challenging 4.5km drive. Our algorithm achieves accurate localization results: a mean lateral absolute error of 14.35cm and a mean longitudinal absolute error of 18.63cm.**

I. INTRODUCTION

Accurate vehicle localization is a challenging problem for autonomous vehicles. Future driver assistance systems require a centimeter level of accuracy in the ego pose estimation. Common approaches address this problem using a global navigation satellite systems (GNSSs). However this approach suffers from multi-path and shadowing effects especially in urban environments. Some expensive approaches that incorporate a high precision GPS with an inertial measurement unit (IMU) may only reach these accuracies in open sky environments. Moreover, this solution is prohibitively expensive for mass production. Thus a cost-effective localization solutions in GPS-denied situation is required.

Recently, methods that use a previously acquired map for localization are emerging [4], [28], [19]. First a detailed map of the environment is built from sensors and stored for future use. In this stage a vehicle (usually designated as a service vehicle) equipped with calibrated stereo cameras and a high precision GPS travels a certain route, extracts 3D features from consecutively images and stores them in a database along with their exact reference ego poses. At runtime, as the vehicle drives over the mapped routes, it estimates its position by querying the map for candidates close to the current ego pose and finds corresponding features between the current image and the reference image. Usually this process is followed by robust estimation methods from the RANSAC family [9]. Finally the global ego pose is obtained based on the current relative estimation and the already known database high precision pose. Other map based approaches [22], [24] incorporate a laser scanner that provides high precision range measurements and very convincing results. However these approaches are also very expensive and raise the question of cost efficiently when talking about mass production.

In this paper we present a framework that inherits the characteristics of map-based approaches that creates off-line a predefined accurate map which contains a set of vehicle locations of a specific reference driving route. Each location point from the set has extracted landmarks features that were viewed from its position associated with their 3D locations in the world. For this task we have driven a service vehicle equipped with a high precision GNSS+INS receiver to derive the ego pose, A calibrated stereo camera rig is used for the 3D location estimation of a feature set using the triangulation method [13]. In the on-line localization stage, we drive the same route using only the input of a monocular camera setup and a common vehicle GPS to localize with respect to this map. The main contribution of this paper is incorporating a recently purposed soft optimization robust estimation method, called SOREPP [11] that utilizes relevant pose priors in the area of the desired solution for achieving high performance reliable matching estimation. It exploits all the correspondence landmarks without depending on their inlier fraction. This approach is capable eventually to yield a very efficient solution whose runtime is independent on the inlier fractions of the matched landmarks.

The rest of the paper is organized as follows; Related work is reviewed in Section II. The mapping part is described in Section III. The on-line localization is elucidated in Section IV. Our experimental setup and evaluation results are both described in Section V. Finally a conclusion and future work are summarized in Section VI.

II. RELATED WORK

The problem of recovering the camera ego-motion is highly correlated with a recent works on Simultaneous

Localization And Mapping (SLAM) [16], [12], [31], [17], [7], [30], [27], Visual Odometry (VO) [10], [15], [1] and map based approaches for localization [28], [3], [26], [6], [24], [22], [20].

SLAM is the process by which a mobile robot can build a map of the environment and at the same time use this map to estimate its own location. The past decade has seen rapid and exciting progress in solving the SLAM problem. Since SLAM, in its native form, scales quadratically with the number of landmarks in the map, a great majority of works have focus on improving computational efficiency while ensuring consistent accurate positions and map estimations. Early solutions started with extended Kalman filter implementations [27] but the focus has more recently shifted towards the bundle adjustment domain [30]. Although a great success of reducing the computational complexity the challenges are still exist especially for real-time implementations.

On the other hand, visual odometry is the process of estimating the egomotion of the robot using the input of a single or multiple cameras attached to it. It aims at recovering only the robot path incrementally, pose after pose, and potentially optimizing only over the last n poses of the path such as [14] that introduced a windowed bundle adjustment approach based on a sliding window optimization. Although VO has reached maturity that allowed us to successfully use it for certain classes of applications, it has an intrinsic problem of drift in long range navigation that are caused by an error accumulation since visual odometry is based on relative measurements. Conversely, our method does not suffer from the drift problem.

Other approaches [28], [22], [24], [29], [18], [4], [3], [20] to tackle the localization problem are mapped based approaches which are usually very accurate and achieve high efficiency reliable solutions. Since these approaches inherently rely on predefined maps, these maps have to be very accurate, usually off-line recorded and followed with a high precision navigation system. For example in the work of [34] an off-line optimization such as Bundle Adjustment (BA) [32] was introduced to refine their map's precision.

The work of Levinson et al. [22] uses a rotating laser scanner to extract 3D points and match them to a 3D point cloud map while also considering remittance of laser beam as an extra measure. Moosmann et al. [24] also uses a scan matching approach that achieved improved localization accuracy over high precision GPS. Our approach is different and uses images of a single camera during localization, the landmark map is sparser and the sensor is considerably less expensive.

The works of Badino et al. [4], [3] describe the entire image by whole image SURF features that significantly reduce the mapping database storage. They introduced a topological localization method where visual and laser range finder features are fused in a histogram filter to yield the nearest pose of a previously recorded mapping trajectory during online operation. Although they demonstrated high efficacy capabilities across wide environmental changes, including lighting differences, seasonal variations, and occlusions they achieved an average accuracy of ~ 1 meter which is insufficient for some autonomous vehicle assignments due to the fact that intelligent vehicles undoubtedly depend on accurate ego localization solutions.

Our work is motivated in part by the work of Lategahn et al. [18] that also exploits a previously computed landmark map and a monocular camera for localization except that they also added a fused inertial measurement unit (IMU) for final refinement. They present significant results and robustness especially for different day times, but they didn't provide a customary experimental results for comparison. They also published another work [20] that exploits only a vision system for localization without relying on GNSS navigation system by utilization holistic features or static objects on the route as map reference points, then localized relative to this map objects with high accuracy. This approach however, requires a supervised labeling procedure of the static objects during the map creation. Our method is different since our monocular visual system provides a final accurate solution without any fusion dependencies on complex refinement steps of an inertial measurement unit, and especially the need for supervised activities such as explicitly labeling semantic objects during the map creation is eliminated. In particular, our work can be adopted for urban navigation assignment such that of traveling a certain predefined bounded area even in GPS denied situations. For example, assuming we consider an autonomous vehicle that has to report on-line its self position in a certain urban bounded environment. To this end an accurate map of the entire environment is created off-line by using a service vehicle with a high precision navigation system and calibrated stereo cameras. Then our method is capable to localize the vehicle with respect to this map.

III. MAP BUILDING

We construct an off-line map by navigating a predefined route only once using our service vehicle equipped with high precision Real Time Kinematics (RTK) navigation system and a calibrated stereo camera rig that can be shown in Figure 10. This navigation system comprises a GNSS+INS receiver that is located in the vehicle and an additional ground-based Differential GPS (DGPS) reference station. The whole system has the capability of delivering up to several centimetres level accuracy in off-line post processing conditions. In the first step, the GPS raw data and the synchronized stereo sequence of the trajectory is recorded. Then a post processing tightly coupling optimization is performed off-line that eventually yields an accurate ego pose estimations fused with each image pair of the stereo sequence. After this stage, salient features are extracted from each image pair. Then, right image features are matched to the left image of each pair based on the epipolar constraint which is already known from the stereo calibration. Finally, we store a regularly sample set of the sequential accurate ego poses of the vehicle as a reference database mapping of the trajectory. In Addition, for each such sample we also store the best left image features descriptors and their associated 3D landmarks which are estimated by the stereo pairs using the

triangulation method [13]. In order to refine our 3D location estimation of the landmarks and also the vehicle ego poses, we employed bundle adjustments [32] optimization off-line to enhance these accuracies and eventually derive a high precision map.

A number of feature descriptor types could be used for localization [8]. Among these, SURF features have shown to be robust in outdoor environments [33], [2]. For high robust performance we chose the Upright-SURF (U-SURF) descriptor [5], [3] which is invariant to scale and to rotations of the vertical axis. U-SURF provides improved speed and is robust to rotations of up to $\pm 15°$.

IV. ONLINE LOCALIZATION

The output of the online localization stage is to sequentially provides the estimated position of the vehicle in real time based on the reference off-line learned database map.

We divide the localization procedure into six steps, each step will be individually explained below: A) We first query the database for relevant m candidates based on a coarse initial position from the current GPS measurement and the previous estimation. Note that the GPS reading is used only for candidates selection. B) We retrieve landmarks in the immediate vicinity of the current test image and match them to the landmarks of the m candidates using a fast approximation with multiple randomized k-d trees [25]. C) We select the best appropriate reference candidate to be used for localization estimation procedure based on its matching score ratio. D) We use the input of the matched landmarks between the current test and the best-candidate pair to estimate the epipolar geometry using a recently introduced robust estimation algorithm called SOREPP [11] that eventually retrieves the unscaled relative translation and rotation matrix. E) We estimate the scale using re-projection error minimization approach that basically yields the estimated distance between the two involved images based on the already known database of 3D points as viewed from the best-candidate. F) We finally calculate the global position based on the current relative estimation from SOREPP, the estimated scale and the known best candidate reference position.

In the sequel, the current ego pose is defined by a $4X3$ matrix, denoted by q_t, that consists of the global test rotation matrix R_t and the global test translation vector t_t. Analogously, the global ego pose of the best candidate from the database is denoted by q_c, that also consists of the global candidate rotation matrix R_c and the global candidate translation vector t_c. Throughout the rest of the article we assume poses to be parameterized by $4X3$ homogeneous matrices with $3X3$ rotation matrix R and $3X1$ translation vector t.

A. Select Relevant Candidates

For each test frame during the driving, we select the $m = 4$ closest candidates from the database within a certian radius r based on the combination of the current coarse GPS reading and the previous position estimation. Since our testing GPS has a measurement error in the order of few meters, We choose a radius of $r = 15[m]$ for this selection procedure. Figure 1 shows the data-base candidates poses $q_1, ..., q_4$ as viewed from the vicinity of the test pose q_t.

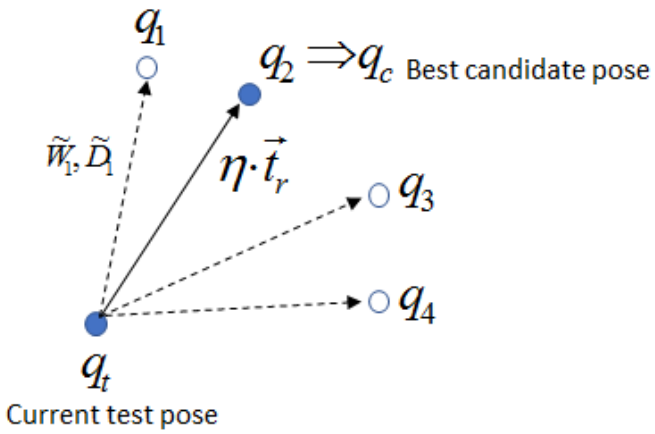

Fig. 1. Selecting the best candidate: q_t is the current test pose. $q_1 ... q_4$ are the relevant candidates poses which were queried from the data-base in the vicinity region of the test pose q_t. In this example q_2 was selected as the best candidate, η indicates the scale estimation and t_r is the estimated relative translation vector.

B. Match Features

We use the U-SURF descriptor as mentioned in Section III to match landmarks between the current test image to each one of the m candidates using a fast approximate nearest neighbors that is based on the randomized kd-tree algorithm [25]. This is a common approach of approximate nearest neighbor search, in which suboptimal neighbors are sometimes returned. The advantage of this approach is that it can be orders of magnitude faster than exact search, while still providing near-optimal accuracy. This approximation is sufficient for selecting the best candidate.

C. Selecting The Best Candidate

If there is not enough camera motion between two frames, the computation of the epipolar geometry is an ill-conditioned problem. Specifically, we would like to select a candidate q_i so that there is as largest camera motion as possible between its position and the current test image while still being able to match the features. For this reason we would like to use a candidate selection mechanism by introducing a two-term energy score function for each matching image pair. The first term, denoted by W_i, is the mean matching ratio of all the associated matching ratios in each i test-candidate pair:

$$W_i = \frac{1}{N} \sum_{k=1}^{N} w_i(k), \tag{1}$$

where N is the number of correspondences and $w_i(k)$ is the ratio between the top match distance and the second best match distance of the $k - th$ correspondence of the i test-candidate pair as described in [23]. Specifically, we

would like to select a candidate that particulary has matching features with low ratios that indicates its quality with respect to the other candidates. The second term, denoted by D_i, is the median of all the disparities distribution of the features in each i test-candidate pair. We would like to select a candidate that contains features with high disparities while at the same time maintains a substantial amount of features. The best candidate is then given by minimization of:

$$BestCandidate = \arg\min_{i \in 1..m} \left\{ \frac{\tilde{W}_i}{\tilde{D}_i} \right\}, \tag{2}$$

where $\tilde{W}_i$ and $\tilde{D}_i$ are the normalized versions of the above mean ratio W_i and median disparities D_i of the i test-candidate pair respectively. The distribution of the features in the image also has an impact of the selection procedure. A good candidate to be considered has a uniform distribution of the features in the image plane and also has features close to the camera since these feature play an important role for longitudinal localization estimation. In Figure 1 for example, candidate q_2 which has the lowest energy score is selected as the best candidate.

D. Soft Optimization Robust Estimation Based on Pose Priors

After carefully choosing the best candidate in the previous section, in this part we aim to robustly estimate the localization of the vehicle relative to the chosen already known best candidate position. We follow the work of [11], called SOREPP (Soft Optimization Robust Estimation using Pose Priors). This work is a recently purposed estimation algorithm designed to exploit pose priors. It assumes that, given a set of inliers, there exists a procedure which can estimate the parameters of a model that optimally explains or fits this data. It sparsely samples the pose space around the measured pose and for few a promising solutions applies a robust optimization procedure. It uses all the putative correspondences simultaneously even though many of them are outliers, yielding a very efficient solution whose runtime is independent of the inlier fractions.

As mentioned above in Section IV, R_c and R_t are the candidate and test rotation matrices respectively indicating the vehicle orientation in global coordinate system. In the same way t_c and t_t represent the candidate and test translation vectors respectively indicating the positioning of the vehicle in same global coordinate system. We define the relative rotation matrix by R_r and the relative translation vector by t_r of the test-candidate pair. Specifically, they can be defined as follows:

$$R_r = R_t R_c^T, t_r = R_c(t_t - t_c). \tag{3}$$

We describe the relative translation vector t_r using polar coordinates, defining α to be the relative horizontal angle and β to be the relative vertical angle:

$$\alpha = \arctan(t_r(x), t_r(y)), \tag{4}$$

$$\beta = \arcsin\left(\frac{t_r(z)}{\|t_r\|}\right). \tag{5}$$

We define an auxiliary vector of unknowns s to be composed of five angles:

$$s \equiv (\psi_r, \theta_r, \phi_r, \alpha, \beta)^T, \tag{6}$$

where ψ_r, θ_r and ϕ_r, are the relative heading, pitch and roll Euler angles respectively that uniquely compose R_r. We define a M-estimator Gaussian score based on the Sampson distance [13], which approximates the geometric distance for every putative correspondence k:

$$g(k, s) = \exp\left(\frac{-d(k, s)^2}{2\sigma_h^2}\right), \tag{7}$$

where $d(k, s)$ denotes the Sampson distance for a certain pose vector s associated with each k correspondence. σ_h specifically applies the soft thresholding for this score. For a true relative pose, inliers receive high scores close to 1, while outliers far away from the model would receive low scores close to 0.

SOREPP eventually seeks to minimize the following objection function:

$$\hat{s} = \arg\min_s \left\{ c\left(\sum_{k \in \Omega_{all}} \tilde{p(k)}(1 - g(k, s))\right) + (\lambda(s))^2 \right\}, \tag{8}$$

where c is a certain weighting parameter. $\tilde{p(k)}$ is a rough approximation of the probability of being an inlier. Practically, it can be defined as the normalized version of the inversely proportional to the square of correspondence ratios ,$w(k)$, from the matching procedure of the test best-candidate pair. That can be calculated as :

$$\tilde{p(k)} = 1 - \frac{w(k)^2}{\sum_{l \in \Omega_{all}} w(l)^2} \tag{9}$$

The second term $\lambda(s)$ is a regularization term that derives the Mahalanobis distance between the current pose and the given prior camera pose s_0:

$$\lambda(s) = \frac{1}{|s|}\sqrt{(s - s_0)^T {\Sigma_{s0}}^{-1}(s - s_0)}, \tag{10}$$

where Σ_{s0} is a predefined prior covariance matrix of the camera that contains the expected errors of each parameter.

This objective function (8) combines the minimization of the Sampson distances for many putative correspondences as possible while keeping the solution close to the pose prior s_0. The minimization can be done using any standard optimization method. In our implementation we use Levenberg-Marquardt [21]. The SOREPP algorithm is simple and fast, as well as robust to low inlier fractions and significant pose noise.

The inputs for the SOREPP algorithm are: 1) the approximate features correspondence of the test-candidate pair from Section IV-B and 2) the pose prior s_0 calculated as the

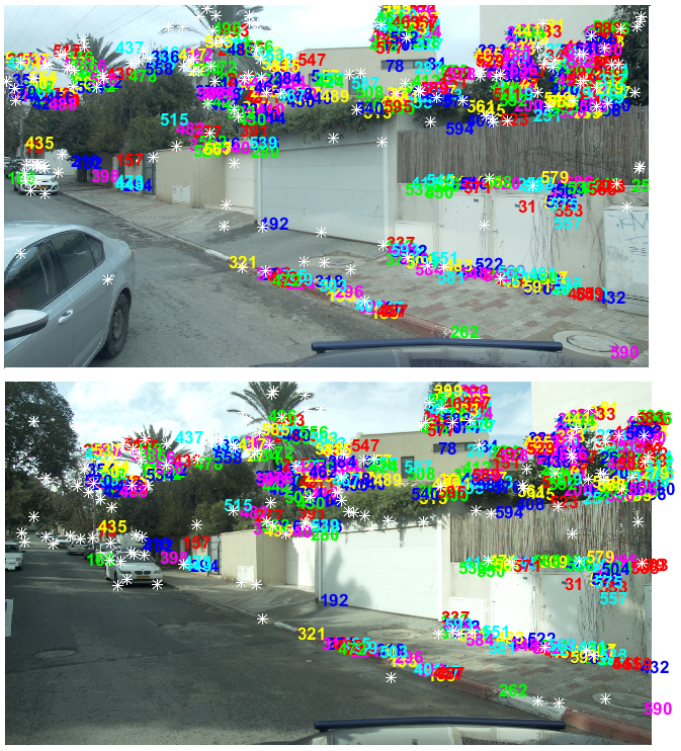

Fig. 2. Online Sequence: The current navigation image (top). The best candidate database image (bottom). Numerative correspondence landmarks indicate a successful matches after SOREPP using our approach, whereas white correspondence landmarks are the initial approximated correspondences before SOREPP that flagged as outliers

average between the current GPS measurement pose and the previous estimated pose. The output will eventually be the relative pose estimation between the current ego pose of the vehicle and the best candidate pose i.e. the relative translation vector t_r and the relative rotation matrix R_r. Figure 2 shows a typical test-candidate pair image. In this example the numerated correspondences indicate successful matches after the SOREPP solution, while the white correspondences indicate the initial approximate features set before SOREPP that eventually determined them as outliers.

E. Estimating the Scale

In order to obtain the global ego pose of the vehicle, we have to estimate the geometric scale distance between the current test position and the known best candidate position since the estimated translation vector from the previous Section IV-D, t_r, is up to scale. To this end we employ a re-projection error minimization function that minimizes the sum of squared distances between the projection of the already known 3D landmark locations in the world of the best candidate and their associated features on the test image plane. Specifically, given the feature correspondences between the current test image and the best candidate image and their known associated 3D positions from the mapping stage, we reproject the 3D positions back to the test image plane, attempting to minimize the actual error distances between the estimated pixel positions and their observation

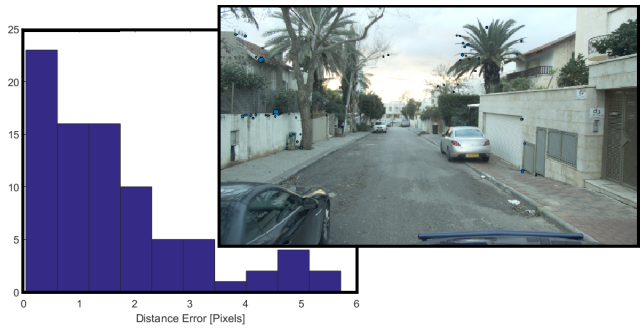

Fig. 3. Histogram of the reprojection errors. The test image landmarks that have significant contribution to the reprojection equation are marked in the image by circles. Circle size corresponds to the pixels error level of the back projection.

positions on the image plane. More concisely, an error function of the sums of squared back projection errors is then defined by:

$$E(\eta) = \sum_{i=1}^{N} \zeta_i \|\pi(l_i, K, s, \eta) - z_i)\|^2, \tag{11}$$

where z_i denotes the pixel locations in the test image, π is a standard camera projection function as given in [13] that takes the already known 3D landmarks points $l_i = (X_i, Y_i, Z_i)$ from the database, the camera intrinsic parameters K, the already estimated relative camera pose vector s of (6) and the unknown scale η and returns its pixel position on the image plane. ζ_i denotes normalized weights that are particulary taken as proportional to the landmarks inverse depthes, $1/Z_i$. Finally, the absolute scale η is optimally found by minimizing the sum of squared reprojection errors (11) using the Levenberg-Marquardt algorithm [21], which iteratively finds a minimum by linearizing the function to be minimized in the neighborhood of the current estimate. Figure 3 shows results of the reprojection error estimation and its calculated histogram for a specific frame.

F. Global Positioning

The global ego pose of the vehicle is finally obtained by concatenation of the rotation matrix and translation vector:

$$q_t = [R_t, t_t], \tag{12}$$

where t_t is the global translation vector and R_t is global rotation matrix of the vehicle defined as:

$$\begin{aligned} t_t &= t_c + \eta R_c^T t_r, \\ R_t &= R_r R_c. \end{aligned} \tag{13}$$

V. Experimental Results

A. Setup

We conducted our experiments using our service vehicle equipped with a mounted stereo cameras rig and a Novatel SPAN-CPT navigation system that can be shown in Figure 10. The cameras were configured to acquire 1920X1080 pixels images at 15Hz. For the purpose of analysis and

evaluation, we selected a complex, 4.5km route that contains a variety of environments, ranging from complex urban to residential and parklike spaces (See figure 4). The route includes man-made and natural structures: buildings, traffic signs, trees, open spaces, and multiple slopes, as well as moving objects such as vehicles, pedestrians, cyclists, and motorcyclists. The trajectory makes a loop, meaning that the vehicle must face all orientations during the trajectory. This is important in order to test the robustness of our method to illumination artifacts, such as specularities made by direct sunlight exposure. Figure 5 shows such an example during our localization stage.

The route map and ground truth information were obtained by our Novatel SPAN-CPT navigation system that is capable of delivering up to several centimetre level accuracy of the vehicle's position. This level of accuracy is eventually obtained using DGPS and an off-line post-processing optimization software that enables tight coupling of the GNSS and IMU measurements and delivers the most satellite observations and the most accurate, continuous stable solution possible.

B. Results

Next we present experiments on real world data to assess and evaluate our method. To this end the mapping data-base sequence and the test sequence were recorded in the same route but at a different time of day.

The ground truth and estimated global localization results are shown in Figure 4. The storage of the data-base mapping of 4.5km route requires roughly 250MB. The algorithm is running at 10Hz which is suitable for standard laptop real-time requirements.

For evaluating our performance quantitatively and computationally, we ran our localization algorithm one time with SOREPP and second time with the 8-Points RANSAC [13] on the same route. The statistical means and variances of the associated estimated parameters of the localization results are summarized in Table I both with SOREPP and the 8-Points RANSAC. As can be seen our algorithm with SOREPP achieved accurate localization results of mean lateral absolute error of 14.35cm and mean longitude absolute error of 18.63cm over a challenging 4.5km route that outperforms the standard robust estimation with RANSAC.

TABLE I
STATISTICAL RESULTS

	SOREPP	**8-Points RANSAC**
Mean Frame Computation time	**0.1213** [Sec]	0.3511 [Sec]
Mean Lateral Absolute Error	**14.35** [Cm]	30.65 [Cm]
Mean Longitude Absolute Error	**18.63** [Cm]	49.13 [Cm]
Mean Heading Absolute Error	**0.357** [Deg]	0.931 [Deg]
STD Lateral Absolute Error	**19.07** [Cm]	59.23 [Cm]
STD Longitude Absolute Error	**30.69** [Cm]	71.48 [Cm]
STD Heading Absolute Error	**0.6919** [Deg]	1.32 [Deg]

Figure 6 shows the estimated longitudinal error histogram (left) and the estimated latitudinal histogram (right). The

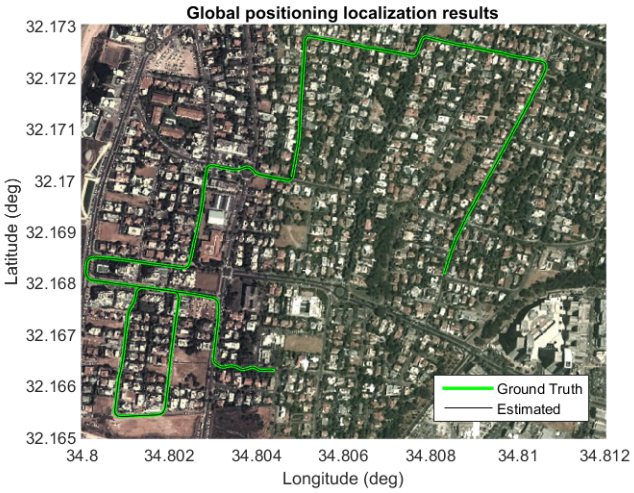

Fig. 4. The global localization results on a satellite map. The ground truth trajectory (green) and the estimated trajectory (thin black line that appeared inside the green line)

estimated longitudinal error corresponds to the part of the error along the driving direction, whereas the latitudinal error corresponds to the complementary part of the error orthogonal to the driving direction.

The orientation estimation of the vehicle achieves mean heading absolute error of 0.35deg. Figure 7 shows the sequential heading estimation in each frame of the online localization compared to their ground truth values. In addition the heading error is summarized as histogram in Figure 8.

Figure 9 shows a typical result of the on-line localization system that contains the estimated parameters and illustration of the estimated vehicle position and orientation on a Google map.

VI. CONCLUSION AND FUTURE WORK

We have presented a system for vehicle ego pose estimation in urban environment using a single camera. The vehicle ego pose is localized relative to a previously high precision data-base map. This map-based localization method doesn't suffer from accumulation of errors like visual odometry methods and can be easily adopted to long range navigation assignments.

We also use a recently purposed soft optimization robust estimation method, called SOREPP that utilizes relevant pose priors for achieving high performance reliable estimation during the matching procedure utilizing all the correspondence landmarks without depending on their inlier fraction (like RANSAC-based algorithms). The idea to use such an approach leads to a very efficient solution whose runtime is independent on the inlier fractions of the landmarks matching.

Our on-line localization stage has no dependence on complex infrastructure, needs no workspace modification or special stereo calibration and even not highly dependent on GPS accuracy. It only requires a monocular camera setup, a standard vehicle GPS and runs on an every day laptop. We

Fig. 5. An example of our method coping with challenging illumination artifacts made by direct sunlight exposure. Top: The current navigation image. Bottom: the best candidate from the database. The color numbers denote the feature correspondences.

evaluate our algorithm on real world data and achieve high performance real-time accurate promising results.

In particular, our work can be adopted for urban navigation assignments such as that of traveling a certain predefined route even in GPS denied situations. For example, assuming we consider an autonomous vehicle that has to report online its self position in a certain urban bounded environment. To this end an accurate map of the entire environment is created off-line using a service vehicle with high precision navigation system and calibrated stereo cameras, then incorporating our method to localize the vehicle with respect to this map.

We believe that future research will be focus on the possibilities of dropping the dependencies on the GPS positions using a probabilistic scheme and also develop new methods to cope with heavy traffic situations that might affect the performance of the system. Another issue that is opened for research is how to efficiently reduce the data-base storage capacity.

ACKNOWLEDGMENT

We would like to thank Viki Kogan for her help with the initialization steps of this project.

REFERENCES

[1] M. Agrawal and K. Konolige. Real-time localization in outdoor environments using stereo vision and inexpensive gps. In *The 18th International Conference on Pattern Recognition*, volume 3, pages 1063–1068, 2006.

[2] A. Ascani, E. Frontoni, A. Mancini, and P. Zingaretti. Feature group matching for appearance-based localization. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3933–3938, Sept 2008.

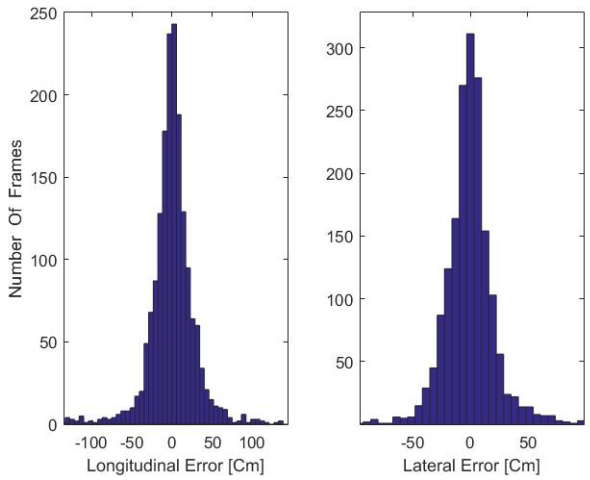

Fig. 6. Left: The estimated longitudinal error histogram. Right: The estimated latitudinal error histogram.

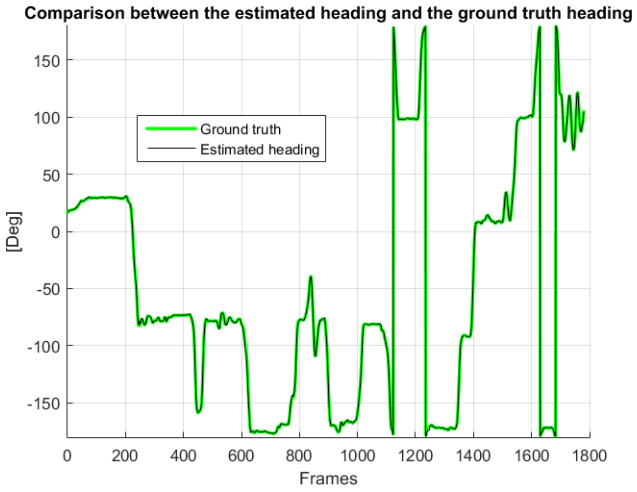

Fig. 7. The estimated heading results compared to the ground truth values.

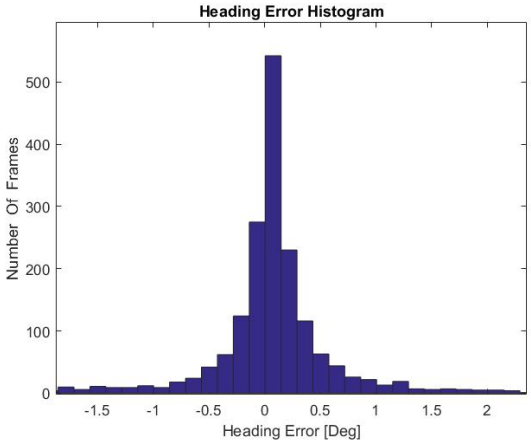

Fig. 8. The estimated heading error histogram.

Fig. 9. Online sequence: A typical result of the localization stage. Left: the current test navigation image. Right: the best candidate from the database. Below: a map illustration of the vehicle estimated position and orientation and the associated estimated parameters on the right.

Fig. 10. Our service research vehicle that constructs the reference database map and the ground truth information. It equipped with a mounted stereo camera rig and Novatel SPAN-CPT high precision navigation system.

[3] H. Badino, D. Huber, and T. Kanade. Visual topometric localization. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 794–799, June 2011.

[4] Hernan Badino, Daniel Huber, and Takeo Kanade. Real-time topometric localization. In *International Conference on Robotics and Automation*, May 2012.

[5] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, June 2008.

[6] Michael Bosse and Robert Zlot. Continuous 3d scan-matching with a spinning 2d laser. In *IEEE International Conference on Robotics and Automation ICRA, 2009*, pages 4312–4319, May 2009.

[7] Zhenhe Chen, Jagath Samarabandu, and Ranga Rodrigo. Recent advances in simultaneous localization and map-building using computer vision. *Advanced Robotics*, 21(3):233–265, 2007.

[8] Aldo Cumani and Antonio Guiducci. Comparison of feature detectors for rover navigation. In *Proceedings of the 11th WSEAS International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering*, MMACTEE'09, pages 126–131, Stevens Point, Wisconsin, USA, 2009. World Scientific and Engineering Academy and Society (WSEAS).

[9] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[10] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *Intelligent Vehicles Symposium (IV), 2011 IEEE*, pages 963–968, June 2011.

[11] Y. Goldman, E. Rivlin, and I. Shimshoni. Robust epipolar geometry estimation using noisy pose priors. *Submitted to Pattern Recognition*, 2015.

[12] G. Grisetti, R. Kummerle, C. Stachniss, and W. Burgard. A tutorial on graph-based slam. *Intelligent Transportation Systems Magazine, IEEE*, 2(4):31–43, winter 2010.

[13] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[14] K. Konolige and M. Agrawal. Frameslam: From bundle adjustment to real-time visual mapping. *IEEE Transactions on Robotics*, 24(5):1066–1077, Oct 2008.

[15] Kurt Konolige, Motilal Agrawal, and Joan Solà. Large-scale visual odometry for rough terrain. In *The 13th International Symposium Robotics Research, ISRR 2007, November 26-29, 2007 in Hiroshima, Japan*, pages 201–212, 2007.

[16] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3607–3613, May 2011.

[17] H. Lategahn, A. Geiger, and B. Kitt. Visual slam for autonomous ground vehicles. In *IEEE International Conference on Robotics and Automation (ICRA), 2011*, pages 1732–1737, May 2011.

[18] H. Lategahn, M. Schreiber, J. Ziegler, and C. Stiller. Urban localization with camera and inertial measurement unit. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 719–724, June 2013.

[19] H. Lategahn and C. Stiller. City gps using stereo vision. In *IEEE International Conference on Vehicular Electronics and Safety (ICVES), 2012*, pages 1–6, July 2012.

[20] H. Lategahn and C. Stiller. Vision-only localization. *IEEE Transactions on Intelligent Transportation Systems*, 15(3):1246–1257, June 2014.

[21] Kenneth Levenberg. A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Mathematics*, (2):164–168, 1944.

[22] Jesse Levinson and Sebastian Thrun. Robust vehicle localization in urban environments using probabilistic maps. In *IEEE International Conference on Robotics and Automation, ICRA 2010, Anchorage, Alaska, USA, 3-7 May 2010*, pages 4372–4378, 2010.

[23] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004.

[24] Frank Moosmann and Christoph Stiller. Velodyne SLAM. In *Proceedings of the IEEE Intelligent Vehicles Symposium*, pages 393–398, Baden-Baden, Germany, June 2011.

[25] Marius Muja and David G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Applications, VISAPP*, pages 331–340, 2009.

[26] A. Napier, G. Sibley, and P. Newman. Real-time bounded-error pose estimation for road vehicles using vision. In *13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010*, pages 1141–1146, Sept 2010.

[27] P. Pinies and J.D. Tardos. Large-scale slam building conditionally independent local maps: Application to monocular vision. *IEEE Transactions on Robotics*, 24(5):1094–1106, Oct 2008.

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

[29] Eric Royer, Maxime Lhuillier, Michel Dhome, and Jean-Marc Lavest. Monocular vision for mobile robot localization and autonomous navigation. *Journal of Computer Vision*, 74(3):237–260, 2007.

[30] Gabe Sibley, Christopher Mei, Ian Reid, and Paul Newman. Vast scale outdoor navigation using adaptive relative bundle adjustment. *International Journal of Robotics Research*, 29(8):958 – 980, July 2010.

[31] H. Strasdat, A.J. Davison, J.M.M. Montiel, and K. Konolige. Double window optimisation for constant time visual slam. In *IEEE International Conference on Computer Vision (ICCV), 2011*, pages 2352–2359, Nov 2011.

[32] Bill Triggs, Philip McLauchlan, Richard Hartley, and Andrew Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice, Lncs*, pages 298–375. Springer Verlag, 2000.

[33] Christoffer Valgren and Achim J. Lilienthal. Sift, surf & seasons: Appearance-based long-term localization in outdoor environments. *Robot. Auton. Syst.*, 58(2):149–156, February 2010.

[34] Zhiwei Zhu, Taragay Oskiper, Supun Samarasekera, Rakesh Kumar, and H.S. Sawhney. Real-time global localization with a pre-built visual landmark database. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pages 1–8, June 2008.