VIZARD: Reliable Visual Localization for Autonomous Vehicles in Urban Outdoor Environments

Mathias Bürki[1,2], Lukas Schaupp[1], Marcin Dymczyk[2], Renaud Dubé[2], Cesar Cadena[1], Roland Siegwart[1], and Juan Nieto[1]
[1]Autonomous Systems Lab, ETH Zürich, {firstname.lastname}@mavt.ethz.ch
[2]Sevensense Robotics AG, {firstname.lastname}@sevensense.ch

***Abstract*— Changes in appearance is one of the main sources of failure in visual localization systems in outdoor environments. To address this challenge, we present VIZARD, a visual localization system for urban outdoor environments. By combining a local localization algorithm with the use of multi-session maps, a high localization recall can be achieved across vastly different appearance conditions. The fusion of the visual localization constraints with wheel-odometry in a state estimation framework further guarantees smooth and accurate pose estimates. In an extensive experimental evaluation on several hundreds of driving kilometers in challenging urban outdoor environments, we analyze the recall and accuracy of our localization system, investigate its key parameters and boundary conditions, and compare different types of feature descriptors. Our results show that VIZARD is able to achieve nearly 100% recall with a localization accuracy below $0.5m$ under varying outdoor appearance conditions, including at night-time.**

I. Introduction

Localization is a pivotal capability of any autonomous vehicle. By knowing their precise location, vehicles are able to plan a path to a next goal location, navigate safely in the environment, and eventually successfully complete their mission. Especially for autonomous vehicles in urban environments, localization is challenging, as GNSS based localization systems fail to provide reliable and precise enough localization near buildings due to multi-path effects, or in tunnels or parking garages due to a lack of visible satellites. Alternative exteroceptive sensor modalities are therefore necessary to accomplish this task, of which LiDARs and cameras have received most attention in recent years [1], [2]. While LiDARs have become more suited for mass market adoption, we believe there are still significant advantages with camera-based localization systems, despite the challenges related to long-term appearance change in outdoor environments. Cameras remain considerably more cost-effective than LiDAR sensors, allowing them to be deployed in multitudes and in a flexible way on a large quantity of vehicles. Furthermore, they can be used for sensing both appearance and geometric information of the environment, and are often better suited for global localization and loop-closure detection, which are necessary capabilities for bootstrapping any local localization algorithm, and to maintain geometrically consistent maps in lifelong operation [1].

For these reasons, we have developed a visual localization system, dubbed VIZARD, for the self-driving cars in the *UP-Drive* project[1], with the following main features:

1) We employ map-tracking, a local localization algorithm able to generate both accurate $6DoF$ pose estimates and achieve high localization recall.
2) Multi-session mapping techniques enable us to successfully tackle the challenge of long-term appearance change in outdoor environments, and even allow for localizing in night-time conditions.
3) The use of binary descriptors and an efficient sensor fusion backend renders real-time localization with CPU-only hardware set-ups feasible.

Fig. 1. We aim at accurately localizing the *UP-Drive* vehicle depicted in the upper-left corner in a map of visual features depicted on the right side. Features are extracted from images of the surround-view camera system (lower-left corner) and matched against $3D$ landmarks in the map. Inlier matches, centered on the estimated $6DoF$ pose of the vehicle in the map, are illustrated as dark yellow lines on the right side.

In a thorough evaluation of all crucial aspects of our localization system using two long-term outdoor dataset collections, one of them publicly available, we carefully analyze the most important parameters in our pipeline, compare the use of different binary descriptors, investigate key performance metrics such as localization accuracy and recall and relate to a state-of-the-art metric global localization algorithm. We see the main added value of this paper in sharing with the community the insights gained in this long-term study.

The contributions of this paper are thus as follows:

- We thoroughly study the critical parameters of our localization system, analyze their boundary conditions, and share our gained insights.

[1]The **UP-Drive** project is a research endeavor funded by the European Commission, aiming at advancing research and development towards fully autonomous cars in urban environment. See www.up-drive.eu.

- From a comparison of the localization performance using different binary descriptors, we show which descriptors are best suited for map-tracking across long-term appearance change.
- In an extensive evaluation on multiple long-term dataset collections, we demonstrate state-of-the-art localization performance across vastly different appearance conditions in outdoor environments, including at night time.

II. RELATED WORK

Visual localization systems can be divided into two main categories. Global localization systems are able to retrieve the location of a robot with no prior knowledge of the robot's pose. In contrast to that, local localization systems exploit a motion model to compute a prior on the robot location, thereby reducing the search space in the map.

Global Localization: Early visual global localization systems have been presented in the context of offline geometric scene reconstruction from a large number of images collected from varying viewpoints [3], [4]. These works led the foundation for many subsequent $6DoF$ global localization algorithms, and have been improved in numerous follow-up works [5]–[9]. More recently, deep learning techniques have given rise to novel global localization algorithms with remarkable robustness against drastic appearance change [10], [11]. They require, however, high-end GPUs in order to achieve real-time operation.

In general, the aforementioned global localization algorithms are capable of achieving reliable localization across significant appearance change in outdoor environments. However, as shown in [12], they often fall short of providing high recall with localization accuracies below $0.5m$, and are thus not well suited for our application, where we aim at permanently localizing our vehicle with sufficient accuracy to prevent deviation from the road lane boundaries. Note that there has also been a substantial amount of work on global localization in the realm of place recognition, or image retrieval [13]–[17]. These approaches, however, only provide a best matching image candidate in a map, instead of a $6DoF$ metric pose, and are thus addressing a different problem than ours.

Local Localization: Local localization algorithms take prior information on the robot pose into account, in order to reduce the localization search space and increase recall. This is well motivated in practice, as subsequent localization attempts of a mobile robot are far from independent, but in fact highly correlated in space, with the incremental motion between images often observable, although with drift, from odometry sensors such as wheel-odometry or IMUs. As a consequence, instead of regarding localization as an independent module, it can be directly integrated into the state estimation framework that optimizes the robot's pose in its environment by fusing odometry measurements and localization constraints. The ORB-SLAM [18] framework with its *localization mode* offers a local localization system similar to ours. They lack, however, the capability to integrate multiple sessions into a map, which greatly limits the robustness towards appearance change in outdoor environments. Lategahn and Schiller present a hierarchical visual localization system for outdoor environments that combines a global with a local localization module [19]. They achieve robustness against appearance change by employing DIRD descriptors [20]. Their experimental evaluation, however, only spans across six weeks, and it thus remains unclear, how well their system performs over long-term appearance change. Instead of employing an illumination invariant descriptor, Paton et al. use color-constant images to gain robustness against appearance change [21]. However, color-constancy primarily removes shadows under sunlight, but does not tackle other variations in appearance, such as seasonal change, or transitions from day to night-time.

Multi-Session Mapping: A common technique to achieve robustness against arbitrary long-term appearance change incorporates visual cues from multiple sorties through the environment in the map. We refer to this as multi-session mapping. Schneider et al. have presented a state estimation framework that fuses visual-inertial odometry with metric global localization [22]–[24]. While their mapping framework allows merging several sessions, they use a feature based global localization algorithm which prohibits sufficient localization recall in outdoor environments. Paton et al. present a visual localization system using multi-session mapping in [25]. Their application is, however, restricted to a teach-and-repeat scenario. In contrast to that, we employ loop-closure detection and bundle adjustment in order to get geometrically consistent multi-session maps, which adds additional flexibility in route planning and navigation. The "Experienced-Based Mapping" framework developed by Churchill et al. maintains separate map instances for diverse appearance conditions [26]. This allows visual localization in arbitrarily appearance conditions in an elegant and efficient manner. However, the maintenance of separate maps for differing conditions renders it impossible to share visual cues between sessions, which can increase recall. Furthermore, an integration of the localization module into a complete navigation stack is more challenging, as the visual pose estimates are expressed with respect to separate, disconnected coordinate frames.

Similarly to our localization system, the works presented in [27]–[30] use a local localization algorithm with multi-session maps for localization. As opposed to our work, Mühlfellner et al., refrain from fusing their visual pose estimates with wheel-odometry, which limits the accuracy and smoothness of their pose estimation framework, while Sons et al. do not report on localization accuracy and recall in long-term experiments.

III. METHODOLOGY

The VIZARD system consists of the following main components, presented at the beginning of this section. a) We employ a state estimation framework for fusing wheel-odometry and visual localization constraints. b) Our map-tracking module matches keypoints extracted from current

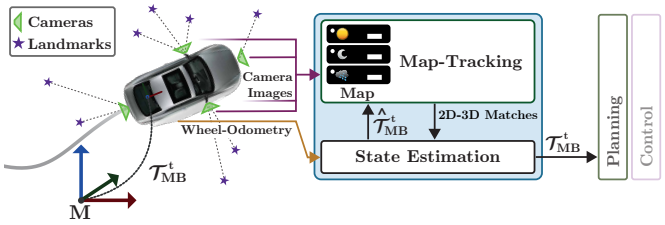

Fig. 2. The map-tracking module extracts 2D features from current camera images, and matches them with 3D map landmarks locally in image space using a pose prior $\hat{T}^t_{MB}$. The state estimation module fuses the visual 2D-3D matches with the current wheel-odometry measurement to obtain a current vehicle pose estimate T^t_{MB}.

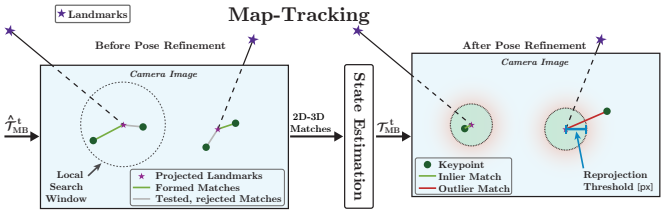

Fig. 3. The map-tracking module extracts 2D features from current camera images, and matches them with 3D map landmarks locally in image space using a pose prior $\hat{T}^t_{MB}$. The state estimation module fuses the visual 2D-3D matches with the current wheel-odometry measurement to obtain a current vehicle pose estimate T^t_{MB}.

camera images to landmarks from the map. Furthermore, key information regarding our mapping pipeline is provided at the end of this section, and a schematic overview of VIZARD can be found in Figure 2.

A. State Estimator

At the core of our localization system we employ a state estimation framework in information form, the dual representation of the (Extended) Kalman-Filter [31], [32]. Our state representation entails an estimate of the current transformation between the vehicle body coordinate frame $\mathcal{F}_B$ and the map reference frame $\mathcal{F}_M$ for each time-step t: $\mathbf{x}_t := [T^t_{MB}]$. Note that T_{MB} is an element of $SE(3)$, and thus represents all six degrees of freedom. The corresponding rotations are represented by unit quaternions. At every time-step t, a set of n simultaneously recorded camera images, and a relative odometry transformation measurement $\bar{T}^{odo}_{B_{t-1}B_t}$ are received. A new state is created by forward-propagating the previous state estimate using the odometry measurement: $\hat{T}^t_{MB} := T^t_{MB}\bar{T}^{odo}_{B_{t-1}B_t}$. It is used both in the map-tracking module as a pose prior, and as an initial linearization point in the filter update. After finding 2D-3D matches in the current set of images using map-tracking, the states are updated by retrieving the *MAP* estimate of the following cost function:

$$\begin{aligned} c(T^t_{MB}, T^{t-1}_{MB}) :=& \parallel f_{prior}(T^{t-1}_{MB}) \parallel^2_{P^{-1}_{t-1}} \\ &+ \parallel f_{odo}(T^t_{MB}, T^{t-1}_{MB}) \parallel^2_Q \\ &+ \sum_{i=1}^{m} \varphi(f_{loc}(T^t_{MB})) \end{aligned}$$

The prior pose and odometry factors, f_{odo} and f_{prior} respectively, follow a standard quadratic loss expression, while the localization re-projection factors f_{loc} employ a *Huber* robust cost function φ to account for possible wrong keypoint-landmark associations. All factors follow a standard graph SLAM formulation, as described in [1]. We retrieve the *MAP* estimate by iteratively minimizing the cost function c using the Levenberg-Marquardt in the GTSAM framework [33].

B. Map-Tracking

At every timestep t, the forward-propagated pose $\hat{T}^t_{MB}$ represents a rough estimate of the vehicle's location at time t in the map. With this, we can retrieve all landmarks from the map that have been observed from within a given distance around $\hat{T}^t_{MB}$. Using $\hat{T}^t_{MB}$ and the extrinsics calibration between the vehicle body and the individual camera frames, the landmark $3D$ points are projected into the current set of images, and matched with extracted keypoints in the following way: A landmark and a keypoint are only considered as a match candidate if their image space distance is smaller than $40px$. This avoids forming geometrically inconsistent matches. Further, a keypoint is preferably matched with the candidate landmark whose FREAK [34] descriptor is closest to the FREAK descriptor of the keypoint, using the Hamming distance metric. A *descriptor distance threshold* $\delta[px]$ is employed to limit the distance between the two descriptors, thus ensuring appearance consistency. The resulting 2D-3D matches are fed back into the state estimator where they form the visual localization constraints for the state update at time t. After optimizing the vehicle pose T^t_{MB}, the geometric consistency of every localization constraint is re-evaluated. For this, a *reprojection threshold* ρ[px] is used to distinguish between inlier and outlier landmark observations. While the localization factors of outlier observations are removed, the localization factors of inlier observations are marginalized out together with the previous pose T^{t-1}_{MB}. An illustration of the map-tracking algorithm can be found in Figure 3.

C. Mapping

A base-map is built by tracking and triangulating local features along the trajectory of the first-session dataset. The resulting $3D$ landmark points are added to the map together with their median feature descriptors. Subsequently, more map sessions are added by localizing further datasets against the available (multi-)session map using map-tracking. Note that all landmarks in the resulting multi-session map are expressed in one common frame of reference $\mathcal{F}_M$. Similar local localization and mapping algorithms have been used in our previous work [27], [35], to which we kindly refer the interested reader for more details.

IV. Evaluation

This section presents evaluation results on the following three key aspects. a) In a parameter study, the optimal values for the most important parameters of our localization system are identified. b) We further investigate the influence of different binary descriptors on the localization performance.

c) In long-term experiments across vastly different appearance conditions in outdoor environments, the localization accuracy and recall using map-tracking are evaluated, and compared with the accuracy and recall resulting from using a global localization algorithm.

The subsequent section first describes the *UP-Drive* vehicle platform, including the sensor set-up, computing infrastructure, and provides details on the online operation. Additional sections are devoted to a brief description of the three dataset collections, and the evaluation metrics used in our experiments.

A. The UP-Drive Platform

The *UP-Drive* vehicle is equipped with a surround-view camera system consisting of four cameras with fish-eye distorted lenses. Images are recorded at $30Hz$ with a resolution of 640×400 pixels in gray-scale. Furthermore, wheel tick encoders and a low-end IMU provide odometry measurements, which are fused with the visual localization constraints as described in Section III-B. The vehicle and sample images from the camera system are depicted in Figure 1. Localization is run in real-time at $10Hz$ on a consumer-grade computer with an Intel i7 CPU and 16GB of RAM. In particular, no GPU is required, neither for mapping, nor for localization. Furthermore, for bootstrapping map-tracking, a position prior is generated with a consumer-grade GPS sensor, while the orientation hypothesis is generated from orientations of near-by map poses.

B. Dataset Collections

1) UP-Drive: The *UP-Drive* dataset collection consists of 32 drives on the Volkswagen factory premises in Wolfsburg, Germany, recorded between December 2017 and December 2018. The total driving distance is approximately $300km$. The scenery resembles an urban environment, with busy streets, buses, zebra crossings, and pedestrians[2]. This dataset collection not only covers seasonal appearance changes and a wide range of different weather conditions, it also contains datasets recorded at dusk and night-time. Five datasets, three from day-time, one at dusk, and one at night, are used to build a multi-session map. The remaining 27 datasets are used for evaluating the localization.

2) NCLT : The *NCLT* [36] dataset collection consists of 27 recordings collected with a Segway platform on the Michigan University campus between January 2012 and April 2013. Analogous to the *UP-Drive* datasets, odometry poses based on wheel-tick encoders and an IMU sensor are fused in the state estimation framework. A *Ladybug 3* camera system is used, collecting images at $5Hz$ which are undistorted and down-scaled to a resolution of 808×616 pixels prior to being fed into our framework. The visited routes vary considerably from dataset to dataset. However, there is an approximately $750m$ long outdoor segment that is traversed, with some minor deviations, in almost all datasets

[2]Sample images can found online at `https://github.com/ethz-asl/up-drive_visual_dataset/wiki/Sample-Images`

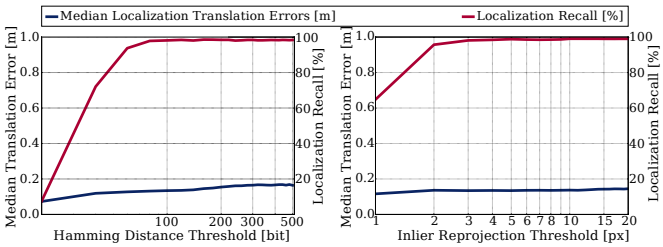

Fig. 4. Localization recall $\mathbf{r}_{mt}$ (red) and median translation accuracy $\tilde{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xyz}}$ (blue) on the *NCLT* dataset from January 8^{th} 2012, in relation to increasing values of the *descriptor distance threshold* δ (left), and *reprojection threshold* ρ (right) respectively, on a logarithmic scale. Even for very high values of δ and ρ, the vehicle remains accurately localized.

in either one or the opposite direction. We therefore use this sub-segment of the campus for building a multi-session map using seven of the datasets. The remainder of the datasets are used for evaluating the localization. Similar to the *UP-Drive* datasets, the *NCLT* datasets cover seasonal and weather changes over an annual cycle.

3) KITTI : We further use $Sequence$ 00 of the *KITTI* [37] visual odometry benchmark dataset in our evaluation. It is the only *KITTI* dataset with significant segments of the trajectory revisited. We split the dataset in two parts, and use the first 170 seconds for mapping, and the remainder for localization. As opposed to the *UP-Drive* and the *NCLT* datasets, the appearance conditions in the *KITTI* drive thus remain similar between mapping and localization.

C. Metrics

1) Localization Recall: We measure localization recall $\mathbf{r}[\%]$ as the distance traveled while localized in relation to the total distance traveled in the respective dataset. Localization at time t is deemed successful if there are at least 10 inlier landmark observations after the pose optimization.

2) Localization Accuracy: The $6DoF$ localization accuracy is evaluated for each successfully localized set of images along the trajectory of a dataset by comparing the relative transformation between the estimated pose T^t_{MB} and the nearest vertex in the map, with the same quantity estimated by a reference solution [38]. For the *NCLT* datasets, ground-truth poses are available, which we employ to evaluate both the translation accuracy $\mathbf{p}^{\mathbf{e}}_{\mathbf{xyz}}$[m], and orientation accuracy $\theta^{\mathbf{e}}_{\mathbf{xyz}}$[deg]. Note that the availability of ground-truth poses is a unique feature of *NCLT*, and the primary reason why we have decided to evaluate on the *NCLT* datasets, in addition to our self-collected *UP-Drive* datasets.

For the *UP-Drive* and *KITTI* datasets, no ground-truth poses are available. Both dataset collections provide, however, poses estimated by an *RTK GPS* sensor, which we use for producing a rough estimate of the localization accuracy on these datasets. Since the *RTK GPS* altitude estimates are unreliable, we only report on planar $\mathbf{p}^{\mathbf{e}}_{\mathbf{xy}}$[m] and lateral translation errors $\mathbf{p}^{\mathbf{e}}_{\mathbf{y}}$[deg] on the *UP-Drive* and *KITTI* datasets.

D. Map-Tracking Parameter Study

As described in Section III-B, there are two main parameters guiding the formation of 2D-3D localization constraints

in the map-tracking module, namely the *descriptor distance threshold* δ[bits], and the *reprojection threshold* ρ[px]. While the *descriptor distance threshold* ensures appearance consistency by setting an upper bound on the descriptor distance for matching 2D keypoints with a 3D landmarks, the *reprojection threshold* enforces geometric consistency by discarding localization constraints if their respective reprojection error after the pose update is more than ρ pixels.

In Figure 4, the localization recall and median localization error are shown for increasing values of δ, and ρ respectively, for the *NCLT 2012-01-08* dataset. A fixed value of $\delta = 100$bits, and $\rho = 3$px is used unless the respective parameter is varied as indicated on the x-axis. As expected, localization recall quickly rises with increasing δ and ρ. Interestingly, the localization accuracy remains approximately constant, even for high values of δ and ρ. This may appear counter-intuitive at first. A descriptor distance threshold greater than 25% of the total descriptor length clearly allows for many wrong matches to be formed, and eventually ought to lead to false positive localizations. In order to understand why this scenario does not occur, it is important to note that, as described in Section III-B, the *descriptor distance threshold* only serves to discard matches whose descriptor distance is above δ bits. If there are multiple match candidates for a given keypoint in the image, the matching algorithm still attempts to pick the landmark with the lowest descriptor distance. Therefore, as long as there *are* sufficiently many correct matches that can be formed, our algorithm *will* find them, even with a very lean *descriptor distance threshold* δ.

A similar effect exists for the *reprojection threshold* too. As long as the pose prior is close to correct and there are sufficiently many valid 2D-3D matches, localization will not deviate from the correct trajectory, even with a very high ρ threshold.

Hence, as long as the vehicle is correctly localized, and there are enough valid localization matches *possible*, our localization system will *remain* correctly localized,

However, too high a value for δ and ρ may indeed derail the localization system if the pose prior is sufficiently wrong. In the remainder of this section, we therefore aim at finding the range of values for δ and ρ that guarantee no false-positive localization, even if the pose prior is wrong. Knowing this range is important in two ways. Firstly, it defines a safe operating space for choosing δ and ρ where a positive localization feedback, such as a certain number of inlier landmark observations, can be trusted. Secondly, it reveals a maximum degree of pose prior disturbance that can be tolerated when bootstrapping the map-tracking algorithm with any kind of auxiliary global localization input such as consumer grade GPS, or a place-recognition module. In order to evaluate these properties, we have conducted a parameter sweep experiment, varying both values for δ and ρ, as well as increasing the disturbance of the prior pose in yaw-angle, longitudinal, and lateral dimension separately. The resulting range of safe operating conditions is shown in Figure 5. The colors indicate the maximum disturbance, before either bootstrapping map-tracking is no longer possible, or, marked with an 'X', bootstrapping resulted in false-positive localization instead. It can be seen that for all three modes for disturbance, there is a safe range for both δ, and ρ, that guarantee convergence to correct localizations, even for considerably inaccurate prior poses with up to 10 degrees in yaw angle, and $3m$ meters in longitudinal and lateral direction. Furthermore, taking the results from both Figure 4, and 5, we find with $\delta = 100bits$, and $\rho = 3.0px$, a safe choice of parameters yielding maximum recall and sufficient robustness for bootstrapping map-tracking with a consumer-grade GPS sensor.

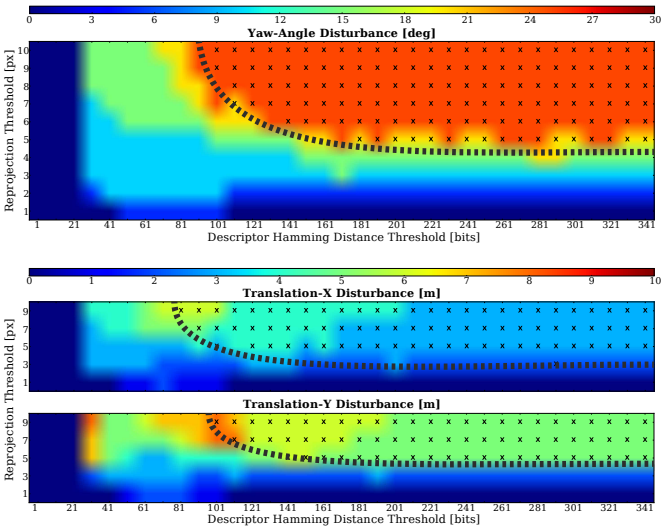

Fig. 5. Sensitivity of the map-tracking pose prior in relation to increasing values for the *descriptor distance threshold* δ, and *reprojection threshold* ρ. The colors represent the maximum admissible degree of disturbance in yaw angle (top), longitudinal (middle), and lateral direction (bottom) leading to convergence of the localization to the true pose. The parameter combinations marked with 'X' denote unsafe operating regions, where high prior pose disturbances lead to false-positive localizations. In the remaining operating regions, localization fails if the prior pose disturbance is larger than the degree represented by the respective color. All three modes of disturbances reveal a safe region for the choice of δ and ρ allowing for guaranteed convergence towards the correct pose, while tolerant to significant disturbance in the prior pose.

E. Binary Descriptor Comparison

In addition to the *descriptor distance threshold* and the *reprojection threshold*, the type of descriptor is another pivotal design choice, as it influences not only the localization recall, but also the size of the map. We restrict ourselves to the use of binary descriptors, as they can be matched very efficiently on a CPU-only platform, and compare the localization performance for three popular choices of binary local feature descriptors, namely FREAK [34], BRISK [39], and ORB [40]. Krajnik et al. have evaluated the influence of various local feature descriptors for visual teach-and-repeat in [41]. They have, however, employed a global matching algorithm to find correspondences between two images recorded at the same location. In contrast to that, our map-tracking algorithm employs a pose prior and performs a local search in the image space. This variation in methodology leads to differing results as compared to [41].

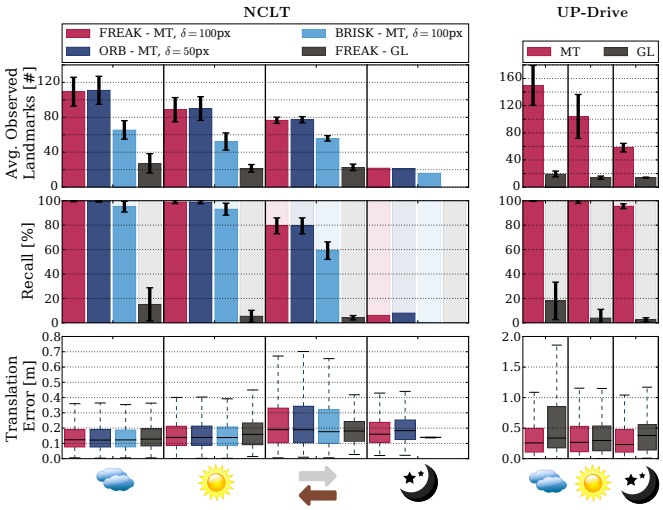

Fig. 6. Average number of observed landmarks (top), localization recall (middle), and translation accuracy on the *NCLT* datasets (left), and *UP-Drive* datasets (right). For the *NCLT* datasets, the translation localization accuracy is evaluated using the *ground-truth* poses, while for the *UP-Drive* datasets, the planar translation errors with respect to the *RTK GPS* poses are shown. The datasets are grouped into categories according to appearance conditions (cloudy or rainy, sunny, and night-time) and traversal direction (indicated by the two opposing arrows). The localization performance using map-tracking (MT) is compared with global localization (GL). On the *NCLT* datasets, the map-tracking performance is further compared with different choices of binary descriptors.

While the evaluation by Krajnik et al. suggests superior performance of BRISK as compared to FREAK and ORB, our experiments reveal worse performance of map-tracking with BRISK than with FREAK or ORB. This emphasizes the strong influence of the specific feature matching algorithm with respect to the localization performance using different types of features. In Table I, the localization recall, the average number of observed landmarks, and the localization accuracy, are presented for different choices of descriptors, aggregated over all day-time *NCLT* datasets. Note that the *descriptor distance threshold* δ is set at 100bits for the two 64 byte long descriptors FREAK and BRISK, and at 50bit for the 32 byte long ORB descriptors, thereby allowing the same relative fraction of bits to be different when forming localization matches in all three cases. The performance using FREAK and ORB is nearly identical. This is remarkable, as the descriptor size of the latter is only half of that of FREAK. With BRISK, on the other hand, the average number of observed landmarks and the localization recall is significantly worse. However, the impact on the localization accuracy is marginal, as only the pose estimates of successful localizations are considered.

A more detailed evaluation of the descriptor comparison can be found in Figure 6, which shows the aforementioned metrics evaluated separately for groups of datasets formed according to the four categories exhibiting differing localization performance. The category *Cloudy* includes four, and the category *Sunny* 12, datasets labeled as (partially) cloudy, and sunny respectively, according to [36]. The category *Opposite* contains the two day-time datasets *2012-11-04*, and *2013-02-23* traversing the map in the opposite direction, while

	FREAK	**ORB**	**BRISK**
$\mathbf{r}_{mt}[\%]$	**96.89** +/- 6.62	**96.76** +/- 6.61	89.75 +/- 12.0
$\mathbf{obs}_{\varnothing}[\#]$	**92.97** +/- 49.94	**94.28** +/- 51.38	56.28 +/- 33.12
$\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xyz}}[\mathrm{m}]$	0.14 [0.32]	0.14 [0.32]	0.14 [0.3]

TABLE I

DESCRIPTOR COMPARISON ON THE *NCLT* DATASETS, SHOWING THE AVERAGE RECALL WITH MAP-TRACKING $\mathbf{r}_{mt}[\%]$, THE AVERAGE NUMBER OF OBSERVED LANDMARKS $\mathbf{obs}_{\varnothing}[\#]$, WITH STANDARD DEVIATIONS DENOTED BY "+/-", AND THE MEDIAN TRANSLATION LOCALIZATION ACCURACY $\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xyz}}[\mathrm{m}]$, WITH THE 90 PERCENTILE DENOTED IN SQUARE BRACKETS.

the *Night* category represents the only night-time dataset from December 1^{st} 2012. The loss in recall with BRISK is primarily attributed to the two datasets traversing the map in opposite direction, where the recall with BRISK is approximately 20% lower than with FREAK or ORB. Contrary to that, the average number of observed landmarks remains roughly the same with BRISK across all three day-time categories, while FREAK and ORB observe significantly more landmarks when traversing the map in the predominant direction, both under cloudy skies, and in sunny conditions.

Based on these experiences, we suggest to use ORB as a binary descriptor for map-tracking, or FREAK in case there are no restrictions with respect to the map size.

F. Localization Accuracy and Recall

In order to fully rely on our visual localization system to control the car in the *UP-Drive* project, a high localization recall with an accuracy below $0.5m$ is paramount, as only short driving segments with no localization may be bridged with wheel-odometry before the car may deviate from its designated lane. We compare the localization recall and accuracy of our localization system using map-tracking with the metric global localization algorithm based on the work presented in [24] and available in the *maplab* framework [22]. We refer to the results using this algorithm with GL in the respective figures and tables. Both algorithms, that is map-tracking and global localization, operate on the same multi-session maps, using the same landmarks. Note, however, that the global localization algorithm is fundamentally different to the map-tracking module presented in this paper, as in contrast to the former, the latter is able to exploit a pose prior. By including this comparison, we aim at highlighting the gain in localization recall attainable by using a local localization algorithm, as opposed to relying only a global localization algorithm. Map-tracking does, however, require *some* global localization module for bootstrapping, or re-localizations. As described in Section IV-A, a consumer grade GPS sensor serves this role on the *UP-Drive* vehicles.

The localization recall with map-tracking $\mathbf{r}_{mt}[\%]$, and with global localization $\mathbf{r}_{gl}[\%]$, and the localization accuracy are shown in Table II, aggregated over all datasets of the three collections. Note that the *NCLT* night-time dataset from December 1^{st} is excluded in the table. While map-tracking attains close to 100% recall on all three dataset collections, global localization fails for extended periods on the *NCLT* and *UP-Drive* datasets which both exhibit pronounced

	$\mathbf{r}_{mt}$[%]	$\mathbf{r}_{gl}$[%]	$\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xyz}}$ / $\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xy}}$, $\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{y}}$		$\bar{\theta}^{\mathbf{e}}_{\mathbf{xyz}}$
NCLT	**96.89** +/-6.62	7.49 +/-8.59	0.14 [0.32]		1.23 [1.8]
UP-Drive	**99.23** +/-1.75	8.94 +/-12.62	0.26 [0.88]	0.12 [0.58]	0.21 [0.33]
KITTI	**96.05**	94.24	0.43 [0.8]	0.31 [0.62]	0.26 [0.59]

TABLE II

THE AGGREGATED LOCALIZATION PERFORMANCE ON THE *NCLT* , *UP-Drive* , AND *KITTI* DATASET(S), SHOWING AVERAGE LOCALIZATION RECALL WITH MAP-TRACKING $\mathbf{r}_{mt}$, AND WITH GLOBAL LOCALIZATION $\mathbf{r}_{gl}$, AND THE MEDIAN TRANSLATION ($\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xyz}}$) AND ORIENTATION ($\bar{\theta}^{\mathbf{e}}_{\mathbf{xyz}}$) ACCURACY. FOR *UP-Drive* AND *KITTI*, PLANAR $\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{xy}}$ AND LATERAL $\bar{\mathbf{p}}^{\mathbf{e}}_{\mathbf{y}}$ ERRORS ARE SHOWN INSTEAD OF FULL $3DoF$ TRANSLATION ERRORS. STANDARD DEVIATIONS ARE DENOTED BY "+/-", AND THE 90 PERCENTILE IS SHOWN IN SQUARE BRACKETS.

appearance change. On the *KITTI* drive, however, the appearance condition only undergo minor change, and global localization achieves with 94% a similarly high recall as map-tracking. This illustrates the challenge in finding enough correct feature matches with a global localization algorithm in multi-session maps that cover outdoor environments with various different appearance conditions. Solely relying on a global localization algorithm in these environments may thus not be sufficient to guarantee reliable localization in real-world applications. As our results show, exploiting a pose prior can help to significantly increase the reliability of the localization.

We further note that the planar median localization accuracy in *UP-Drive* and *KITTI* are below $0.5m$. Note that due to the different kind of reference sensors, these numbers are not directly comparable with the localization accuracy attained on the *NCLT* datasets, with the latter exhibiting a median translational accuracy of $11cm$. Furthermore, the *KITTI* vehicle is equipped with only a forward facing camera, while the *UP-Drive* vehicle has a surround view camera rig. This results in less strictly constraint position estimates on the *KITTI* dataset, which translate into significantly lower planar and lateral localization accuracy. For the driving performance, the lateral errors are most important. On the *UP-Drive* datasets, the median lateral error is below $15cm$, which is sufficient for a smooth steering of the car.

The median orientation errors are less effected by the difference in the camera rigs, and are well below 0.5 degrees for both the *UP-Drive* and *KITTI* datasets. In contrast to that, the orientation errors on the *NCLT* datasets are higher due to more vivid roll and pitch motions of the Segway platform, as compared to the car platforms in case of *UP-Drive* and *KITTI*.

A more detailed analysis of the localization recall and planar accuracy on the *NCLT* and *UP-Drive* datasets is shown in Figure 6, with datasets grouped into categories as described in Section IV-E. There are 10 drives of the *UP-Drive* dataset collection in the *Cloudy*, 15 in the *Sunny*, and two in the *Night* category respectively. Recordings in rainy conditions are categorized as *Cloudy*, since there is little difference in performance on rainy datasets as opposed to in dry cloudy conditions. Map-tracking reaches virtually 100% recall with a median localization accuracy of around $10cm$ for all the *NCLT* day-time datasets that traverse the map in the primary direction. The same high recall is also achieved for all day-time *UP-Drive* datasets, with a planar median localization accuracy with respect to RTK GPS of approximately $20cm$. The additional challenge for visual localization in sunny conditions is, however, reflected in a lower average number of observed landmarks in case of map-tracking, and in a significantly worse recall using global localization. Recall using map-tracking remains, however, unaffected.

In contrast to that, map-tracking performs significantly worse on the two *NCLT* datasets that traverse the map in the opposite direction, with considerably lower recall, and slightly lower localization accuracy. This is understandable, given that there is only one map session in opposite direction, whereas there are six traversing the map in the primary direction. However, this also reveals the limitations of matching landmarks projected into the cameras field-of-views under considerable viewpoint change. Here it is important to note the asymmetry of the *Ladybug* camera rig when traversing in the opposite direction, as the surround view is covered by an odd number of five cameras.

Furthermore, the only *NCLT* night-time dataset from December 1^{st} 2012 fails to localize along most parts of the trajectory. Not only is this the only available recording under night-time conditions, but the Segway also traverses the map in the opposite direction, further exacerbating localization. A lack of more recordings from dusk or night-time renders it impossible to augment the multi-session map with the appearance conditions at night-time, and thus the conditions in this dataset lie outside the appearance coverage of the map. In contrast to the *NCLT* datasets, the *UP-Drive* datasets contain multiple recordings under both dusk and night-time conditions, allowing to extend the appearance coverage of the multi-session map with these conditions. Therefore, localization at night is successful in this case, even though the respective average recall is slightly less than 100% for the *UP-Drive* night-time datasets. This minor drop in recall is mainly attributed to the night-time recording from December 11^{th}, which only attains a recall of 92%. Sample images of the route segment where localization fails on this dataset are depicted in Figure 7. In this part of the route, the car is driving up North on a ramp crossing numerous rail tracks. With a lack of both street lamps and near-by building structures, there are hardly any stable visual cues in this section, and our localization system fails to match sufficiently many landmarks from the map. Only later along the ramp, once artificial lighting on the railing to the left and right of the road boundary is present, localization is picked up

Fig. 7. On the left, a sample image of a trajectory segment that fails to localize at night. A lack of structure and street lighting renders it unfeasible to match a sufficient number of map landmarks. A few meters later, street lighting is present (right side), and localization is picked up again.

again. This example demonstrates the current limitations of visual localization in night-time conditions. Even with high-performance CMOS cameras providing remarkably bright images at night, a certain amount of artificial street lighting and human made structure in the vicinity is required.

V. CONCLUSIONS

This paper presented a reliable visual localization system for urban outdoor environments. An extensive evaluation on several hundreds of kilometers of real-world driving conditions over the course of more than a year has demonstrated that our localization system is able to meet the requirement of high localization recall at high accuracy. Thereby, the appearance conditions encountered in our experiments not only cover various challenging weather conditions, wet road surfaces, sun reflections, and seasonal changes, but also night-time conditions. A comparison with a state-of-the-art global metric localization algorithm has revealed a large increase in recall attainable by instead employing a local localization algorithm, such as the map-tracking algorithm described in this paper. Additionally, a comparison of binary feature descriptors suggests superior performance of map-tracking when using FREAK or ORB, as compared to using BRISK. In a thorough parameter study, we have further investigated the boundary conditions of our map-tracking module and validated a safe range for selecting the most critical parameters in order to guarantees reliable localization.

ACKNOWLEDGMENT

This project has received funding from the EU H2020 research project under grant agreement No 688652 and from the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 15.0284.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *TRO*, 2016.

[2] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual Place Recognition: A Survey," *TRO*, 2016.

[3] G. Schindler, M. Brown, and R. Szeliski, "City-scale location recognition," in *CVPR*, 2007.

[4] S. Agarwal, N. Snavely, I. Simon, S. M. Seitz, and R. Szeliski, "Building Rome in a day," in *ICCV*, 2009.

[5] Y. Li, N. Snavely, and D. P Huttenlocher, "Location Recognition using Prioritized Feature Matching," *ECCV*, 2010.

[6] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *ICCV*, 2011.

[7] L. Liu, H. Li, and Y. Dai, "Efficient global 2d-3d matching for camera localization in a large-scale 3d map," in *ICCV*, 2017.

[8] S. Lynen, M. Bosse, P. Furgale, and R. Siegwart, "Placeless Place-Recognition," in *IC3DV*, 2014.

[9] M. Cummins and P. Newman, "Appearance-only SLAM at large scale with FAB-MAP 2.0," *IJRR*, 2011.

[10] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," *arXiv*, 2017.

[11] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," *arXiv*, 2018.

[12] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic *et al.*, "Benchmarking 6dof outdoor visual localization in changing conditions," in *CVPR*, 2018.

[13] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *ICRA*, 2012.

[14] T. Naseer, L. Spinello, W. Burgard, and C. Stachniss, "Robust visual robot localization across seasons using network flows." in *AAAI*, 2014.

[15] W. Maddern, M. Milford, and G. Wyeth, "Cat-slam: probabilistic localisation and mapping using a continuous appearance-based trajectory," *IJRR*, 2012.

[16] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016.

[18] R. Mur-Artal and J. D. Tardós, "Visual-inertial monocular slam with map reuse," *RAL*, 2017.

[19] H. Lategahn and C. Stiller, "Vision-only localization," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1246–1257, 2014.

[20] H. Lategahn, J. Beck, and C. Stiller, "DIRD is an illumination robust descriptor," in *IV*, 2014.

[21] M. Paton, K. MacTavish, C. J. Ostafew, and T. D. Barfoot, "It's not easy seeing green: Lighting-resistant stereo Visual Teach & Repeat using color-constant images," in *ICRA*, 2015.

[22] T. Schneider, M. Dymczyk, M. Fehr, K. Egger, S. Lynen, I. Gilitschenski, and R. Siegwart, "maplab: An open framework for research in visual-inertial mapping and localization," *RAL*, 2018.

[23] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *IJRR*, 2017.

[24] S. Lynen, T. Sattler, M. Bosse, J. A. Hesch, M. Pollefeys, and R. Siegwart, "Get out of my lab: Large-scale, real-time visual-inertial localization." in *RSS*, 2015.

[25] M. Paton, K. Mactavish, M. Warren, and T. D. Barfoot, "Bridging the appearance gap: Multi-experience localization for long-term visual teach and repeat," in *IROS*, 2016.

[26] W. Churchill and P. Newman, "Experience-based Navigation for Long-Term Localisation," *IJRR*, 2013.

[27] P. Mühlfellner, M. Bürki, M. Bosse, W. Derendarz, R. Philippsen, and P. Furgale, "Summary Maps for Lifelong Visual Localization," *JFR*, 2016.

[28] P. Muehlfellner, P. Furgale, W. Derendarz, and R. Philippsen, "Evaluation of fisheye-camera based visual multi-session localization in a real-world scenario," in *IV*, 2013.

[29] M. Lauer, C. G. Keller, C. Stiller *et al.*, "Mapping and localization using surround view," in *IV*, 2017.

[30] M. Sons and C. Stiller, "Efficient multi-drive map optimization towards life-long localization using surround view," in *ITSC*, 2018.

[31] M. Burri, M. Bloesch, D. Schindler, I. Gilitschenski, Z. Taylor, and R. Siegwart, "Generalized information filtering for mav parameter estimation," in *IROS*, 2016.

[32] H. Strasdat, J. M. Montiel, and A. J. Davison, "Visual slam: why filter?" *IVC*, 2012.

[33] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," Georgia Institute of Technology, Tech. Rep., 2012.

[34] A. Alahi, R. Ortiz, and P. Vandergheynst, "Freak: Fast retina keypoint," in *CVPR*, 2012.

[35] M. Burki, I. Gilitschenski, E. Stumm, R. Siegwart, and J. Nieto, "Appearance-based landmark selection for efficient long-term visual localization," in *IROS*, 2016.

[36] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and lidar dataset," *IJRR*, 2016.
[37] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *IJRR*, 2013.
[38] W. Burgard, C. Stachniss, G. Grisetti, B. Steder, R. Kümmerle, C. Dornhege, M. Ruhnke, A. Kleiner, and J. D. Tardós, "A comparison of SLAM algorithms based on a graph of relations," in *IROS*, 2009.
[39] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *ICCV*, 2011.
[40] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *ICCV*, 2011.
[41] T. Krajník, P. Cristoforis, K. Kusumam, P. Neubert, and T. Duckett, "Image features for visual teach-and-repeat navigation in changing environments," *RAS*, 2017.