

Efficient 2D-3D Matching for Multi-Camera Visual Localization

Marcel Geppert¹, Peidong Liu¹, Zhaopeng Cui¹, Marc Pollefeys^{1,2}, Torsten Sattler¹

Abstract—Visual localization, *i.e.*, determining the position and orientation of a vehicle with respect to a map, is a key problem in autonomous driving. We present a multi-camera visual inertial localization algorithm for large scale environments. To efficiently and effectively match features against a pre-built global 3D map, we propose a prioritized feature matching scheme for multi-camera systems. In contrast to existing works, designed for monocular cameras, we (1) tailor the prioritization function to the multi-camera setup and (2) run feature matching and pose estimation in parallel. This significantly accelerates the matching and pose estimation stages and allows us to dynamically adapt the matching efforts based on the surrounding environment. In addition, we show how pose priors can be integrated into the localization system to increase efficiency and robustness. Finally, we extend our algorithm by fusing the absolute pose estimates with motion estimates from a multi-camera visual inertial odometry pipeline (VIO). This results in a system that provides reliable and drift-less pose estimations for high speed autonomous driving. Extensive experiments show that our localization runs fast and robust under varying conditions, and that our extended algorithm enables reliable real-time pose estimation.

I. INTRODUCTION

Visual localization is the problem of estimating the position and orientation, *i.e.*, the camera pose, from which a given query image was taken. This problem plays a key role in autonomous navigation, *e.g.*, for self-driving cars [9] and in Simultaneous Localization and Mapping (SLAM) [27]. It is also encountered in many 3D computer vision algorithms such as Structure-from-Motion (SfM) [33], camera calibration [9], and Augmented Reality [23], [26].

State-of-the-art approaches for visual localization are *structure-based*, *i.e.*, they explicitly or implicitly use a 3D model to represent the scene. Explicit methods typically employ a sparse 3D point cloud constructed via SfM [18], [23], [30], [38], [44], allowing them to associate each 3D point with one or more local image descriptors. For a given query image, they establish a set of 2D-3D correspondences by comparing the descriptors of local features extracted from the image with the 3D point descriptors. Using these matches, they then estimate the camera pose of the query by applying an n -point-pose solver [10], [16], [17] inside a RANSAC loop [8]. In contrast, implicit approaches [4], [7], [25], [37] forego explicit descriptor matching. Instead, they directly learn the 2D-3D matching function by learning a mapping from image patches to 3D scene point coordinates. Again, the resulting 2D-3D correspondences are used for RANSAC-based pose estimation. Implicit approaches can achieve a

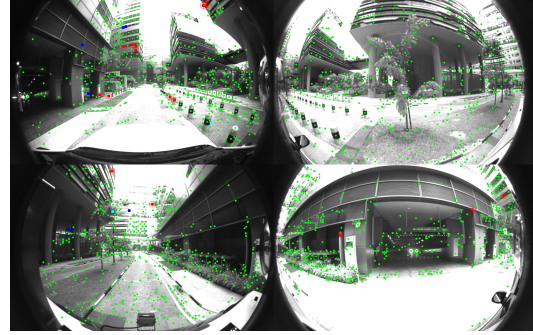


Fig. 1. Visual localization based on 2D-3D matching against a 3D scene model for a multi-camera system. We use one fisheye camera mounted on each side of a car. We show the correct matches (red) with their pose inlier threshold (red circles), outlier 2D-3D matches (dark blue), outlier 3D-2D matches (light blue), and unmatched image features (green).

higher pose accuracy compared to explicit ones [4], [7]. Yet, they currently do not scale to larger outdoor scenes [4], [35].

Most explicit structure-based localization methods focus on the monocular (single image) case, *e.g.*, Augmented Reality on smart phones and tablets [3], [15], [23], by developing more strategies for more efficient matching [19], [30] or for scaling to larger or more complex scenes [20], [32], [38], [44]. Yet, many robotics applications, especially self-driving cars [9], [36], benefit from using a multi-camera systems that covers the full 360° field-of-view (FoV) around the robot. It has also been shown that cameras covering a larger FoV can be localized more accurately [2] and that multi-camera systems significantly boost localization performance in challenging conditions [31].

Existing work on multi-camera localization has mainly focused on stereo SLAM [13], [21], [28], camera calibration [12], [13], and camera pose estimation [6], [17], [39], [41]. The latter two types of approaches model multi-camera systems as a generalized camera [29], *i.e.*, a camera with multiple centers of projection, to derive (minimal) solvers for pose estimation. Yet, one central aspect of multi-camera localization has received little attention: Using multiple images leads to more features that need to be considered during feature matching and thus to significantly longer run-times.

This paper aims to close this gap in the literature by focusing on efficient 2D-3D matching for multi-camera systems. To this end, we make the following main contributions: 1) We develop a prioritized descriptor matching scheme for multi-camera systems. Our strategy is based on Active Search [30], an efficient prioritization scheme developed for monocular cameras. We show that a fast variant of Active Search, which leads to unstable pose estimates for a single image, is very well suited for multi-camera systems. 2) We

¹Computer Vision and Geometry Group, Department of Computer Science, ETH Zürich, Switzerland

²Microsoft, Switzerland

interleave prioritized matching with camera pose estimation. In contrast to standard schemes, which terminate search once a fixed number of matches has been found, our approach terminates as soon as sufficiently many geometrically consistent matches have been found. 3) Inspired by approaches for geometric outlier filtering [38], [44], we develop an efficient geometric verification step that can be used to integrate potential pose priors. This allows us to avoid comparing descriptors for geometrically implausible matches, which can make our search both more efficient and robust. These later two contributions are not restricted to the multi-camera case but also applicable in the monocular scenario. 4) We show how to combine our approach with a VIO pipeline, enabling our system to provide accurate, drift-free pose estimates in real-time on a car.

II. RELATED WORK

In the following, we review related work from the areas of visual localization and multi-camera pose estimation.

Efficient visual localization approaches aim at accelerating the localization process [4], [7], [14], [19], [25], [30], [40], [43]. Most related to our approach are explicit methods based on prioritized matching [19], [30]. These methods aim at designing a prioritization function such that features that are more likely to yield 2D-3D matches are considered first. Once a fixed number of correspondences has been found, matching is terminated and RANSAC-based pose estimation is performed. In this paper, we build upon Active Search [30]. We show that a variant of it that is more efficient but leads to inferior results for monocular images is actually well-suited for multi-camera systems. We adapt the prioritization scheme to encourage distributing matches over many images in the camera system. We also propose an adaptive criterion that terminates matching once a certain number of correct matches is found rather than stopping search after finding a fixed number of (possibly incorrect) correspondences.

Scalable visual localization. In larger or more complex scenes, which are often characterized by more ambiguous scene elements, it is hard to distinguish between correct and incorrect matches based on descriptor comparisons alone. State-of-the-art methods for scalable localization thus relax the matching criteria and perform geometric reasoning to handle the resulting large amounts of wrong matches [5], [38], [42], [44]. As a result, they are often too slow for real-time processing. In this paper, we propose a geometric filter based on a potentially available pose prior, *e.g.*, from VIO-based camera tracking or via a GPS sensor. We show that this filter can be implemented very efficiently, allowing us to perform it before descriptor matching. This leads to faster matching times, but also makes matching more robust as we can again relax the matching criteria.

Learning-based methods integrate machine learning into the localization process. This is usually done by either learning the 2D-3D matching stage [4], [7], [25] or by directly regressing the camera pose [14], [40], [43]. However, recent work shows that these methods are either less accurate

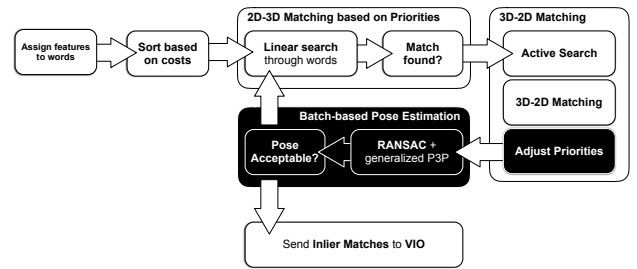


Fig. 2. The localization pipeline. White boxes denote components based on the Active Search approach from [30]. Black boxes correspond to our modifications made to adapt to the multi-camera system.

than feature-based methods such as the one presented in this paper [43] or do not scale to larger scenes [4], [14], [31], [35]. As such, we do not consider them further in this work.

Multi-camera pose estimation algorithms model multi-camera systems mathematically as a generalized camera [29]. Given a set of 2D-3D matches between features in the images of the multi-camera system and 3D scene points, they then estimate the pose of the generalized camera [6], [17], [39], [41]. These approaches can easily replace monocular n -point-pose solvers inside the RANSAC loop. Yet, we are only aware of a single multi-camera localization approach [31]. This method first applies Active Search on all images independently and then performs pose estimation as a post-processing step. In contrast, this paper proposes a variant of Active Search that jointly considers all images and we show that this approach leads to significantly faster run-times.

III. ALGORITHM

Our system consists of two main functional modules, the localization and pose fusion module. The localization module provides absolute pose estimates based on a prebuilt map. It uses two main processes, the prioritized feature matching, and the iterative RANSAC pose estimation. An overview over our approach is shown in Fig. 2. Furthermore, the localization module provides a pose prior-based candidate filtering solution that increases its efficiency and robustness when a pose prior is available. To provide reliable drift-less global pose estimation at camera framerate, we further fuse the 3D-2D feature matchings from our localization algorithm with the estimated local motions from VIO via a pose graph. We will describe the details as follows.

A. Map creation

We represent our global map with a sparse 3D point cloud. Each 3D point from the global map associates with one or multiple 2D feature descriptors, such that 3D-2D feature matching can be performed for visual localization.

We build our map based on the COLMAP SfM pipeline [34] by the aid of image synchronized GPS/INSS poses. In particular, since the captured full framerate images provide redundant information, we subsample a set of representative image frames for 3D reconstruction by using the pose prior provided by GPS/INSS. A guided 3D reconstruction based on COLMAP is then performed on these selected frames.

We extract SIFT features and perform pairwise feature matching among nearby images. The camera poses are

initialized with the pose priors from GPS/INSS. This step makes our approach significantly more scalable and robust compared to performing standard SfM from scratch. We perform multiple iterations of 3D point triangulation, point subsampling, bundle adjustment and outlier filtering. During bundle adjustment, we fix the extrinsic parameters among the cameras of the multi-camera system. Subsampling the point cloud before bundle adjustment significantly reduces the optimization complexity and runtime. After the camera poses are optimized, to increase the density of the final map, we sample 2D-3D correspondences from previously filtered feature matches and add back to the map. We also further perform multiple iterations of structure only optimization to improve the accuracy of the estimated 3D point positions.

For feature description and retrieval we closely follow the original Active Search approach. We divide the feature descriptor space into a visual vocabulary and assign each feature that is associated to a 3D point to its closest word. We average all descriptors that are assigned to both the same word and the same point, and store this average as a descriptor for the respective point. When matching an image feature to a point later, we can find the feature's corresponding word, and then perform linear search through the point descriptors assigned to this word.

B. Visual localization

We design our online visual localization algorithm by extending the Active Search approach from [30] for a multi-camera system. Our algorithm consists of the prioritized feature matching and iterative RANSAC pose estimation. To minimize the number of required feature matches and dynamically adapt the matching efforts to the surrounding environment, we run both components in parallel and add new feature matches iteratively to the pose estimation until a valid pose is found, or no more image features are available.

Prioritized Feature Matching for Multi-Camera-Systems:

Using all features that can be found in a whole frame usually leads to far more correspondences than required to reliably localize frame. In an ideal case, three 2D-3D correspondences are sufficient to compute the pose. However, due to the presence of outliers a larger number of constraints is required to verify the estimated pose. Still, this number is far below the total number of features in a frame, and can be reduced further if a subset of correspondences with higher inlier ratio than the whole set is selected.

Active Search implements an ordered feature matching strategy that aims to prioritize more unique, and therefore simpler to match features. To this end, each feature is first assigned to a word in the map's visual vocabulary, and their matching order is then determined by sorting the features w.r.t. to the number of 3D point descriptors assigned to the same word. A lower number of assigned descriptors suggests that the feature is more unique and has a higher likelihood of being matched correctly. As a side effect, performing linear search is faster with fewer descriptors, even if no match can be found.

Using feature matches with different viewing directions avoids errors due to quasi-degenerate correspondence sets, and generally leads to more stable pose estimates. As we concentrate on a multi-camera setup, there is an easy way to enforce a more balanced spatial feature distribution by enforcing a balanced feature distribution over the different cameras. This obviously only holds under our current assumption that the cameras have no overlapping fields of view. We add an additional matching cost factor

$$c_I = \frac{\log m_I + 1}{\log 6} + 1. \quad (1)$$

to direct the prioritization algorithm to prefer a more balanced feature distribution. For each image I , the cost factor c_I increases with the number of matched features m_I . The term grows rapidly in the beginning, leading the prioritized search to consider other images. Once a few matches are found in all images, the prioritization scheme starts to converge to a situation where all images are (more or less) treated similarly. The feature order for each single image is independent of this term. Therefore we can maintain the order for each image separately, and simply chose the next feature based on the scaled cost of the respective image's next feature.

After selecting a feature for matching, we find the two most similar point descriptors within the same word. Here we ignore the possibility of the best match being associated with another word for the sake of simplicity and runtime efficiency. We reject ambiguous correspondences by performing a bi-directional ratio test. First, we threshold the ratio of descriptor distances between the image descriptor and the two point descriptors as in original Active Search. Afterwards we match the best point descriptor back to the image, and perform a ratio test with the obtained descriptors. We also reject the match if the generating feature is not the best match for the point. The additional ratio test with the image features is especially helpful when a pose prior is used as described in Sec. III-B, since it the image features' density in the descriptor space is not influenced by the cone filter. As Active Search, we use the found correspondences to find spatial neighbors of the 3D point that were observed together during mapping. For each point, we match the point descriptor back to the image's feature descriptors as for the ratio test before, and filter the found matches again by a ratio test. As additional constraint, we also threshold the descriptor distance to the best match to be no more than twice the distance of the generating 2D-to-3D match. While for the monocular camera case, [30] shows that performing this backmatching immediately can lead to correspondence clusters and therefore degenerate configurations, this is not a concern for our system as we explicitly enforce a sufficient feature distribution in the pose estimation.

Iterative RANSAC: For efficient and robust estimation, we utilize an iterative RANSAC strategy for the pose estimation. Instead of attempting to match all descriptors first and then perform RANSAC on the matches, we run both matching and pose estimation with RANSAC in parallel. This way we

can avoid matching more features than necessary. If very few matches are sufficient to find a good pose estimate, we can stop matching as soon as this estimate is found. To this end, we repeatedly match small batches of features to exploit Faiss' parallel search capabilities. After each batch the found matches are immediately handed to the pose estimation. While running RANSAC we maintain the best five hypotheses to avoid the need to repeatedly sample the same configurations during the iterative addition of matches. It is highly likely that a good hypothesis is found very early, but we might not be able to verify it immediately because we do not have enough matches so far. When the number of inliers increases with the number of matches, we gain confidence that the hypothesis is indeed correct. To increase the chance to find a correct hypothesis, we also guide RANSAC to prefer new and consistent data. The first match is sampled randomly from the set of recently added matches. For the following ones we reject the match and keep sampling if the point does not have at least one common observation with an already selected match. We hereby define a common observation as two points being observed in the same frame during mapping, even if the two observations were by two different cameras. Once all recently added matches were sampled multiple times, we switch to a more general method where the first match is sampled randomly from all available matches. This helps to still find a good hypothesis even if it was missed previously. When new matched points are obtained, we will first check the best hypotheses. If the new matches increase the best inlier count and -ratio above the threshold η , we will stop; otherwise we perform RANSAC as described above and update the best hypotheses if necessary. To accept a hypothesis, three requirements need to be fulfilled: The hypothesis has an inlier ratio of at least 20%, a total number of at least 15 inliers, and the inliers distribute over more than half of the available cameras. A match is counted as inlier for the hypothesis if its reprojection error is less than 10 pixels. As soon as a hypothesis is accepted, the parallel feature matching is stopped as well.

Candidate Filtering with Pose Prior: When a prior is known during localization in a large map, the number of match candidates drops significantly since we know that many points cannot be seen. Removing these impossible candidates already before matching limits the necessary effort for descriptor comparisons, and can remove ambiguities from visually similar points in other parts of the map. This improves both efficiency and robustness of the feature matching. Assuming the camera pose is known accurately, a 3D point \mathbf{p} forms an inlier match with a feature f if it is projected within r pixels around f in the image. Equivalently, \mathbf{p} has to lie in a cone along the normalized feature direction \mathbf{f} whose opening angle α is defined by r . This is illustrated in 2D in Fig. 3. Mathematically, \mathbf{p} lies within this cone if

$$\mathbf{p}^T \mathbf{f} \leq \|\mathbf{p}\| \cos \alpha, \quad (2)$$

assuming \mathbf{p} is given in the camera's coordinate system. We slightly adapt this formulation to account for the uncertainty

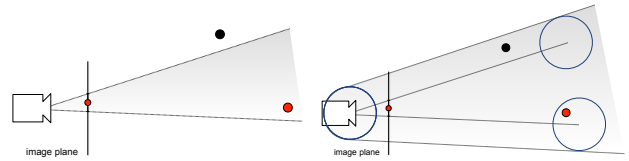


Fig. 3. Using pose priors in the localization pipeline: (left) Assuming that the pose of the camera is known precisely, a 3D point (red) is an inlier to the pose if it projects within r pixels of its corresponding image measurement (red point on image plane). This is equivalent to the point lying inside a 3D cone defined by the image measurement and the inlier threshold r . (right) Uncertainty on the camera pose can be incorporated by enlarging the cone accordingly. The figure shows the case that the position of the camera is known up to a radius (shown by the circle around the camera). The case of uncertainty in the camera orientation works similarly.

in the pose prior. For simplicity, we make 2 assumptions on this uncertainty: 1) The real camera position lies within a distance d around the prior position. This makes the set of possible positions form a sphere around the prior position. 2) The real heading direction lies within a heading cone of apex angle 2θ around the prior heading direction. This way, the heading uncertainty can trivially be incorporated by increasing the cone opening angle

$$\alpha' = \alpha + 2\theta. \quad (3)$$

The camera position uncertainty can be translated into uncertainty of the 3D point positions. We can assume a sphere of radius d around each point, and compute if this sphere intersects with the cone. A problem that arises with this approach is that it changes the density of the descriptor space by filtering spatially unreasonable points. While we can assume that the best match is not filtered out (otherwise would be a wrong match anyway), it is possible that the next best matches are removed by the spatial constraint. This impacts the result of the ratio test when performed only on the remaining descriptors. It should also be noted that this filtering approach introduces a hard limit on the uncertainty that can be handled. If the prior pose error is outside our thresholds, the correct matches cannot be found any more. For our experiments we therefore use a large but fixed uncertainty of 50 m radius around the given position, and 10° around the given heading. These values could, however, dynamically be adapted if the prior pose uncertainty can be reliably estimated.

C. Pose fusion with VIO

Although the pose estimation from our online localization algorithm is accurate, it outputs outlier pose estimations occasionally. It is due to the unreliable feature matchings for certain challenging conditions. Furthermore, the localizer can only localize the vehicle at a low frequency up to 4Hz due to its computational complexity and feature extraction runtime. In contrast, VIO is able to estimate accurate local vehicle motions at image frame rate (~ 30 Hz). Therefore, we integrate our localizer with a multi-camera VIO pipeline from [22] such that the fused pose can be used for reliable high speed autonomous driving.

A sliding window factor graph is used to fuse the estimated global poses from our localizer and the estimated local

poses from the VIO algorithm. In particular, we initialize the first vehicle global pose node of the factor graph with the estimated pose from the localizer. A new vehicle pose node will only be created whenever we have new measurements from the localizer. Since the VIO runs at a much higher rate than the visual localizer, we concatenate all VIO estimations within the time interval of two consecutive pose nodes to constrain their relative pose. Instead of using the global poses estimated from our localizer for additional constraints, we use the 2D-3D feature matchings provided by our localizer for tightly coupled fusion. The factor graph is optimized by minimizing following cost function

$$\hat{\mathbf{T}}_i, \hat{\mathbf{T}}_{i+1}, \dots, \hat{\mathbf{T}}_{i+N} = \underset{\mathbf{T}_i, \mathbf{T}_{i+1}, \dots, \mathbf{T}_{i+N}}{\operatorname{argmin}} \quad (4)$$

$$\sum_{j=i}^{i+N-1} \left\| \mathbf{T}_{j+1}^{-1} * \mathbf{T}_j * \Delta \mathbf{T}_j^{j+1} \right\|_{\Sigma_0}^2 + \quad (5)$$

$$\sum_{j=i}^{i+N-1} \sum_{k \in \Omega_j} \left\| \mathbf{u}_k - \pi(\mathbf{P}_k, \mathbf{T}_j) \right\|_{\Sigma_1}^2. \quad (6)$$

Here, N is the sliding window size of the factor graph, $\hat{\mathbf{T}}_i, \hat{\mathbf{T}}_{i+1}, \dots, \hat{\mathbf{T}}_{i+N}$ are the estimated global vehicle poses, $\Delta \mathbf{T}_j^{j+1}$ is the relative pose constraint from VIO between pose node j and $j+1$, Σ_0 is the covariance matrix for the relative pose, Ω_j is a set of 2D-3D feature matches found by localization for the j^{th} vehicle pose node, and \mathbf{u}_k and \mathbf{P}_k are the 2D and 3D feature measurements of k^{th} correspondence, respectively. Σ_1 is the covariance matrix for re-projection errors, and $\pi(*)$ is the projection function of the camera corresponding to \mathbf{u}_k . The pose graph is implemented and optimized in the Google Ceres framework [1].

The above cost function will only be optimized once the visual localizer provides new measurements. Image frame rate vehicle poses are computed by integrating relative poses estimated by VIO with respect to the latest vehicle global pose node of the factor graph.

IV. EXPERIMENTS

A. Dataset selections

We evaluate our algorithm on two datasets, *One North* and *RobotCar Seasons*. The One North dataset was recorded in Singapore's One North district as part of the AutoVision project [11]. The dataset contains the traversal of a 5 km trajectory at different times. We use 3 sets of synchronized frames to create the map and evaluate the localization offline, and a live data stream to evaluate the pose fusion with VIO. The images of the map creation show strong sunlight and hard shadows that influence the appearance of the scene. The 'sunny' traversal was recorded directly after the mapping set and therefore shows very similar conditions. The 'overcast' traversal, instead, shows significantly different illumination. The live data stream was also recorded during the 'overcast' traversal. We observed that the ground truth for the One North dataset, which is inferred from GPS and IMU is in itself inconsistent, and shows deviations compared to the map. This has an impact on the errors reported in the

quantitative analysis. The RobotCar Seasons dataset [31] is a subset of the RobotCar dataset [24] which was collected in Oxford under changing weather conditions over the course of one and a half years. The evaluated conditions include among others strong sunlight, rain, and snow. As the ground truth for this benchmark is withheld from the public, we evaluate the localization only without using a pose prior. The benchmark provides already extracted image features. For the sake of comparability we do not extract features on our own, but use the provided ones. However, it should be mentioned that the algorithms parameters were optimized with a few hundred features per image in mind, while the benchmark provides several thousand features per image.

B. Evaluation metrics

We are interested in two characteristics of the system, accuracy and runtime. To evaluate the accuracy, the RobotCar Seasons evaluation defines multiple error classes expressed as heading and position error thresholds, and reports the percentage of pose estimates with errors below the respective thresholds. We adopt this approach for our One North evaluation as well. However, due to errors in the ground truth poses as mentioned in above, we limit our position error evaluation to the ground plane and ignore potential altitude errors. For the runtime analysis, we ignore the time required for feature extraction from the images, as this is not part of our contribution, and only report the time from starting the feature matching until a pose is found or the system reports that it cannot localize, respectively.

C. Parameter settings

For the One North evaluation we use a camera setup of four wide-angle cameras, one on each side of the vehicle platform [11]. For the offline localization, we evaluate a frame after every meter driven. When evaluating the pose prior, we add random noise to the ground truth to obtain a prior as it could also be observed from a low quality GPS system. We sample a heading deviation $e_H \sim \mathcal{N}(0, 0.28)$ radians and rotate around the vertical axis. We also sample an offset distance $e_P \sim \mathcal{N}(0, 100)$ meters and shift the pose in a (uniform) random direction in the horizontal plane. For the RobotCar benchmark we use the three cameras also used in [31], located at the left, right and rear of the vehicle.

D. Experimental evaluations

Localization accuracy: Localization with the 'sunny' traversal can find a pose for more than 97% of all query frames. When employing the pose prior this number increases further, but at the same time, the amount of noise in the estimated poses increases as well.

A quantitative analysis is provided in Table I. Noticeable is the almost complete absence of poses with very low errors, *i.e.*, less than 5 degree heading / 0.5 m position deviation. We suspect this to be caused by misalignments of the map and ground truth. For the 'overcast' traversal, the localization accuracy significantly decreases, and in some sections of the trajectory localization fails consistently. Using the pose prior again increases the percentage of localized poses, but also

TABLE I

PERCENTAGES OF POSES WITHIN DIFFERENT ERROR THRESHOLDS.

deg / m	2 / 0.25	5 / 0.5	Errors 10 / 5	15 / 10	20 / 20
sunny	0.1	1.5	96.8	97.6	97.9
sunny+prior	0.2	1.6	95.0	97.9	98.8
overcast	0.0	0.0	39.6	51.9	60.7
overcast+prior	0.0	0.5	43.0	62.6	75.2

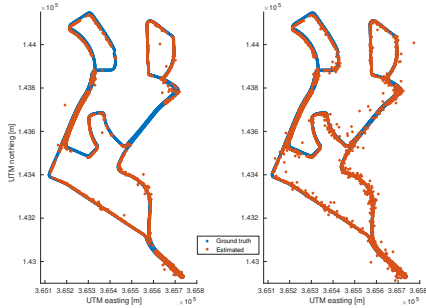


Fig. 4. Estimated vs. ground truth poses for the ‘overcast’ trajectory without (left) and with (right) a pose prior. noise as shown in Fig. 4. Notice that the failure to localize every frame as well as inaccurate poses can be handled by integrating localization into a VIO system, as shown below.

For the RobotCar Seasons benchmark, we compare our method to both the original monocular Active Search algorithm, as well as a version that performs Active Search on each image individually and then performs multi-camera pose estimation (ActiveSearch+GC). For low error classes, our approach performs significantly worse than both Active Search references, but the difference shrinks with higher error thresholds. The percentages of poses in the different error classes are shown in Table II.

Localization runtime: Due to the dynamic number of matches, our algorithm’s runtimes are highly dependent on the recognizability of the scene. This can be seen in Table III.

The ‘sunny’ traversal can easily be located within our map with very few, good feature matches, which leads to very low runtimes of 55ms and 48ms on average, respectively. For the ‘overcast’ images, more or even all image features need to be matched, resulting in significantly higher runtimes. In this case, using the candidate filtering also leads to significantly lower runtimes, whereas its impact for ‘sunny’ traversal, where very few features need to be matched anyway, is mostly negligible.

Analyzing the runtimes for the RobotCar Seasons experiments shows that our approach easily outperforms the Active

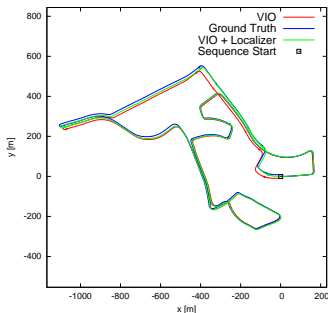


Fig. 5. Trajectory estimated by pure VIO and fusion of VIO and localizer from the ‘overcast’ dataset vs. ground truth. The fused pose actually reaches the starting point, whereas there is an offset in the GPS ground truth.

TABLE II

SUMMARIZED ERRORS FOR ALL DAY AND NIGHT SCENES WITH DIFFERENT CAMERA SETUPS.

	m	all day .25 / .50 / 5.0 2 / 5 / 10	all night .25 / .50 / 5.0 2 / 5 / 10
ActiveSearch	deg	35.6 / 67.9 / 90.4	0.9 / 2.1 / 4.3
ActiveSearch + GC		45.5 / 77.0 / 94.7	2.7 / 6.9 / 12.1
Ours		15.5 / 40.5 / 85.8	0.3 / 1.9 / 8.2

TABLE III

LOCALIZATION RUNTIMES FOR THE ONE NORTH EXPERIMENTS WITHOUT FEATURE EXTRACTION IN MS.

	All frames		Localized frames	
	mean	median	mean	median
‘sunny’	55	43	52	43
‘sunny+prior’	48	40	46	39
‘overcast’	146	151	103	93
‘overcast+prior’	106	78	77	65

Search multi-camera implementation and is only slightly slower than the monocular camera version. Active Search and multi-camera Active Search takes 291ms and 879ms on average, while our solution takes 371ms. However, the average runtimes for our system are significantly influenced by the frames that could not be localized, as our system tries to match all available image features in this case. This could easily be avoided by setting a (large) upper limit on the number of matched features. If we consider localized frames for our solution, our system is with 239ms on average almost 20% faster than the monocular Active Search implementation.

Pose fusion evaluation: Fusing the localizer’s absolute pose estimates with relative pose estimates from VIO enables us to provide the current pose with the frequency and smoothness required for autonomous driving. Fig. 5 shows the positions estimated by pure VIO on the ‘overcast’ traversal of the OneNorth dataset, which experiences drift over the course of the trajectory, together with the pose fusion result. Here, the drift can successfully be corrected using localization against the map. Notice also that integrating localization results into the VIO pipeline allows us to accurately track the pose over the full sequences, even though localization fails or is inaccurate in some parts (c.f. Fig. 4).

V. CONCLUSION

In this paper, we have proposed a novel visual localization system tailored to multi-camera setups. Our approach does not require preset parameters for the number of features to match, and is therefore highly adaptive to different scenes while minimizing the feature matching effort. We also presented an efficient candidate filtering step based on pose priors. By fusing our pose estimates with relative motion estimates from a VIO system we are able to provide accurate absolute poses in real time. Our results thus clearly demonstrate the feasibility of real-time, camera only localization for autonomous driving, even in larger scenes.

In the future, we plan to optimize parameters and our implementation to leverage the full potential of the approach. Adjusting the pose prior to the uncertainty of the VIO pipeline, rather than using a fixed large uncertainty radius, seems especially promising.

REFERENCES

- [1] S. Agarwal, K. Mierle, and Others. Ceres solver. <http://ceres-solver.org>.
- [2] C. Arth, M. Klopschitz, G. Reitmayr, and D. Schmalstieg. Real-Time Self-Localization from Panoramic Images on Mobile Devices. In *ISMAR*, 2011.
- [3] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide area localization on mobile phones. In *ISMAR*, 2009.
- [4] E. Brachmann and C. Rother. Learning Less is More - 6D Camera Localization via 3D Surface Regression. In *CVPR*, 2018.
- [5] F. Camposeco, T. Sattler, A. Cohen, A. Geiger, and M. Pollefeys. Toroidal Constraints for Two-Point Localization under High Outlier Ratios. In *CVPR*, 2017.
- [6] F. Camposeco, T. Sattler, and M. Pollefeys. Minimal Solvers for Generalized Pose and Scale Estimation from Two Rays and One Point. In *ECCV*, 2016.
- [7] T. Cavallari, S. Golodetz, N. A. Lord, J. Valentin, L. Di Stefano, and P. H. S. Torr. On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation. In *CVPR*, 2017.
- [8] M. Fischler and R. Bolles. Random Sampling Consensus: A Paradigm for Model Fitting with Application to Image Analysis and Automated Cartography. *Commun. ACM*, 24:381–395, 1981.
- [9] C. Häne, L. Heng, G. H. Lee, F. Fraundorfer, P. Furgale, T. Sattler, and M. Pollefeys. 3D visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *IMAVIS*, 68:14 – 27, 2017.
- [10] R. Haralick, C.-N. Lee, K. Ottenberg, and M. Nölle. Review and analysis of solutions of the three point perspective pose estimation problem. *IJCV*, 13(3):331–356, 1994.
- [11] L. Heng, B. Choi, Z. Cui, M. Geppert, S. Hu, B. Kuan, P. Liu, R. Nguyen, Y. C. Yeo, A. Geiger, G. H. Lee, M. Pollefeys, and T. Sattler. Project AutoVision: Localization and 3D Scene Perception for an Autonomous Vehicle with a Multi-Camera System. *arXiv:1809.05477*, 2018.
- [12] L. Heng, P. Furgale, and M. Pollefeys. Leveraging Image-based Localization for Infrastructure-based Calibration of a Multi-camera Rig. *Journal of Field Robotics*, 32(5):775–802, 2015.
- [13] L. Heng, G. H. Lee, and M. Pollefeys. Self-calibration and visual slam with a multi-camera system on a micro aerial vehicle. *Autonomous Robots*, 39(3):259–277, 2015.
- [14] A. Kendall and R. Cipolla. Geometric loss functions for camera pose regression with deep learning. In *CVPR*, 2017.
- [15] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *ISMAR*, 2007.
- [16] Z. Kukelova, M. Bujnak, and T. Pajdla. Real-Time Solution to the Absolute Pose Problem with Unknown Radial Distortion and Focal Length. In *ICCV*, 2013.
- [17] G. H. Lee, B. Li, M. Pollefeys, and F. Fraundorfer. Minimal solutions for the multi-camera pose estimation problem. *IJRR*, 34(7):837–848, 2015.
- [18] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide Pose Estimation Using 3D Point Clouds. In *ECCV*, 2012.
- [19] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *ECCV*, 2010.
- [20] L. Liu, H. Li, and Y. Dai. Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map. In *ICCV*, 2017.
- [21] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys. Towards Robust Visual Odometry with a Multi-Camera System. In *IROS*, 2018.
- [22] P. Liu, M. Geppert, L. Heng, T. Sattler, A. Geiger, and M. Pollefeys. Towards robust visual odometry with a multi-camera system. In *IROS*, 2018.
- [23] S. Lynen, T. Sattler, M. Bosse, J. Hesch, M. Pollefeys, and R. Siegwart. Get Out of My Lab: Large-scale, Real-Time Visual-Inertial Localization. In *RSS*, 2015.
- [24] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.
- [25] D. Massiceti, A. Krull, E. Brachmann, C. Rother, and P. H. Torr. Random Forests versus Neural Networks - What's Best for Camera Relocalization? In *ICRA*, 2017.
- [26] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt. Scalable 6-DOF Localization on Mobile Devices. In *ECCV*, 2014.
- [27] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. *TRO*, 31(5):1147–1163, 2015.
- [28] R. Mur-Artal and J. D. Tardós. ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. *TRO*, 33(5):1255–1262, 2017.
- [29] R. Pless. Using Many Cameras as One. In *CVPR*, 2003.
- [30] T. Sattler, B. Leibe, and L. Kobbelt. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *PAMI*, 39(9):1744–1756, 2017.
- [31] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla. Benchmarking 6DOF Outdoor Visual Localization in Changing Conditions. In *CVPR*, 2018.
- [32] T. Sattler, A. Torii, J. Sivic, M. Pollefeys, H. Taira, M. Okutomi, and T. Pajdla. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization? In *CVPR*, 2017.
- [33] J. L. Schönberger and J.-M. Frahm. Structure-From-Motion Revisited. In *CVPR*, June 2016.
- [34] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [35] J. L. Schönberger, M. Pollefeys, A. Geiger, and T. Sattler. Semantic Visual Localization. In *CVPR*, 2018.
- [36] U. Schwesinger, M. Buerki, J. Timmer, S. Rottman, L. Wolf, L. M. Paz, H. Grimmet, I. Posner, P. Newman, C. Häne, L. Heng, G. H. Lee, T. Sattler, M. Pollefeys, M. Allodi, F. Valenti, K. Mimura, B. Goebelsmann, and R. Siegwart. Automated Valet Parking and Charging for e-Mobility: Results of the V-Charge Project. In *IV*, 2016.
- [37] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images. In *CVPR*, 2013.
- [38] L. Svärm, O. Enqvist, F. Kahl, and M. Oskarsson. City-Scale Localization for Cameras with Known Vertical Direction. *PAMI*, 39(7):1455–1461, 2017.
- [39] C. Sweeney, V. Fragoso, T. Höllerer, and M. Turk. Large Scale SfM with the Distributed Camera Model. In *3DV*, 2016.
- [40] A. Valada, N. Radwan, and W. Burgard. Deep Auxiliary Learning For Visual Localization And Odometry. In *ICRA*, 2018.
- [41] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg. A Minimal Solution to the Generalized Pose-and-Scale Problem. In *CVPR*, 2014.
- [42] C. T. Viktor Larsson, Johan Fredriksson and F. Kahl. Outlier Rejection for Absolute Pose Estimation with Known Orientation. In *BMVC*, 2016.
- [43] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-Based Localization Using LSTMs for Structured Feature Correlation. In *ICCV*, 2017.
- [44] B. Zeisl, T. Sattler, and M. Pollefeys. Camera pose voting for large-scale image-based localization. In *ICCV*, 2015.