Technical report from Automatic Control at Linköpings universitet

State-of-the-art Report of Research about Multi Sensor Image-based Navigation

Jeongmin Kang, Zoran Sjanic, Gustaf Hendeby

Division of Automatic Control

E-mail: jeongmin.kang@isy.liu.se, zoran.sjanic@isy.liu.se, gustaf.hendeby@isy.liu.se

13th April 2023

Report no.: LiTH-ISY-R-3109

Address:
Department of Electrical Engineering
Linköpings universitet
SE-581 83 Linköping, Sweden

WWW: http://www.control.isy.liu.se

Abstract

This report aims to describe the latest research and method development of image-based multi sensor fusion navigation and summarizes open aerial datasets which can support the latest research related to this project. It supports the initial setting of the direction of the algorithm development in the early stage of the project.

The Multi Sensor Image-based Navigation project aims to study and develop the methods focusing on image-based multisensor navigation in order to acquire a precise localization of the aircraft. GNSS-based localization and navigation systems are sensitive to disturbances and jamming, hence the capability to provide reliable position accuracy without GNSS is a key element to develop the navigation systems.

The output of this project can be utilized in a wide range of applications, such as aircraft operation in GNSS denied environments or urban air mobility context.

Keywords: sensor fusion, aerial navigation, image processing, simultaneous localization and mapping, visual-inertial navigation, deep learning

State-of-the-art Report of Research about Multi Sensor Image-based Navigation

Jeongmin Kang, Zoran Sjanic, and Gustaf Hendeby
Dept. of Electrical Engineering, Linköping University, Linköping, Sweden

Summary

This report aims to describe the latest research and method development of image-based multi sensor fusion navigation and summarizes open aerial datasets which can support the latest research related to this project. It supports the initial setting of the direction of the algorithm development in the early stage of the project.

Contents

1 Introduction

1.1 Project Information

The Multi Sensor Image-based Navigation project aims to study and develop the methods focusing on image-based multisensor navigation in order to acquire a precise localization of the aircraft. GNSS-based localization and navigation systems are sensitive to disturbances and jamming, hence the capability to provide reliable position accuracy without GNSS is a key element to develop the navigation systems.

The output of this project can be utilized in a wide range of applications, such as aircraft operation in GNSS denied environments or urban air mobility context.

1.2 Objective of the Report

This report aims to describe the latest research and method development of image-based multi sensor fusion navigation, and especially reviews landmark papers related to the project. In addition, the latest studies in which deep learning methods are also described to investigate the latest research trends.

Camera sensors used for image-based navigation include monocular, stereo, fisheye, and RGB-D, etc., cameras. This report focuses on monocular camera-based navigation studies from the viewpoint that the research using the monocular camera sensor precedes the development of core technology. Recently, visual-inertial navigation in which measurements from an inertial measurement unit (IMU) and a camera are fused shows high performance and is receiving a lot of attention. This report also summarizes the landmark papers utilizing IMU sensors. In addition, open aerial datasets which can support the latest research related to this project are summarized. It also sets the initial direction of the research, and the dataset and methods can be expected to be referenced and utilized in developing algorithms, within the project.

The rest of this report is organized as follows. First, the background of image-based navigation is reviewed in Chapter 2, and open datasets are summarized in Chapter 3. The key papers of the image-based navigation are divided into feature-based, direct, visual-inertial, and deep learning applied methods, and the methods are reviewed in Chapter 4 to 7. Finally, Chapter 8 concludes and discusses future research directions.

2 Background

The latest image-based navigation research showing high performance has been initially led by the visual simultaneous localization and mapping (SLAM) based on image processing with photogrammetry. The SLAM methods recognize the surrounding environment to build a map and estimates the state of a sensor simultaneously. The classical age saw the introduction the probabilistic estimation formulations for SLAM [1, 2], including approaches based on Kalman filter, extended Kalman filter, and particle filter. The main formulations connected to efficiency and data association [3, 4] are well described. The front end of the SLAM is related to research fields such as computer vision and signal processing. The back end consists of mix of the geometry, optimization, and probabilistic estimation [5]. A method for estimating the trajectory based on image information was introduced as visual odometry (VO) [6]. Building a globally consistent representation using the generated environment map information is the difference between SLAM and odometry [7, 8]. The VO and visual SLAM are interdependent methods, and the definition may vary depending on the researcher. The VO is regarded as the localization part of SLAM in this report.

The initial image-based SLAM aimed to estimate camera pose and generate a 3D map is known as structure from motion in computer vision. These methods use image processing techniques to extract feature points to match between different image frames and then estimate camera pose based on photogrammetry. The basic information that is used is a camera model, calibration, and 2D-2D/3D-3D/3D-2D motion from image correspondences. The final pose of the camera and structure are refined by optimization methods such as bundle adjustment and pose graph. However, the disadvantage is that the motion is only recovered up to an unknown scale factor depending on depth ambiguity. For this reason, navigation systems using only a camera sensor still have a limited accuracy that is lower than that of positioning using GNSS fusion or Light detection and ranging (LiDAR) sensors. The scale can be determined from direct measurement by other sensors. Here, as one of the alternatives, the scale factor can be determined by using the IMU. The IMU fusion methods have been proposed as loosely or tightly coupled systems. In loosely coupled systems, the poses are estimated by an independent algorithm, then the vision and inertial measurements are fused in a subsequent estimation. On the other hand, tightly coupled systems fuse the correlations among all the measurements in a single algorithm. Additionally, studies on estimating camera pose and generating dense map information using deep learning methods have been actively conducted. Methods using deep learning have been extensively studied, from being a method to supplement parts of the previous SLAM structure to end-to-end methods that learn the output from the input.

This report summarizes the landmark papers of the latest studies from the background of these technical development and examines the theoretical or practical ways to be able to utilize them in this project from the key methods.

3 Datasets

It is challenging to acquire high quality data for development of the navigation algorithm in both indoor and outdoor environments. Therefore, utilizing open datasets commonly used by researchers can be a method in terms of availability of the data. Ground-level datasets such as Kitti [9], nuScenes [10], Waymo [11], and Argoverse [12] have significant volume and quality in terms of the type of sensor combination, driving scenarios, and data utilization. These datasets can be utilized for various tasks in autonomous driving systems such as tracking, prediction, LiDAR segmentation, panoptic segmentation, planning, scenario-based driving, etc. However, aerial datasets lack the volume compared to the ground-level datasets due to sensor configuration and limitations in data acquisition step. Since it is difficult to change the sensor configuration compared to ground vehicles, this makes the limitations of the data collection for various scenarios. This chapter summarizes the characteristics of currently accessible aerial datasets, with the expectation that the datasets can support the algorithm development in terms of flexibility and utility in the early stage of the project implementation. The overall features of the datasets are listed in Table 1

Table 1. The Characteristics of the Open Dataset

	EuRoC MAV (2016) [13]	**UZH-FPV Drone Racing (2019) [14]**	**Blackbird UAV (2020) [15]**	**NTU VIRAL (2021) [16]**	**KAIST VIO (2021) [17]**
Environment	Indoor	Indoor/Outdoor	Indoor	Indoor/Outdoor	Indoor
Sequences	11	27	186	9	4
Camera	20 Hz, 752*480	30/50 Hz, 640*480	120/60 Hz, 1024*768	10 Hz, 752*480	30 Hz, 640*480
IMU	200 Hz	500/1000 Hz	100 Hz	385 Hz	100 Hz
Ground truth	20 Hz	20 Hz	360 Hz	20 Hz	50 Hz
Others	-	Event Camera 50Hz	Depth camera 60Hz, Segmentation* 60 Hz	Vertical/Horizontal LiDAR, UWB sensors	RGB (640*480) 30 Hz

* Ground truth segmentation images

3.1 EuRoC MAV Dataset

The EuRoC Micro Aerial Vehicle (MAV) dataset [13] used a stereo camera mounted on a micro aerial vehicle in an indoor environment. It contained a comprehensive suite of sensor measurements. A Vicon motion capture system and a Leica laser tracker was used as ground truth, and 11 scenarios were acquired including sequences categorized as easy, medium, and difficult in Machine Hall and Vicon Room. The ROS bag data format is supported.

The EuRoC MAV dataset has limitations in that the trajectories are short, and it provides only indoor environment data. However, it has been widely used as the reference dataset in many studies, and it is useful in this project in terms of data synchronization and algorithm development.

3.2 UZH-FPV Drone Dataset

The UZH-FPV drone dataset [14] provides drone racing data with more aggressive motion trajectories. It contains flight distances of 10 km in 27 sequences, captured on a first-person-view (FPV) racing quadrotor flown by pilot. The examples of the environment and trajectory are shown in Figure 1. The ROS bag data format is supported. A competition based on 6 sequences is posted on the webpage. A leader board of the top ranked methods obtaining the best accuracy is maintained. However, the ground truth data of the leader board is not public on these sequences.

This dataset provides indoor and outdoor environmental data for a forward facing and a 45° downward facing camera. Compared to the ground level datasets, the trajectories are short. However, it can be utilized to analyze the influence of the aggressive motion on the algorithm.

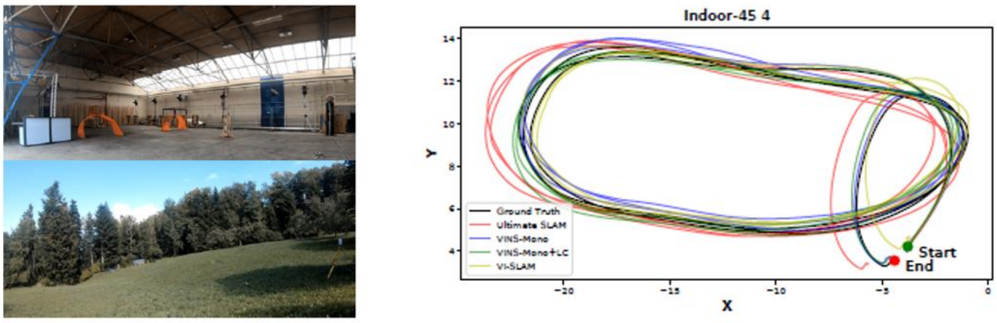

Figure 1 The examples of the environment (left) and trajectory (right) [14].

3.3 Blackbird UAV Dataset

The Blackbird dataset [15] contains 18 different trajectories at varying maximum speeds through 5 different visual environments. It also includes synchronized motion-capture ground truth data with inertial measurements using camera exposure timestamps. Examples of the environment and trajectory are shown in Figure 2.

This UAV dataset provides high-rate measurements, and can be used to develop visual inertial navigation, 3D reconstruction, and depth estimation algorithms. It contains trajectories for various sequences and conditions, but it also lacks longer trajectories (the max distance is up to 860 meters).

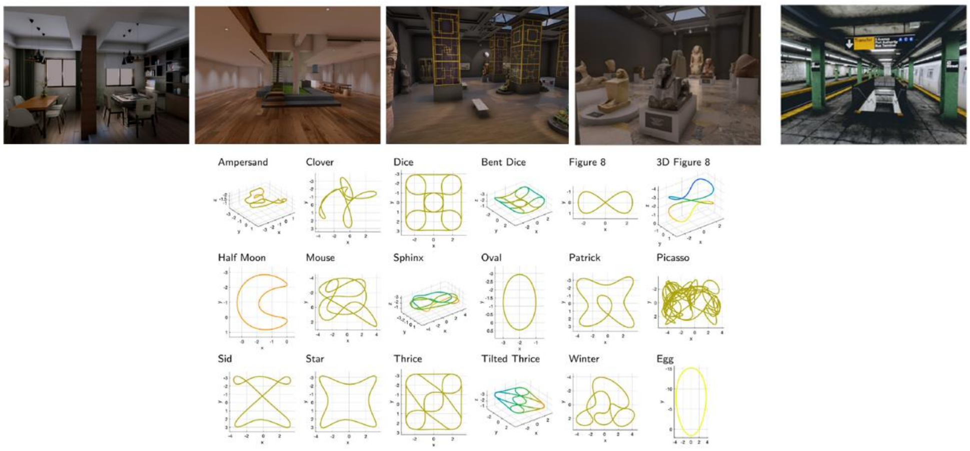

Figure 2 The examples of the environment (top 1 row) and trajectory (bottom) [15].

3.4 NTU VIRAL Dataset

The NTU visual-inertial-ranging-lidar (VIRAL) dataset [16] includes camera, IMU, LiDAR, and Ultra-wideband (UWB) ranging units. The overall sensor suite is a configuration that conforms to an autonomous vehicle, and the experiments are performed in a low-texture condition

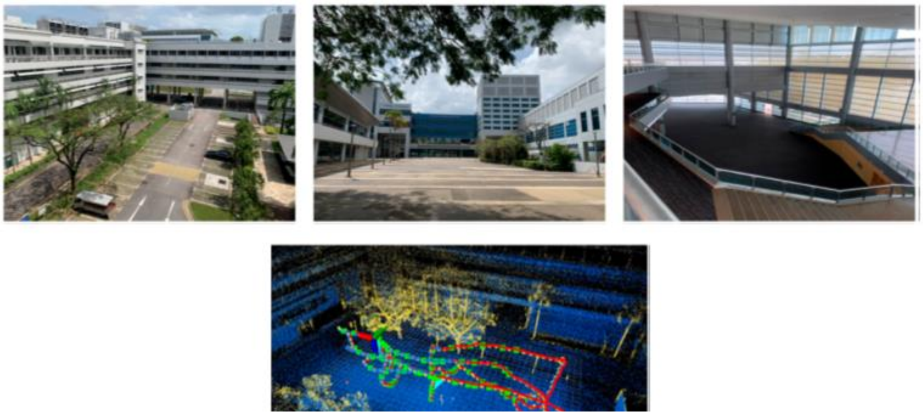

Figure 3 The examples of the environment (top) and trajectory (bottom) [16].

environment of a building area. With a sensor configuration oriented toward autonomous drones, data composed of various hardware among aviation datasets is provided and ROS bag format is supported. The examples of the environment and trajectories are shown in Figure 3.

The reference experimental results are provided with the latest vision and LiDAR based methods that make up the state of the art. Outdoor and indoor environments are provided, but the lengths of the trajectories are limited to within several hundred meters.

3.5 KAIST VIO Dataset

The KAIST visual-inertial odometry (VIO) dataset [17] provides ROS bag format for 4 different trajectories in the laboratory environment, and examples of the environment and trajectories are shown in Figure 4. Comparisons of results from visual-inertial-based state-of-the-art methods including VINS-Mono [18], VINS-Fusion [19], Kimera [20], ALVIO [21], Stereo-MSCKF [22], ORB-SLAM2 [23] stereo, and ROVIO [24] are provided. By comparing the algorithm performances in three types of NVIDIA Jetson platforms, it has the advantage of being a reference for analyzing the performance of the current state-of-the-art algorithm and hardware platform. However, it has the disadvantage of providing only sequences of the data in laboratory environment.

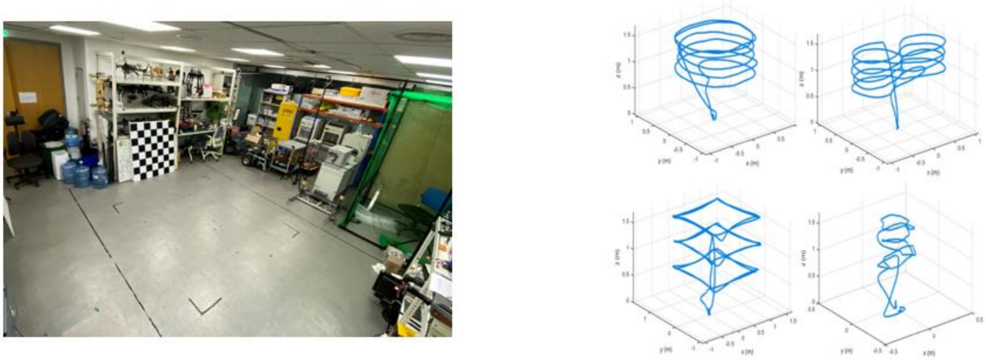

Figure 4 The examples of the environment (left) and trajectory (right) [17].

3.6 General Comment of the Datasets

The characteristic of the cited datasets is that they use low resolution images compared with datasets from autonomous vehicles. Considering the high-altitude aerial scenario, it will be necessary to take into account the resolution for extracting the image information required for navigation, and mutually consider the computational cost of high resolution of the image processing algorithms.

4 Feature-based Methods

Image-based navigation has been developed with feature based detection methods that uses extracted features from images. Recently, methods that directly utilize image pixels without extracting features have been actively studied but the basic structure of the algorithm is related to the existing methods, therefore, it is meaningful to understand the research flow of the feature-based methods.

To briefly recap of the feature detection flow, since the initial corner detection proposed by Maravec [25], the feature extracting method has been developed through interaction of the algorithm and computing power. The corner/edge detector proposed by Harris and Stephens [26] is one of the methods most used so far. The concept of visual odometry, first introduced by Nister [27] using a 5-point RANSAC algorithm, established the initial idea of vision-based navigation. After that, the SIFT method with scale-invariance and rotation-invariance was presented by Lowe [28]. This method is accurate but relatively slow. The FAST method proposed by Rosten [29] improved the processing speed compared to the Harris and SIFT methods. Meanwhile, to improve the performance of the descriptor, a more memory efficient binary descriptor, the BRIEF descriptor was proposed by Calonder [30]. The ORB feature, an oriented FAST and rotated BRIEF to improve both feature detection and descriptor performance, is proposed by Rublee [31]. It shows scale and rotation invariance performance and is the feature detector in ORB-SLAM [32] which is one of the state-of-the-art visual SLAM. More recently, Hamming Binary Search Tree (HBST) [33], a fast matching of binary feature descriptors using kd-tree, was proposed. In summary, a method is selected that meets the requirements of the algorithm to be developed in terms of accuracy of SIFT and speed of ORB.

Mono-SLAM [34], an EKF-based SLAM, was presented as a visual navigation system to obtain the camera pose by deriving translation between image frames. It extracts 2D features and mapping it to 3D based on a Kalman filter. However, it requires an object of prior knowledge in the initialization step and is limited to a small area.

PTAM [35] which uses the FAST corner detector and optimizes the map with bundle adjustment was proposed as a method to improve performance. This approach splits tracking and maps them into two threads and uses a keyframe idea in the mapping step. However, it is limited to small scale operation and manual initialization is required.

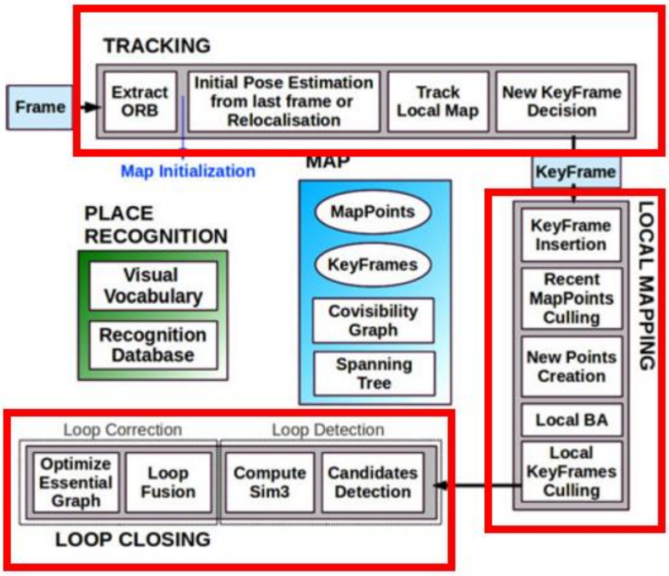

Figure 5 The basic threads of ORB-SLAM [32].

ORB-SLAM [32] is a method actively used in research and the industrial setting, showing excellent performance. It is a feature-based method. It uses ORB feature and overcomes the shortcomings of manual initialization of PTAM by extracting a keyframe-based homography and fundamental matrix with automatic initialization. It draws attention as an optimization-based pipeline with 3 parallel threads structure of tracking, local mapping, and loop closing as shown in Figure 5. This architecture is computationally efficient and does not require high-end computing power. Based on this structure, an improved version, ProSLAM [36], was proposed. In ORB-SLAM2 [23] and ORB-SLAM3 [37] sensor such as stereo and RGB-D cameras are added, and the source code is continuously updated.

Feature-based methods require more computing resources compared to the direct methods to handle the feature extraction step and generate sparse feature information which can be different from the real environment. The weakest point is poor performance with a low number of features or featureless environment. This is the background why the direct methods that directly utilize image pixels without extracting features has been proposed.

5 Direct Methods

As the performance of computer and camera sensors improves, methods using a direct photogrammetric method that compares the entire image, rather than extracting features, have been widely studied. DTAM [38] using parallel operation and keyframe selection in PTAM has been proposed, and dense measurement information was provided with robust performance in featureless environments with motion-blur. Although this method provides a basic idea for a dense map generation, it is limited to a small area and requires a GPU for real-time operation. However, it is recognized as an early study in the development of the direct methods that can compensate for the shortcomings of the feature-based methods in featureless environments and in the presence of motion blur.

5.1 LSD-SLAM

The large-scale direct monocular SLAM (LSD-SLAM) [39] implemented the direct method in the CPU and reconstructed the semi-dense map, as shown in Figure 6. Tracking of camera frames is calculated with transform estimation of subsequent image frames with extracting lines through image gradient calculation. It has the characteristics of a light algorithm and can be executed in mobile environment with good performance in large scale area. It is defined as semi-dense because it uses line information rather than feature points or entire images. It can provide a method of how to specify the feature density based on the characteristics of the environment.

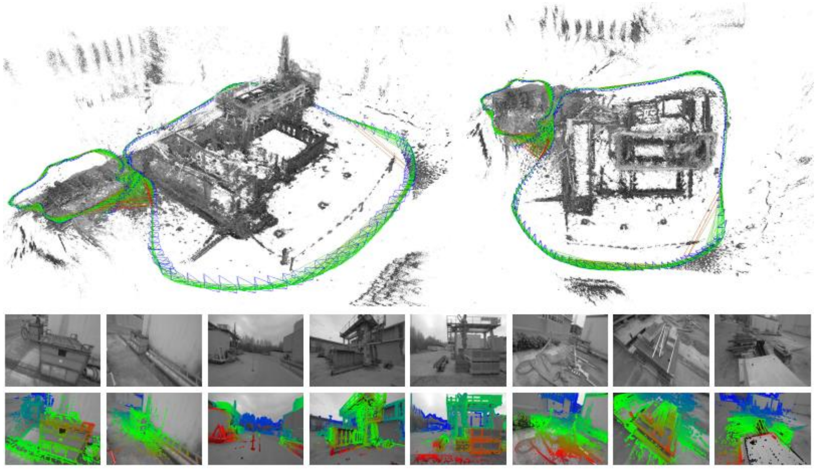

Figure 6 The examples of the LSD-SLAM [39].

5.2 SVO

The semi-direct visual odometry (SVO) [40] method is a method that focuses on increasing the speed of the direct method. It uses the direct photogrammetric method for frame-to-frame tracking and a feature-based method for bundle adjustment. The direct method is used for random selection of sub-image patches without using the entire image, as shown in Figure 7. It led to a speed-up of the direct method, and SOV2 [41] was proposed to extend a multi-camera system including a fisheye camera.

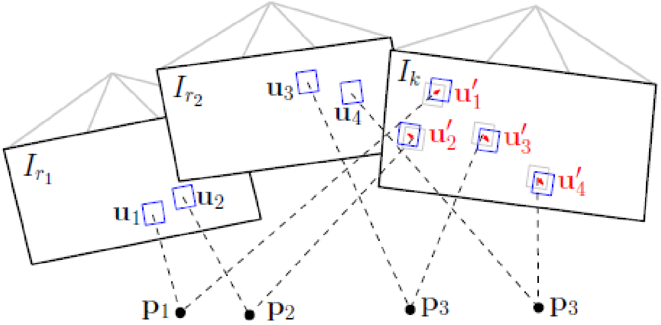

Figure 7 The optimization 2D position of each patch in SVO [40].

5.3 DSO

The direct sparse odometry (DSO) [42] method was first proposed as a sliding window bundle adjustment method for the direct tracking method without using a feature-based method and loop closure, as shown in Figure 8. It provides reliable tracking performance in large scale environments without loop closure, and it has the less translation error than LSD-SLAM and ORB-SLAM especially in featureless environments. Extensions such as separating threads from the DSO and adding other sensors have been proposed, and recently the delayed marginalization visual-inertial odometry (DM-VIO) [43] combined with an IMU was proposed.

When comparing the performance of the sparse, semi-dense, and dense methods according to the distance between frames, the three methods show similar performance when the frame interval is small. Therefore, the dense method tends to be used for 3D reconstruction, surface restoration, and dense map construction. The sparse method is used for general odometry and SLAM purposes. Compared to the feature-based methods, the direct methods skip the feature extraction and matching steps, that is a benefit for high camera frame rate.

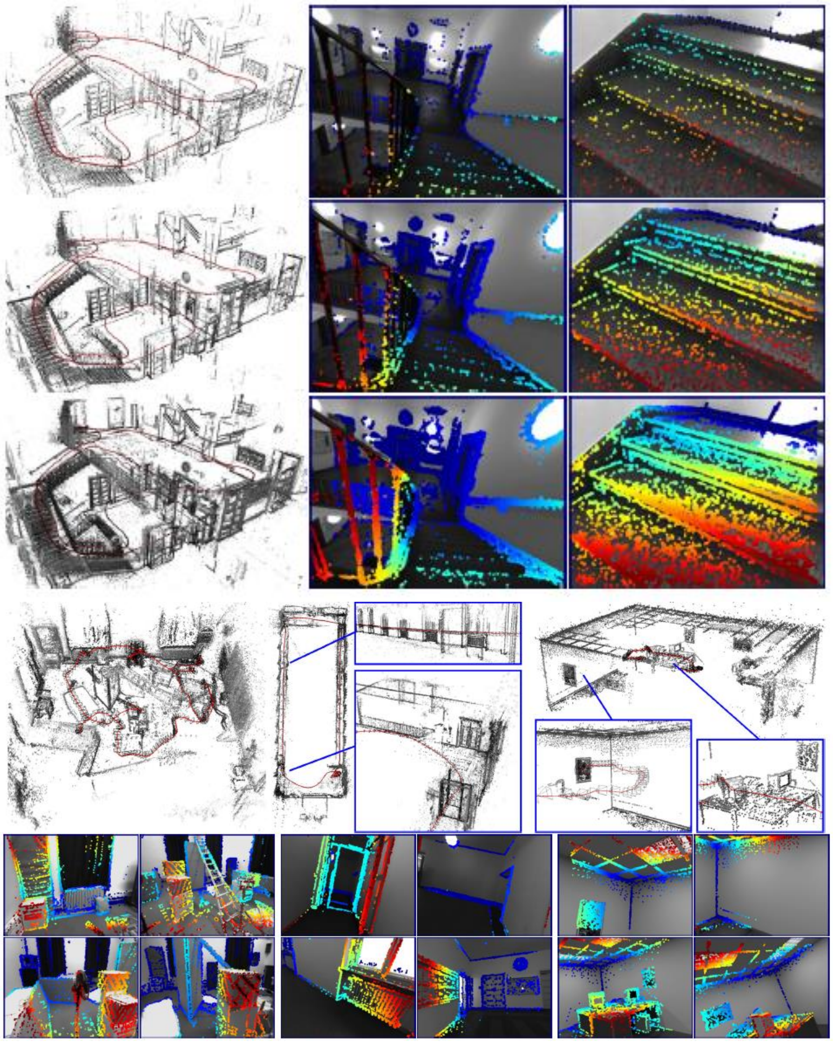

Figure 8 The examples of the DSO [42].

The direct methods need the high camera frame rate. However, fast cameras consume more power, therefore, it requires more from the hardware. Moreover, the monocular-based direct method still has the scale issue. Therefore, visual-inertial methods that integrate the IMU sensor have been proposed to solve the scale issue and fast localization performance considering the hardware performance.

6 Visual-inertial Methods

The visual-inertial method, which is a method that fuses camera and IMU information, has been actively studied by solving the issues of the extrinsic calibration and time synchronization of frame rates. This chapter summarizes the visual-inertial methods in which the IMU sensor and camera are fused.

6.1 MSCKF

The multi-state constraint Kalman filter (MSCKF) [44] has the characteristic that the algorithm is fast because only the last state is updated. Although the number of features has to be limited because the covariance increases quadratically with the number of features, but it provides good performance in an environment with sufficient features and is actively utilized in for example Google's ARCore [45] and Apple's ARKit [46].

The filter-based method has a limitation on the number of features, which has led to the proposal of the optimization approach. In addition, the limited number of features leads to lack of measurements, which reduces the IMU calibration performance. On the other hand, the optimization-based methods have the disadvantage of handling a lot of parameters.

6.2 Rovio/Rovioli/Maplab

The robust visual inertial odometry (Rovio) [24] method was proposed as an EKF-based feature tracking and IMU fusion method. It extracts FAST corners and surrounding sub-images and uses the direct method on the image for tracking. The Rovioli [47] method proposed EKF-based IMU camera filtering, and the Maplab [48] method supported the offline entire map building within the framework.

6.3 VINS-Mono/VINS-Fusion

The IMU pre-integration [49] method which integrates the IMU parameters between camera frames has been proposed as an improvement of the real-time computational performance by reducing the number of parameters. This can reduce the computational burden of the algorithm. The IMU pre-integration is widely used in research of the tightly coupled systems.

The VINS-Mono [18] modularized the pipeline of the visual-inertial method and is considered one of the leading methods in this fields. Tightly coupled pre-integration and loop closure for global optimization were proposed. It optimizes position with sliding window optimization using a feature-based method. It contains loosely coupled bootstrap initialization from arbitrary state and has expanded the hardware configuration with sensors such as stereo-IMU fusion (VINS-Fusion) [19] in large scale environments. The pipeline of the VINS series is cited as the reference pipeline for visual-inertial framework, and the pipeline of the VINS-Mono is shown in Figure 9.

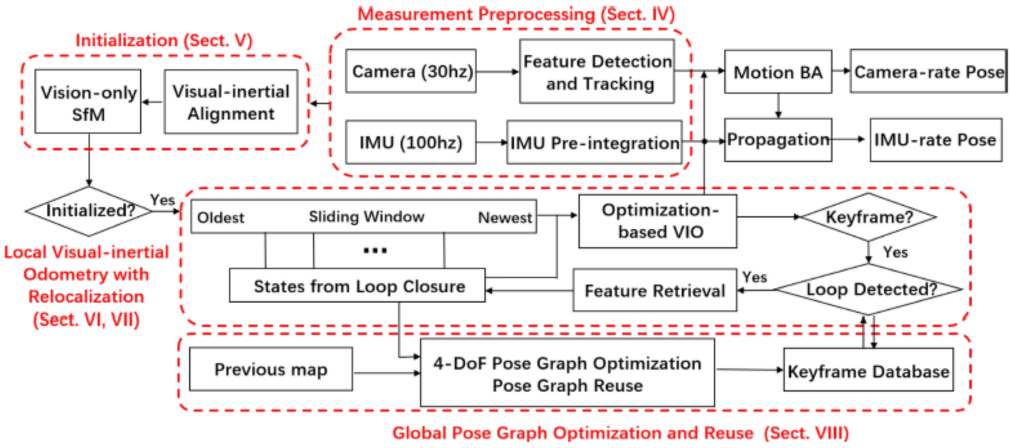

Figure 9 The pipeline of the monocular VINS [18].

The optimization method and filter-based approaches have been actively studied complementary to each other. Therefore, both methods are important for confirming the latest research.

7 Deep Learning applied Methods

With the development of deep learning algorithms, various studies have been conducted to apply learning techniques to the visual navigation problem.

CNN SLAM [50] was proposed by combining deep learning-based depth estimation and the LSD-SLAM backend. It can estimate the camera pose from pure rotation through depth estimation without baseline movement. This method improves the performance of deep learning-based depth estimation at sharp edges with the LSD-SLAM backend. However, its robustness is limited in large scale environments and long-term navigation.

A study applying the deep learning method to the existing feature extraction and descriptor has also been proposed. The SuperPoint [51] method was proposed that uses a deep learning-based corner detector, as shown in Figure 10. It creates a synthetic corner dataset using OpenGL, and the domain is transferred from synthetic to real scene. The self-supervised learning-based corner detector and descriptor are presented in a real scene. The LIFT [52] benchmarked on SIFT was proposed as a method of outputting detection, orientation, and descriptor using a Siamese network and a spatial transformer [53]. Compared to SIFT, it improves the performance in feature extraction and matching rate.

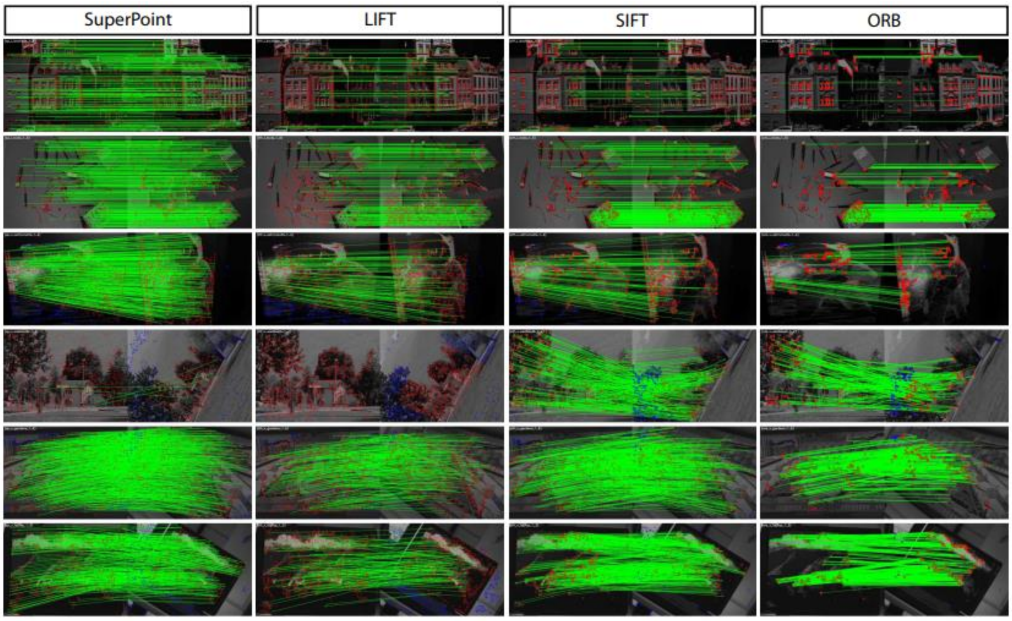

Figure 10 The examples of the feature extraction and descriptor [52].

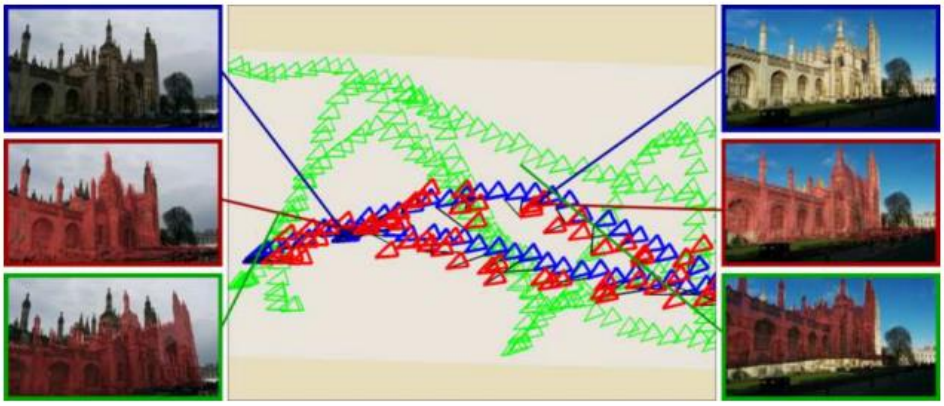

Figure 11 The camera pose prediction (red) from training (green) and testing (blue) of the PoseNet [54].

The PoseNet [54] was proposed as a GoogLeNet-based pose regression. It is trained on SfM data and do not use any map information for pose regression. The example of the PoseNet is shown in Figure 11. With image-input and pose-output structure, it requires sophisticated SfM with existing data in the training step. Therefore, the performance is highly data dependent.

The CubsSLAM [55] was proposed in a more object dependent method, and it uses 3D object detection with bundle adjustment for pose computation, as shown in Figure 12. It is considered a useful method to use when the objects have continuity and representation in the scene.

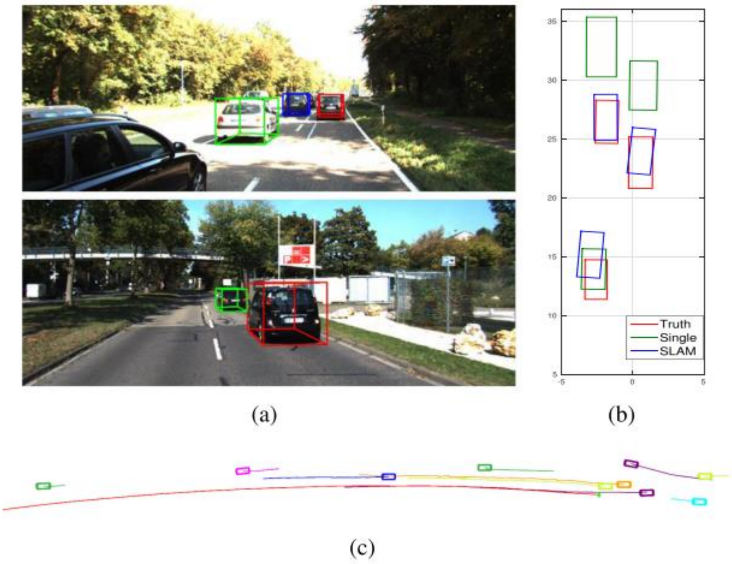

Figure 12 The CubeSLAM. (a) Sample frames, (b) Top view comparison, (c) pose estimation [55].

The DeepFactors [56] method introduces the idea of combining the advantages of previous image-based navigation paradigms. It considers 3 kind of errors, geometric error from feature-based, photometric error from direct method, and geometric error from deep learning depth estimation. It creates a dens-map and focuses on 3D reconstruction. The method shows clean 3D surface reconstruction using the factor graph, but it still has limitation in the large-scale scenes.

Studies on how to apply the deep learning algorithm to navigation system have also been conducted. A study on how intermediate representations influence the result [57] suggests that using an intermediate representations method including the depth estimation or optical flow performed better than end-to-end learning. A study on limitations of CNN-based camera pose estimation [58] suggests that pose estimation in an unknown environment results in interpolation, that can cause performance degradation. By presenting a theoretical model for camera pose estimation, it explains why these methods do not achieve the same level of pose accuracy as 3D structure-based methods. Because of the limitations of the image net dataset for 3D camera pose estimation, it claims that the algorithms need to be improved.

Deep learning applied methods started with replacing obvious components such as feature detection, descriptor, and depth estimation. End-to-end learning methods for direct camera pose estimation have also been studied. The learning-based algorithms are highly data-dependent. The aerial data is relatively sparse compared to the ground-level data, and research on algorithm development that supplements this is needed.

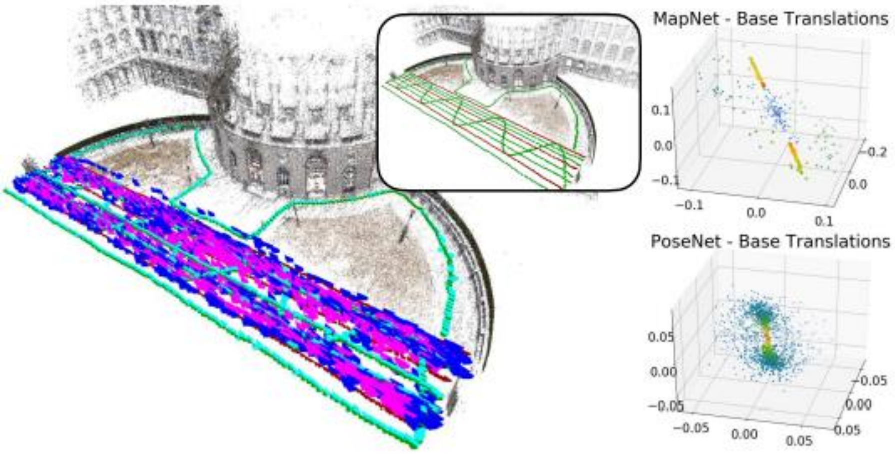

Figure 13 Visualization of the translational errors of the learning-based pose [58].

8 Summary

This report summarized currently available open datasets and described landmark papers related to the project. By reviewing from the traditional image feature extraction methods to the latest methods applying deep learning, it can help set the initial direction of the algorithm development for the composition of image features and navigation structure suitable in the early stage of the project.

References

1. DURRANT-WHYTE, Hugh; BAILEY, Tim. Simultaneous localization and mapping: part I. *IEEE robotics & automation magazine*, 2006, 13.2: 99-110.

2. BAILEY, Tim; DURRANT-WHYTE, Hugh. Simultaneous localization and mapping (SLAM): Part II. *IEEE robotics & automation magazine*, 2006, 13.3: 108-117.

3. THRUN, Sebastian. Probabilistic robotics. *Communications of the ACM*, 2002, 45.3: 52-57.

4. STACHNISS, Cyrill; LEONARD, John J.; THRUN, Sebastian. Simultaneous localization and mapping. In: *Springer Handbook of Robotics*. Springer, Cham, 2016. p. 1153-1176.

5. GRISETTI, Giorgio, et al. A tutorial on graph-based SLAM. *IEEE Intelligent Transportation Systems Magazine*, 2010, 2.4: 31-43.

6. NISTÉR, David; NARODITSKY, Oleg; BERGEN, James. Visual odometry. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. Ieee, 2004. p. I-I.

7. SCARAMUZZA, Davide; FRAUNDORFER, Friedrich. Visual odometry [tutorial]. *IEEE robotics & automation magazine*, 2011, 18.4: 80-92.

8. FRAUNDORFER, Friedrich; SCARAMUZZA, Davide. Visual odometry: Part ii: Matching, robustness, optimization, and applications. *IEEE Robotics & Automation Magazine*, 2012, 19.2: 78-90.

9. GEIGER, Andreas; LENZ, Philip; URTASUN, Raquel. Are we ready for autonomous driving? the kitti vision benchmark suite. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012. p. 3354-3361.

10. CAESAR, Holger, et al. nuscenes: A multimodal dataset for autonomous driving. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020. p. 11621-11631.

11. ETTINGER, Scott, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. p. 9710-9719.

12. CHANG, Ming-Fang, et al. Argoverse: 3d tracking and forecasting with rich maps. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019. p. 8748-8757.

13. BURRI, Michael, et al. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 2016, 35.10: 1157-1163.

14. DELMERICO, Jeffrey, et al. Are we ready for autonomous drone racing? the UZH-FPV drone racing dataset. In: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019. p. 6713-6719.

15. ANTONINI, Amado, et al. The blackbird uav dataset. *The International Journal of Robotics Research*, 2020, 39.10-11: 1346-1364.

16. NGUYEN, Thien-Minh, et al. NTU VIRAL: A visual-inertial-ranging-lidar dataset, from an aerial vehicle viewpoint. *The International Journal of Robotics Research*, 2021, 02783649211052312.

17. JEON, Jinwoo, et al. Run your visual-inertial odometry on NVIDIA Jetson: Benchmark tests on a micro aerial vehicle. *IEEE Robotics and Automation Letters*, 2021, 6.3: 5332-5339.

18. QIN, Tong; LI, Peiliang; SHEN, Shaojie. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 2018, 34.4: 1004-1020.

19. QIN, Tong, et al. A general optimization-based framework for global pose estimation with multiple sensors. *arXiv preprint arXiv:1901.03642*, 2019.

20. ROSINOL, Antoni, et al. Kimera: an open-source library for real-time metric-semantic localization and mapping. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020. p. 1689-1696.

21. JUNG, KwangYik, et al. ALVIO: Adaptive line and point feature-based visual inertial odometry for robust localization in indoor environments. In: *RiTA 2020*. Springer, Singapore, 2021. p. 171-184.

22. SUN, Ke, et al. Robust stereo visual inertial odometry for fast autonomous flight. *IEEE Robotics and Automation Letters*, 2018, 3.2: 965-972.

23. MUR-ARTAL, Raul; TARDÓS, Juan D. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 2017, 33.5: 1255-1262.

24. BLOESCH, Michael, et al. Robust visual inertial odometry using a direct EKF-based approach. In: *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015. p. 298-304.

25. MORAVEC, Hans Peter. *Obstacle avoidance and navigation in the real world by a seeing robot rover*. 1980. PhD Thesis. Stanford University.

26. HARRIS, Chris, et al. A combined corner and edge detector. In: *Alvey vision conference*. 1988. p. 10-5244.

27. NISTÉR, David. An efficient solution to the five-point relative pose problem. *IEEE transactions on pattern analysis and machine intelligence*, 2004, 26.6: 756-770.

28. LOWE, David G. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004, 60.2: 91-110.

29. ROSTEN, Edward; DRUMMOND, Tom. Fusing points and lines for high performance tracking. In: *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*. Ieee, 2005. p. 1508-1515.

30. CALONDER, Michael, et al. Brief: Binary robust independent elementary features. In: *European conference on computer vision*. Springer, Berlin, Heidelberg, 2010. p. 778-792.

31. RUBLEE, Ethan, et al. ORB: An efficient alternative to SIFT or SURF. In: *2011 International conference on computer vision*. Ieee, 2011. p. 2564-2571.

32. MUR-ARTAL, Raul; MONTIEL, Jose Maria Martinez; TARDOS, Juan D. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE transactions on robotics*, 2015, 31.5: 1147-1163.

33. SCHLEGEL, Dominik; GRISETTI, Giorgio. HBST: A hamming distance embedding binary search tree for feature-based visual place recognition. *IEEE Robotics and Automation Letters*, 2018, 3.4: 3741-3748.

34. DAVISON, Andrew J., et al. MonoSLAM: Real-time single camera SLAM. *IEEE transactions on pattern analysis and machine intelligence*, 2007, 29.6: 1052-1067.

35. KLEIN, Georg; MURRAY, David. Parallel tracking and mapping for small AR workspaces. In: *2007 6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE, 2007. p. 225-234.

36. SCHLEGEL, Dominik; COLOSI, Mirco; GRISETTI, Giorgio. Proslam: Graph SLAM from a programmer's Perspective. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018. p. 3833-3840.

37. CAMPOS, Carlos, et al. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 2021, 37.6: 1874-1890.

38. NEWCOMBE, Richard A.; LOVEGROVE, Steven J.; DAVISON, Andrew J. DTAM: Dense tracking and mapping in real-time. In: *2011 international conference on computer vision*. IEEE, 2011. p. 2320-2327.

39. ENGEL, Jakob; SCHÖPS, Thomas; CREMERS, Daniel. LSD-SLAM: Large-scale direct monocular SLAM. In: *European conference on computer vision*. Springer, Cham, 2014. p. 834-849.

40. FORSTER, Christian; PIZZOLI, Matia; SCARAMUZZA, Davide. SVO: Fast semi-direct monocular visual odometry. In: *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014. p. 15-22.

41. FORSTER, Christian, et al. SVO: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 2016, 33.2: 249-265.

42. ENGEL, Jakob; KOLTUN, Vladlen; CREMERS, Daniel. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 2017, 40.3: 611-625.

43. VON STUMBERG, Lukas; CREMERS, Daniel. DM-VIO: Delayed Marginalization Visual-Inertial Odometry. *IEEE Robotics and Automation Letters*, 2022.

44. MOURIKIS, Anastasios I., et al. A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation. In: *ICRA*. 2007. p. 6.

45. VOINEA, Gheorghe-Daniel, et al. Exploring cultural heritage using augmented reality through Google's Project Tango and ARCore. In: *International conference on VR Technologies in Cultural Heritage*. Springer, Cham, 2018. p. 93-106.

46. DILEK, Ufuk; EROL, Mustafa. Detecting position using ARKit II: generating position-time graphs in real-time and further information on limitations of ARKit. *Physics Education*, 2018, 53.3: 035020.

47. BLOESCH, Michael, et al. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research*, 2017, 36.10: 1053-1072.

48. SCHNEIDER, Thomas, et al. maplab: An open framework for research in visual-inertial mapping and localization. *IEEE Robotics and Automation Letters*, 2018, 3.3: 1418-1425.

49. FORSTER, Christian, et al. On-manifold preintegration for real-time visual--inertial odometry. *IEEE Transactions on Robotics*, 2016, 33.1: 1-21.

50. TATENO, Keisuke, et al. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. p. 6243-6252.

51. DETONE, Daniel; MALISIEWICZ, Tomasz; RABINOVICH, Andrew. Superpoint: Self-supervised interest point detection and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018. p. 224-236.

52. YI, Kwang Moo, et al. Lift: Learned invariant feature transform. In: *European conference on computer vision*. Springer, Cham, 2016. p. 467-483.

53. JADERBERG, Max, et al. Spatial transformer networks. *Advances in neural information processing systems*, 2015, 28

54. KENDALL, Alex; GRIMES, Matthew; CIPOLLA, Roberto. Posenet: A convolutional network for real-time 6-dof camera relocalization. In: *Proceedings of the IEEE international conference on computer vision*. 2015. p. 2938-2946.

55. YANG, Shichao; SCHERER, Sebastian. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 2019, 35.4: 925-938.

56. CZARNOWSKI, Jan, et al. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 2020, 5.2: 721-728.

57. ZHOU, Brady; KRÄHENBÜHL, Philipp; KOLTUN, Vladlen. Does computer vision matter for action?. *Science Robotics*, 2019.

58. SATTLER, Torsten, et al. Understanding the limitations of cnn-based absolute camera pose regression. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019. p. 3302-3312.

Avdelning, Institution
Division, Department

Division of Automatic Control
Department of Electrical Engineering

Datum
Date

2023-04-13

Språk
Language

☐ Svenska/Swedish
☒ Engelska/English
☐ ________________

Rapporttyp
Report category

☐ Licentiatavhandling
☐ Examensarbete
☐ C-uppsats
☐ D-uppsats
☒ Övrig rapport
☐ ________________

URL för elektronisk version

`http://www.control.isy.liu.se`

ISBN
—

ISRN
—

Serietitel och serienummer
Title of series, numbering

ISSN
1400-3902

LiTH-ISY-R-3109

Titel
Title

State-of-the-art Report of Research about Multi Sensor Image-based Navigation

Författare
Author

Jeongmin Kang, Zoran Sjanic, Gustaf Hendeby

Sammanfattning
Abstract

This report aims to describe the latest research and method development of image-based multi sensor fusion navigation and summarizes open aerial datasets which can support the latest research related to this project. It supports the initial setting of the direction of the algorithm development in the early stage of the project.

The Multi Sensor Image-based Navigation project aims to study and develop the methods focusing on image-based multisensor navigation in order to acquire a precise localization of the aircraft. GNSS-based localization and navigation systems are sensitive to disturbances and jamming, hence the capability to provide reliable position accuracy without GNSS is a key element to develop the navigation systems.

The output of this project can be utilized in a wide range of applications, such as aircraft operation in GNSS denied environments or urban air mobility context.

Nyckelord
Keywords

sensor fusion, aerial navigation, image processing, simultaneous localization and mapping, visual-inertial navigation, deep learning