State of the Art in Visual Localization for UAVs in GNSS-denied environments

Michael Schleiss
Fraunhofer FKIE
Fraunhoferstr. 20, 53343 Wachtberg
GERMANY

michael.schleiss@fkie.fraunhofer.de

ABSTRACT

The expected proliferation of autonomous aerial vehicles in the near future requires back-up concepts in case of GNSS-failure or attacks in order to ensure robust and safe autonomous aviation. Visual localization has made large progress as an alternative to GNSS-based navigation, especially in the realm of self-driving cars. Can this progress be transferred to the aerial scenario? We present a large-scale outdoor dataset that can help researchers answer this question and measure the state-of-the-art in aerial visual localization.

1 INTRODUCTION

One of the great challenges of autonomy in unmanned vehicles - be it on the ground or in the air - is the precise estimation of the vehicle's own location in absence of external localization sources like GPS or other satellite-based navigation systems (GNSS). Visual localization - an alternative to GNSS using only camera images and supporting sensors like an IMU or an altimeter - has made substantial progress sparked by the enormous interest in autonomous driving in the very recent past. It is unclear, however, whether this progress can be directly translated from ground-based systems such as self-driving cars to aerial systems such as UAVs.

We have developed a benchmark that measures the performance of current camera-based navigation methods under challenging, real flight conditions. This benchmark is based on a large-scale dataset consisting of 1-hour of flight data, spanning a trajectory over 130 km long, collected over an area near the city of Bonn, Germany at an altitude of 300-600 m over ground.

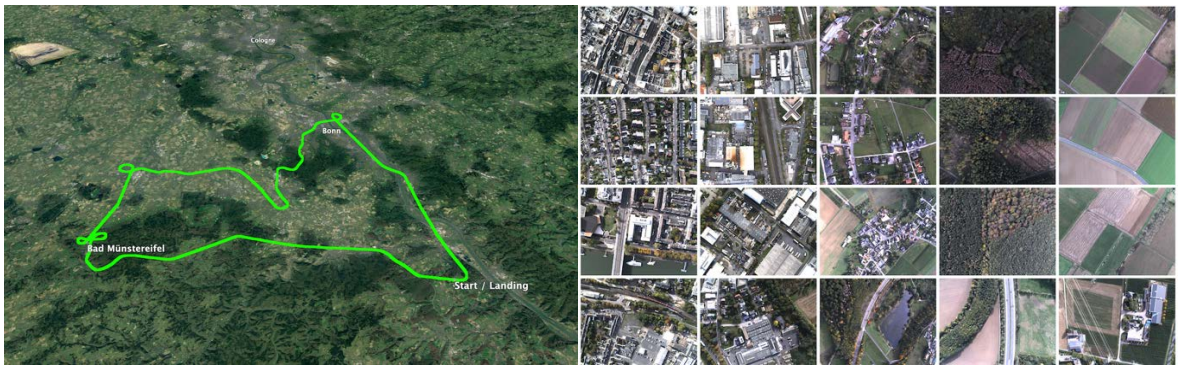

Figure 1: Our dataset is 130km long and spans a trajectory covering various land surface types depicted on the right (columns are sorted f.l.t.r: urban, industrial, rural, forestial and agricultural).

Based on this benchmark we investigate how well visual localization performs in an aerial scenario using the widely adopted ORB-SLAM3 [1] and NetVLAD [2] as examples for state-of-the-art visual localization methods from the domain of autonomous vehicle research. We find that these methods suffer from the change in perspective in aerial situations due to lack of a dominant (upright) orientation especially for downward looking cameras. We report reasonable accuracy close to satellite-based navigation systems over populated areas based on our benchmark. Purely natural, unstructured scenes, i.e. forests, however, still pose a challenge for visual geo-localization.

Our research has applications for autonomous flight in Anti-Access/Area Denial (A2/AD) environment, spoofing detection and increasing robustness in urban environments. It can also be used for cooperative localization in air/ground scenarios. With our research we hope to inform future research steps that are required for establishing camera-based navigation as an alternative to GNSS. The dataset is available publicly and shall serve as a benchmark for future progress in aerial visual localization.

2 RELATED WORK

Visual localization has a long tradition as an alternative to satellite-based navigation systems. It has been successfully incorporated into Tomahawk cruise missiles [3] and on Mars rovers [4, 5] in the 1980s and 1990s for navigation in absence of GNSS. By today visual localization has found many applications in areas such as augmented reality [6] and robot navigation in the form of drones [7], autonomous vessels [8] and self-driving cars [9]. Visual localization is used indoors where satellite navigation is not available as well as outdoors where reception might be degraded due to obstacles, multipathing or it might even be interrupted due to natural causes or malicious acts such as spoofing and jamming. However, depending on the use case and accuracy requirements, the methodologies, sensor and hardware requirements differ vastly. Visual localization techniques can be found in high cost platforms such as the aforementioned cruise missiles, Mars rovers and self-driving cars as well as in low-cost, amateur drones, home appliances and toys. While on the latter type of platforms a local motion estimate might be sufficient the former ones typically require drift-free geo-coordinates on a more global scale.

More formally we distinguish two types of navigation. Relative visual localization refers to motion estimation relative to a local coordinate system where the origin represents an arbitrary starting point with no reference to global geo-coordinates. Typical implementations observe image sequences and try to estimate the motion in between frames. This is also called visual odometry [10]. Relative visual localization or visual odometry is akin to dead reckoning. On the other hand, there is absolute visual localization (AVL) which refers to position estimation in a global coordinate system. In absolute visual localization one typically tries to find known landmarks and compare it to some sort of map with known geocoordinates to deduce the camera's position in world coordinates [12-14]. It therefore provides – in contrast to RVL - a drift-free, global position estimate.

Both types of navigation – RVL and AVL – have their pros and cons. In general, visual localization requires scenes that are predominantly static and sufficiently illuminated. RVL relies only on a sequence of onboard images for motion estimation. As long as the above requirements are met visual odometry is usually able to provide motion estimates. However, due to the incremental nature, it accumulates drift over time – although this drift can be reduced using loop closures. Loop closures are a technique where previously traversed places are detected and then used to optimize the trajectory [11]. This might be enough for a small-scale, local scenario such as indoor navigation where loop closures are likely to happen frequently. In large-scale outdoor scenarios loop closures might never happen. Therefore, we need a different mechanism for eliminating drift. AVL can provide such drift free position estimate, however, it requires additional information because onboard imagery is compared to an offline database, e.g. georeferenced orthophotos, digital elevation models, 3D-point clouds, etc.

Both types of navigation – RVL and AVL – can be fused using filtering techniques, e.g. an Extended Kalman Filter. This resembles the traditional navigation pipeline where an inertial measurement unit (IMU) or an inertial navigation system (INS) provides high-rate relative motion estimates and a satellite-based navigation system provides drift-free, global position estimates which are then fused to a six degrees-of-freedom (6-DoF) global pose estimate.

Instead of replacing both INS and GNSS with their visual counterparts, hybrid systems are also feasible. One can combine visual odometry and an INS into visual-inertial-odometry (VIO) [15] which can reduce drift compared to a vision-only or an INS-only solution and recover metric scale in a monocular camera setting. Another possibility is to augment GNSS with visual localization techniques to increase the accuracy of the pose estimate in scenarios where up to centimeter accuracy is required [16].

Some studies on the accuracy of RVL have been conducted. Delmerico and Scaramuzza [17] presented benchmark comparisons of the state of-the-art VIO algorithms on several hardware platforms (Laptop, Intel NUC, UP Board, and ODROID) using a drone-based indoor dataset. They demonstrated that VIO-algorithms can provide highly accurate position estimates with less than 1% drift per distance travelled in real-time even on low-compute platforms. Schubert et al. provide a VIO-Bechmark that includes various indoor and outdoor sequences from a handheld camera [18]. They confirm the high accuracy of the benchmarked VIO-methods; however, they note that they can become unstable in large-scale trajectories.

In contrast to RVL - to the best of our knowledge - no benchmarks have been performed on AVL in a large-scale aerial setting. Although many such benchmarks exist for the domain of autonomous driving and other ground or close to ground settings e.g. [9, 7] no public dataset exists for performing comparisons. While a lot of progress has been made in the former domains it is unclear if the methods can be directly translated to the aerial settings. The goal of this work is to provide such a dataset and perform evaluations on some baseline methods.

Our work has the following contributions. First, we provide a publicly available dataset for the comparison of visual localization techniques in a large-scale aerial setting. Second, we perform an evaluation based on this benchmark and state of the art RVL and AVL techniques, namely ORB-SLAM3 and NetVLAD.

3 DATASET

The data was collected onboard a lightweight 2-seat fixed-wing aircraft with a nominal cruise speed of around 13a0 km/h. The payload consisted of the following sensors:

- Allied Vision Prosilica 1660C Mono Color Camera, 1600 × 1200 × 3, 25Hz, 2/3” ON Semi KAI-02050 CCD, global shutter, 8mm Schneider-Kreuznach lens, 57◦ HFoV
- SBG Systems Ellipse-D inertial and GNSS navigation system (INS/GNSS), 6-axis IMU, 200 Hz, dual antenna, resolution: <1 m (SBAS) / 0.05°

The INS serves a double function in our case. It provides raw IMU measurements from an inertial measurement unit (IMU) as well as reference poses in 6-DoF that are obtained by fusing IMU and GPS. The INS is able to compute highly accurate positions with the help of satellite-based augmentation systems (SBAS). It reduces the uncertainty of the computed position to less than one meter. All items of the payload were placed close to each other below the wing with the camera facing down except for the GNSS antennas which were mounted on the top of the wing. We measured the distances between all parts of the payload carefully and provide a schematic in Figure 2.

The sensors were logged using a PC with an eight-core Intel Xeon E-2286M Processor, 32 GB DDR4-Ram and two 2 TB SSDs in RAID-0 configuration running Ubuntu Linux 18.04 and ROS Melodic. Camera and

IMU/INS are synchronized through hardware triggering ensuring accurate timestamps. The camera's timestamps correspond to the beginning of exposure. The exposure time was fixed to 5 ms at the day of capture. We used the official ROS implementations of the sensor drivers and adapted them for timestamp synchronization. The camera intrinsics and the extrinsics between IMU and camera were obtained by using the Kalibr calibration toolbox [19].

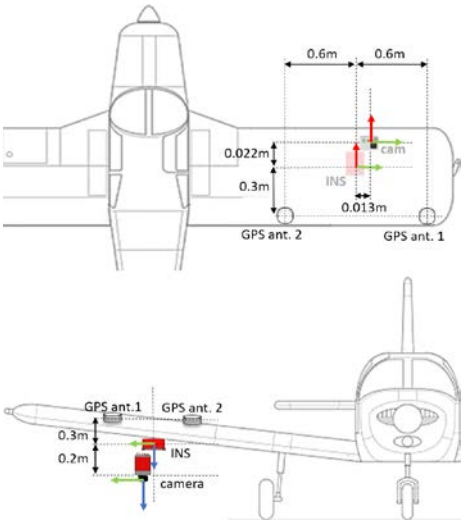

Figure 2: Diagram of the light aircraft (top and front view) and the sensors locations. The distances between the GPS Antennas and INS have been directly measured while the distances between the sensors are obtained through extrinsic calibrations. The coordinate frames show the origin and direction of each sensor mounted to the vehicle with the convention: x-forward (red), y-right (green), z-down (blue).

The dataset contains time-synchronized high-resolution aerial images, GPS and IMU data. It consists of an hour- long recording that spans a trajectory of 138 kilometers and was recorded on October 13, 2020. We aimed to collect a diverse dataset covering different ground surface types. It includes sequences over urban and rural areas as well as areas that are shaped by agriculture, forests or vegetation. See Fig. 5 for examples. The total size of the dataset amounts to roughly 320 GB. The dataset and a detailed description are publicly available and can be accessed at aervisloc.github.io.

4 EXPERIMENTS

We apply two state of the art visual localization techniques to emphasize the utility of our benchmark dataset to autonomous aerial navigation. We combine monocular visual odometry (ORB-SLAM3) with a global place recognition technique named NetVLAD. NetVLAD is a neural network that is trained on thousands of street level images from the city of Pittsburgh. It is based on the idea of image similarity. Images that are taken at the same location but at different points in time should be more similar than images taken at different locations taken at the same time. Based on image similarity we now can compare onboard images to a database of geotagged reference images. This database consists of 3400 orthophotos placed equally spaced alongside the planned trajectory depicted in Figure 1. We use the geotag from the top-1 result as a rough position estimate. Although NetVLAD has been trained on an autonomous driving scenario it has proven its utility in various domains [20] and under difficult circumstances such as seasonal and weather changes [21].

We first conduct experiments using both systems in isolation. ORB-SLAM3 will be evaluated based on average drift per distance travelled. NetVLAD will be evaluated based on the percentage of correctly

retrieved images. An image is successfully retrieved if the top-1 result is within 100m of the vehicles position. We further evaluate place recognition on categories such as urban, forestial and agricultural based on a land cover database [22]. Finally, we evaluate a combination of these two RVL/AVL techniques by fusing the position estimates in an EKF. Our state vector consists of position, rotation and velocity. Velocity is updated based on visual odometry, pose based on an INS using magnetometer and accelerometer and position both based on the current velocity estimate and position estimate from NetVLAD.

5 EVALUATION

Figure 3 shows the difference between the visual odometry only and the fused RVL/AVL result. While visual odometry accumulates hundreds of meters drift the fused results is very close to the ground truth trajectory. Specifically, we can reduce drift from 1.29% of distance travelled to a few meters over the total trajectory of the evaluated subsequence. In the following we report performance when NetVLAD is used in isolation in regard to correctly retrieved database images. NetVLAD is able to retrieve an image in the database that is within 100m of the vehicles position in 69% of the cases. However, the accuracy varies greatly based on what type of land surfaces are encountered. The accuracy drops to only 14% over forestial and 45% over agricultural areas in this subsequence. Also, performance drops significantly if the orientation of the vehicle (yaw angle) is not known beforehand from 69% to 34%.

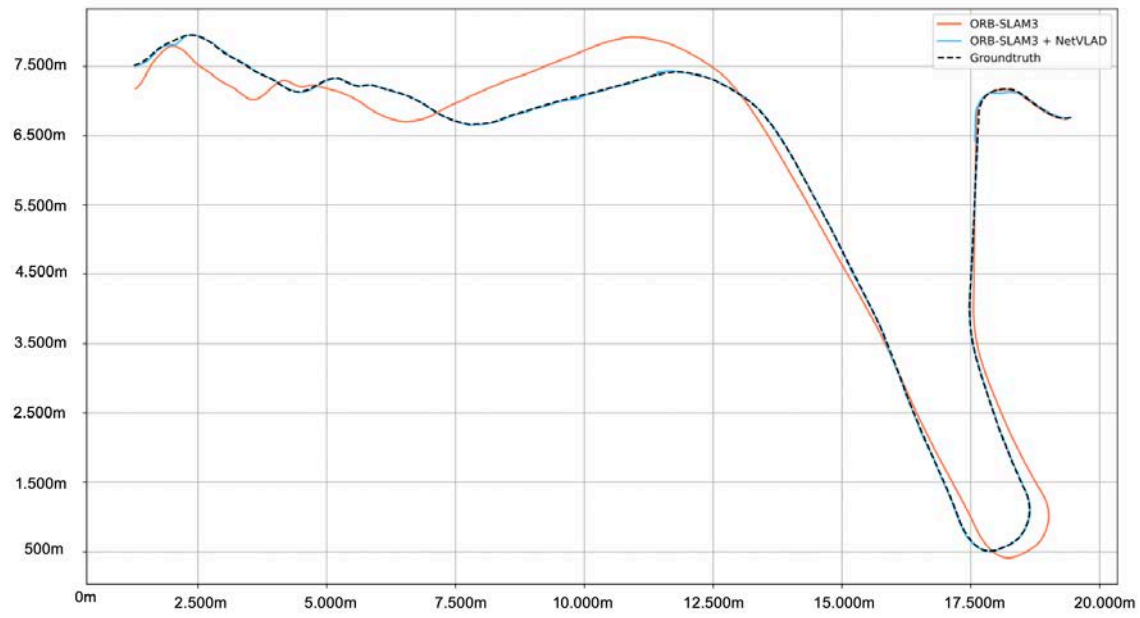

Figure 3: Visualization of the localization performance of two state of the art visual localization algorithms over one of our test sequences. It can be seen from the trajectories above that ORB-SLAM3, a state-of-the-art visual odometry technique, suffers from drift over time (the plane traversed from east to west in this sequence). The combination of ORB-SLAM3 and a geolocalization technique such as NetVLAD results in a trajectory that aligns very well with the ground truth positions.

6 CONCLUSIONS

We have discussed the difference between relative and absolute visual localization. Based on a large-scale benchmark dataset we have evaluated a state-of-the-art representative for both types of navigation. The

results show that these methods are able to provide accurate localization in absence of GNSS in a large-scale outdoor scenario. However, the results also show that under challenging circumstances the robustness of the evaluated AVL method drops significantly. Further research has to be made into topics such as rotation-invariant place recognition and place recognition under drastic appearance changes. The dataset is made publicly available and can be used to measure the future progress in large-scale aerial visual localization.

7 REFERENCES

[1] C. Campos et al.: ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM, IEEE Transactions on Robotics. 2021

[2] A. Relja et al.: NetVLAD: CNN architecture for weakly supervised place recognition, Proceedings of the IEEE conference on computer vision and pattern recognition. 2016

[3] Carr, J. R. & Sobek, J. S. Digital Scene Matching Area Correlator (DSMAC). Image Process Missile Guid 36–41 (1980)

[4] Laubach, L. M. and C. F. O. and G. T. and S. Visual Localization Methods for Mars Rovers using Lander, Rover, and Descent Imagery. In Proceedings of the 4th International Symposium on Artificial Intelligence, Robotics and Automation in Space (1997)

[5] Maimone, M., Cheng, Y. & Matthies, L. Two years of Visual Odometry on the Mars Exploration Rovers. J Field Robot 24, 169–186 (2007)

[6] Klein, G., & Murray, D. Parallel tracking and mapping on a camera phone. In 2009 8th IEEE International Symposium on Mixed and Augmented Reality (pp. 83-86). IEEE. (2009)

[7] Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., ... & Siegwart, R. The EuRoC micro aerial vehicle datasets. The International Journal of Robotics Research, 35(10), 1157-1163. (2016)

[8] Griffith, S., Chahine, G., & Pradalier, C. Symphony lake dataset. The International Journal of Robotics Research (2017)

[9] Maddern, W., Pascoe, G., Linegar, C., & Newman, P. 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research, 36(1), 3-15 (2017).

[10] Scaramuzza, D., & Fraundorfer, F. Visual odometry [tutorial]. IEEE robotics & automation magazine, 18(4), 80-92. (2011)

[11] Bokovoy, A., & Yakovlev, K. Original loop-closure detection algorithm for monocular vslam. In International Conference on Analysis of Images, Social Networks and Texts (pp. 210-220). Springer, Cham. (2017)

[12] Conte, G., & Doherty, P. (2009). Vision-based unmanned aerial vehicle navigation using geo-referenced information. EURASIP Journal on Advances in Signal Processing, 2009, 1-18.

[13] Schleiss, M. Image-based geolocalization for UAVs. In Forum Bildverarbeitung 2020 (p. 401). KIT Scientific Publishing. (2020).

[14] Couturier, A., & Akhloufi, M. A. A review on absolute visual localization for UAV. Robotics and Autonomous Systems, 135, 103666. (2021)

[15] Leutenegger, S., Lynen, S., Bosse, M., Siegwart, R., & Furgale, P. Keyframe-based visual–inertial odometry using nonlinear optimization. The International Journal of Robotics Research, 34(3), 314-334. (2015)

[16] Surber, J., Teixeira, L., & Chli, M. Robust visual-inertial localization with weak GPS priors for

repetitive UAV flights. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 6300-6306). IEEE. (2017)

[17] Delmerico, J., & Scaramuzza, D. A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2502-2509). IEEE. (2018).

[18] Schubert, D., Goll, T., Demmel, N., Usenko, V., Stückler, J., & Cremers, D. The TUM VI benchmark for evaluating visual-inertial odometry. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 1680-1687). IEEE. Chicago. (2018)

[19] Rehder, J., Nikolic, J., Schneider, T., Hinzmann, T., & Siegwart, R. Extending kalibr: Calibrating the extrinsics of multiple IMUs and of individual axes. In 2016 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4304-4311). IEEE. (2016).

[20] Hausler, S., Garg, S., Xu, M., Milford, M., & Fischer, T. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14141-14152). (2021)

[21] Warburg, F., Hauberg, S., Lopez-Antequera, M., Gargallo, P., Kuang, Y., & Civera, J. Mapillary street-level sequences: A dataset for lifelong place recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2626-2635). (2020)

[22] Büttner, G., Feranec, J., Jaffrain, G., Mari, L., Maucha, G., & Soukup, T. The CORINE land cover 2000 project. EARSeL eProceedings, 3(3), 331-346. (2004)