Visual Odometry

Christos Kokas

May 2022

Contents

1 Introduction 4

2 Definitions/Appendix 4

3 Literature Review 5

3.1 ORB_SLAM2 5

3.1.1 Abstract 5

3.1.2 System Overview 6

3.1.3 Key Points 7

3.2 Hydra: A Real-time Spatial Perception Engine for 3D Scene Graph Construction and Optimization 7

3.2.1 Abstract 7

3.2.2 Building 3D scene Graphs 8

3.2.3 Key Points 10

3.2.4 Loop Closure Detection 11

3.2.5 Geometric Verification 11

3.2.6 3D Scene Graph Optimization 11

3.2.7 The Hydra Architecture 12

3.3 VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator 12

3.3.1 Abstract 12

3.3.2 Issues 13

3.3.3 VINS-Mono Features 13

3.4 SVO: Fast Semi-Direct Monocular Visual Odometry 16

3.4.1 Abstract 16

3.4.2 Key Points 16

3.4.3 System Overview 17

3.5 Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping 18

3.5.1 Abstract 18

3.5.2 Key Points 19

3.6 A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors 20

3.6.1 Abstract 20

4 Feature Detection 20

4.1 Interest Point Detection 20

4.1.1 Harris detection 21

5 Ideas 21

6 TO DO 23

1 Introduction

In robotics and computer vision, visual odometry is the process of determining the position and orientation of a robot by analyzing images from the robot's camera. It has been used in a wide variety of robotic applications, such as on the Mars Exploration Rovers [1].

The accurate localization of the robot poses a great challenge for many applications, even more in GPS-Denied environments (e.g. indoors) where GPS systems, such as RTK GPS, have difficulty producing an accurate estimate of the position of the robot. Visual Odometry has gained popularity because it can produce accurate estimates of the position of the camera even in GPS-Denied environments and at the same time be cheaper than other alternatives (e.g. RTK GPS, LiDARS). Combinations of Visual Odometry and other techniques or products, such as Visual Odometry combined with an IMU (also know as Visual Inertial Odometry) or Visual Odometry combined with Deep Learning, can also yield satisfying results.

2 Definitions/Appendix

Bundle Adjustment is simultaneous refining of the 3D coordinates describing the scene geometry, the parameters of the relative motion, and the optical characteristics of the camera(s) employed to acquire the images, given a set of images depicting a number of 3D points from different viewpoints. It amounts to an optimization problem on the 3D structure and viewing parameters (i.e., camera pose and possibly intrinsic calibration and radial distortion), to obtain a reconstruction which is optimal under certain assumptions regarding the noise pertaining to the observed image features.

Reprojection Error is a geometric error corresponding to the image distance between a projected point and a measured one.

Pose Graph contains nodes connected by edges that represents the pose of the robot.

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available.

Voxel each of an array of elements of volume that constitute a notional three-dimensional space, especially each of an array of discrete elements into which a representation of a three-dimensional object is divided.

Euclidean clustering groups points that are close together. You must set a ”closeness” threshold, such that points within this threshold are considered to be part of the same cluster.

DBoW2 DBoW2 is an open source C++ library for indexing and converting images into a bag-of-word representation. It implements a hierarchical tree for approximating nearest neighbours in the image feature space and creating a visual vocabulary.

Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of the estimates.

Parallax Parallax is a displacement or difference in the apparent position of an object viewed along two different lines of sight, and is measured by the angle or semi-angle of inclination between those two lines.

Keyframe is a drawing or shot that defines the starting and ending points of any smooth transition.

Structure from motion (SfM) is a photogrammetric range imaging technique for estimating three-dimensional structures from two-dimensional image sequences that may be coupled with local motion signals.

Feature Correspondence refers to the problem of ascertaining which parts of one image correspond to which parts of another image, where differences are due to movement of the camera, the elapse of time, and/or movement of objects in the photos.

3 Literature Review

In this section papers using Visual Odometry will be presented each using a different method of producing an accurate estimate of the camera.

3.1 ORB_SLAM2

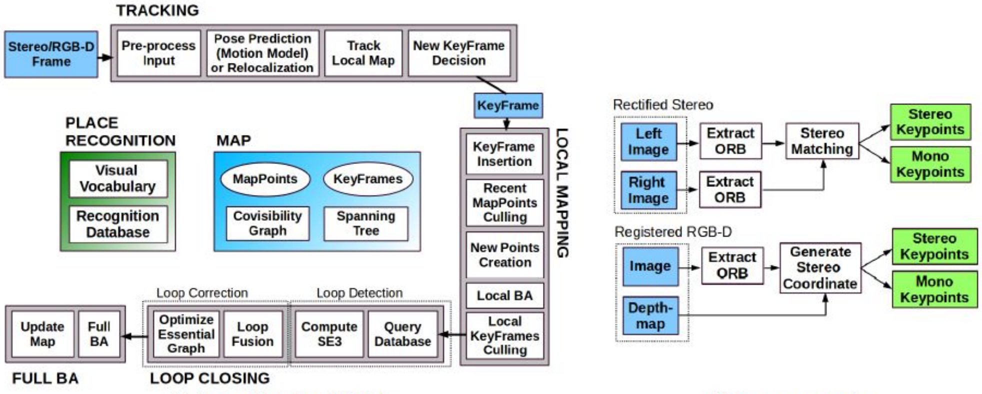
3.1.1 Abstract

We present ORB-SLAM2 a complete SLAM system for monocular, stereo and RGB-D cameras, including map reuse, loop closing and relocalization capabilities. The system works in real-time on standard CPUs in a wide variety of environments from small hand-held indoors sequences, to drones flying in industrial environments and cars driving around a city. Our backend based on bundle adjustment with monocular and stereo observations allows for accurate trajectory estimation with metric scale. Our system includes a lightweight localization mode that leverages visual odometry tracks for unmapped regions and matches to map points that allow for zero-drift localization. The evaluation on 29 popular public sequences shows that our method achieves state-of-the-art accuracy, being in most cases the most accurate SLAM solution. We publish the source code, not only for the benefit of the SLAM community, but with the aim of being an out-of-the-box SLAM solution for researchers in other fields [2].

3.1.2 System Overview

ORB-SLAM2 for stereo and RGB-D camera has three main parallel threads.

- Find Feature Matches to the local map with every frame to localize the camera, and apply motion-only Bundle Adjustment (BA) to minimize the reprojection error. (Tracking)
- Perform local BA to manage and optimize the local map. A local map is a simplified representation of the immediate environment around the robot. (Local Mapping)
- Detect large loops and correct the accumulated drift by performing a pose-graph optimization. (Loop Closure)



(a) System Threads and Modules.

(b) Input pre-processing

Figure 3.1: ORB-SLAM2 Diagram of Parallel Threads

3.1.3 Key Points

ORB-SLAM2's key points include :

- ORB-SLAM2 treats differently close and far points, first described in the work of Paz et al. [3], which indicates that the points' depth cannot be reliably estimated due to little disparity in the stereo camera. If the points' depth is more than ~ 40 times the stereo baseline it cannot be reliably estimated.

- Stereo keypoints (points that are found both in left and right image) contribute to distance from the object as well as rotation and translation estimation. On the other hand, Monocular keypoints (points that are only found in one of the cameras) are used for rotation and translation estimation only.

- For camera pose optimization the Levenberg–Marquardt method is implemented in g2o [4].

- Full BA optimization is used to achieve the optimal solution. This optimization is very costly and therefore is performed on a separate thread allowing the system to continue creating map and detecting loops. If a loop is detected before the Full BA finishes, it is aborted, the loop is closed, and the Full BA optimization is launched again. To merge the updated subset of keyframes and points optimized by the full BA with the non-updated keyframes and points that where inserted while the optimization was running is done by propagating the correction of updated keyframes (i.e. the transformation from the non-optimized to the optimized pose) to non-updated keyframes through the spanning tree. Non-updated points are transformed according to the correction applied to their reference keyframe.

3.2 Hydra: A Real-time Spatial Perception Engine for 3D Scene Graph Construction and Optimization

3.2.1 Abstract

3D scene graphs have recently emerged as a powerful high-level representation of 3D environments. A 3D scene graph describes the environment as a layered graph where nodes represent spatial concepts at multiple levels of abstraction (from low-level geometry to high-level semantics including objects, places, rooms, buildings, etc.) and edges represent relations between concepts. While 3D scene graphs can serve as an advanced "mental model" for robots, how to build such a rich representation in real-time is still uncharted territory.

This paper describes the first real-time Spatial Perception engINe (SPIN), a suite of algorithms to build a 3D scene graph from sensor data in real-time. Our first contribution is to develop real-time algorithms to incrementally construct the layers of a scene graph as the robot explores the environment; these algorithms build a local Euclidean Signed Distance Function (ESDF) around the current robot location, extract a topological map of places from the ESDF, and then segment the places into rooms using an approach inspired by community-detection techniques. Our second contribution is to investigate loop closure detection and optimization in 3D scene graphs. We show that 3D scene graphs allow defining hierarchical descriptors for loop closure detection; our descriptors capture statistics across layers in the scene graph, ranging from low-level visual appearance, to summary statistics about objects and places. We then propose the first algorithm to optimize a 3D scene graph in response to loop closures; our approach relies on embedded deformation graphs to simultaneously correct all layers of the scene graph. We implement the proposed SPIN into a highly parallelized architecture, named Hydra, that combines fast early and midlevel perception processes (e.g., local mapping) with slower highlevel perception (e.g., global optimization of the scene graph). We evaluate Hydra on simulated and real data and show it is able to reconstruct 3D scene graphs with an accuracy comparable with batch offline methods, while running online [5].

3.2.2 Building 3D scene Graphs

A 3D scene graph is a layered graph where nodes represent spatial concepts at multiple levels of abstraction (from low-level geometry to objects, places, rooms, buildings, etc.) and edges represent relations between concepts.

- Euclidean Signed Distance Function (ESDF) : From a field of booleans and produces a field of scalars such that each value in the output is the distance to the nearest "true" cell in the input. Unfortunately, ESDFs scale poorly in the size of the environment [6].

This paper focuses on indoor environments and adopts the 3D scene graph introduced in [7]. The 3D scene graph consists of 5 layers that represent semantics, objects and agents, places, rooms and lastly a building node connecting all rooms. Edges connect nodes within each layer or across layers.

- Mesh and Objects (Layers 1-2) : The construction of the metric-semantic 3D mesh is an extension of the approach in [8]. The authors in [8] use Voxblox to integrate semantically-labeled point clouds into a monolithic Truncated Signed Distance Field (TSDF) and an ESDF of the environment, while also performing Bayesian inference (Defined in Chapter 2) over the semantic label of each voxel. This paper forms a volumetric model of the robot's surroundings within a radius to bound the memory used by the ESDF. Subsequently, they extract he 3D metric-semantic mesh using Voxblox' marching cubes implementation and the places, and they are passed to the Scene Graph Frontend. They segment objects by performing euclidean clustering (EC) of the 3D meshes. The results of the EC are then used to create a bounding box for each putative object.

- Places (Layer 3) : To extract the subgraph of places the authors of this paper use a Generalized Voronoi Diagram (GVD). After the GVD they identify distinctive points and connect them with edges to form the graph of places.

- Rooms (Layer 4) : The segmentation of rooms was created directly from the sparse subgraph of places. Firstly, by inflating objects, small apertures in the environment (i.e., doors) will gradually close, naturally partitioning the voxel-based map into disconnected components (i.e., rooms). Secondly, with a dilation of the map by a distance δ, every place with obstacle distance smaller than δ will disappear from the graph (since it will no longer be in the free space). A visualization of this idea is given in Figure 3.2.

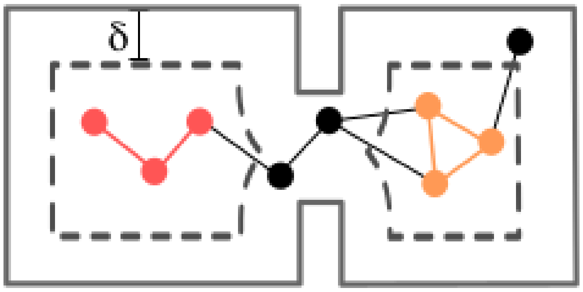


Figure 3.2: Map Dilation for Room Detection

3.2.3 Key Points

This paper overcomes that ESDFs scale poorly depending on the size of the environment by implementing algorithms that reconstruct a local ESDF of the robot's surroundings and incrementally convert the ESDF into a metric-semantic 3D mesh as well as a Generalized Voronoi Diagram, from which a topological graph of places can be quickly extracted. The main contribution of this paper is the development of the first real-time Spatial Perception engINe (SPIN), a suite of algorithms and implementations to build a 3D scene graph from sensor data in real-time.

This paper additionally investigates loop closure detection and optimization in 3D scene graphs. It proposes a hierarchical approach for loop closure detection :

- top-down loop closure detection that uses hierarchical descriptors —capturing statistics across layers in the scene graph— to find putative loop closures
- a bottom-up geometric verification that attempts estimating the loop closure pose by registering putative matches
- optimize a 3D scene graph in response to loop closures, which relies on embedded deformation graphs to simultaneously correct all layers of the scene graph, from the 3D mesh, to places, objects, and rooms.

Lastly the paper proposes a highly paralllelized implementation, named Hydra,that combines fast early and mid-level perception processes (e.g., local mapping) with slower high-level perception (e.g., global optimization of the scene graph).

3.2.4 Loop Closure Detection

Agent nodes describe the robot's trajectory in the 3D scene graph. A keyframe, containing appearance information, is stored for each agent node. Using these agent nodes loop closure detection tries to find a past agent node that matches the last agent node (current robot pose).

For each agent node a hierarchy of descriptors is constructed, that describe statistics of the node's surroundings. The hierarchical descriptors include standard DBoW2 (Described in Chapter 2) appearance descriptor. The appearance descriptor is augmented with an object-based and a place-based descriptor. For loop closure detection, the place-based descriptor is examined and subsequently object-based and then appearance descriptors. If any of the descriptors return a putative result, geometric verification (Described in Paragraph 3.2.5) is performed.

3.2.5 Geometric Verification

After a putative loop closure, bottom-up geometric verification is performed. In particular, whenever there is a match (e.g.,between agent i and agent j) at a given layer, (e.g., between appearance descriptors at the agent layer, or between object descriptors at the object layer), an attempt is made to register frames i and j. For registering visual features standard RANSAC-based geometric verification as in [9] is used. If that fails, TEASER++ [10] is used, discarding loop closures that also fail object registration.

3.2.6 3D Scene Graph Optimization

the Scene Graph Backend (i) optimizes the graph using a deformation graph approach and (ii) postprocesses the results to remove redundant subgraphs corresponding to the robot visiting the same location multiple times.

- Scene Graph Frontend : The frontend builds an initial estimate of the 3D scene graph that is uncorrected for drift. More precisely, the frontend takes as input the latest mesh, places subgraph, objects, and pose graph of the agent. Then, the frontend populates inter-layer edges from each object or agent node to the nearest place node in the active window using nanoflann [11].

- Scene Graph Backend : When a loop closure is detected, the backend optimizes an embedded deformation graph built from the frontend scene graph and then reconstructs the other nodes in the scene graph via interpolation [12].

3.2.7 The Hydra Architecture

The authors the this paper implement their spatial perception engine into a highly parallelized architecture, named Hydra. Hydra involves a combination of processes that run at sensor rate (e.g., feature tracking for visual-inertial odometry), at sub-second rate (e.g., mesh and place reconstruction), and at slower rates (e.g., the scene graph optimization, whose complexity depends on the map size). Therefore these processes have to be organized such that slow-but-infrequent computation (e.g., scene graph optimization) does not get in the way of faster processes. The visualization of Hydra is presented in Figure **??**

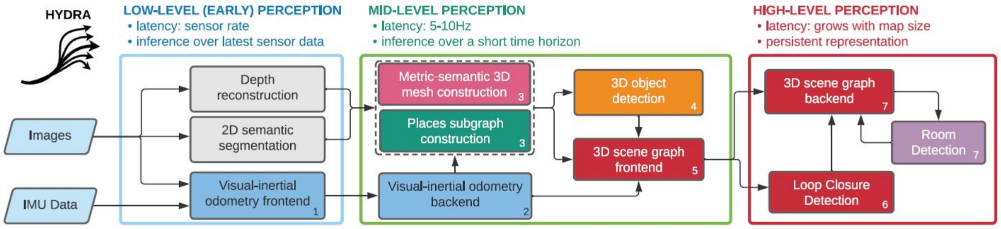

HYDRA
LOW-LEVEL (EARLY) PERCEPTION
• latency: sensor rate
• inference over latest sensor data
MID-LEVEL PERCEPTION
• latency: 5-10Hz
• inference over a short time horizon
HIGH-LEVEL PERCEPTION
• latency: grows with map size
• persistent representation
Images
IMU Data
Depth reconstruction
2D semantic segmentation
Visual-inertial odometry frontend 1
Metric-semantic 3D mesh construction 3
Places subgraph construction 3
Visual-inertial odometry backend 2
3D object detection 4
3D scene graph frontend 5
3D scene graph backend 7
Room Detection 7
Loop Closure Detection 6

Figure 3.3: Hydra's Functional Blocks

3.3 VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator

3.3.1 Abstract

One camera and one low-cost inertial measurement unit (IMU) form a monocular visual-inertial system (VINS), which is the minimum sensor suite (in size, weight, and power) for the metric six degrees-of-freedom (DOF) state estimation. In this paper, we present VINS-Mono: a robust and versatile monocular visual-inertial state estimator. Our approach starts with a robust procedure for estimator initialization. A tightly coupled, nonlinear optimization-based method is used to obtain highly accurate visual-inertial odometry by fusing preintegrated IMU measurements and feature observations. A loop detection module, in combination with our tightly coupled formulation, enables relocalization with minimum computation. We additionally perform 4- DOF pose graph optimization to enforce the global consistency. Furthermore, the proposed system can reuse a map by saving and loading it in an efficient way. The current and previous maps can be merged together by the global pose graph optimization. We validate the performance of our system on public datasets and real-world experiments and compare against other state-of-the-art algorithms. We also perform an onboard closedloop autonomous flight on the microaerial-vehicle platform and port the algorithm to an iOS-based demonstration. We highlight that the proposed work is a reliable, complete, and versatile system that is applicable for different applications that require high accuracy in localization. We open source our implementations for both PCs (`https://github.com/HKUST-Aerial-Robotics/VINS-Mono`) and iOS mobile devices (`https://github.com/HKUST-Aerial-Robotics/VINS-Mobile`) [13].

3.3.2 Issues

- Rigorous Initialization : Difficult to fuse the monocular visual structure with inertial measurements.
- VINSs are highly non-linear : challenges in terms of estimator initialization. (In most cases, the system should be launched from a known stationary posi- tion and moved slowly and carefully at the beginning, which limits its usage in practice.)
- Long-Term Drift : unavoidable for visual-inertial odometry (VIO). In order to eliminate the drift, loop detection, relocalization, and global optimization has to be developed.

3.3.3 VINS-Mono Features

- robust initialization procedure that is able to bootstrap the system from unknown initial states
- tightly coupled, optimization-based monocular VIO with camera–IMU extrinsic calibration and IMU bias correction

- online relocalization and four degrees-of-freedom (DOF) global pose graph optimization
- pose graph reuse that can save, load, and merge multiple local pose graphs

The IMUs usually acquire data at a much higher rate than the camera. Different methods have been proposed to handle the high-rate IMU measurements. The most straightforward approach is to use the IMU for state propagation in EKF-based approaches [14], [15].

In a graph optimization formulation, an efficient technique called IMU preintegration is developed in order to avoid the repeated IMU reintegration. This technique was first introduced in [16], which parameterize the rotation error using Euler angles. Shen et al. [17] derived the covariance propagation using continuous-time error-state dynamics. The preintegration theory was further improved in [18] and [19] by adding posterior IMU bias correction.

The structure of the monocular visual-inertial state estimator is shown in Figure 3.4.

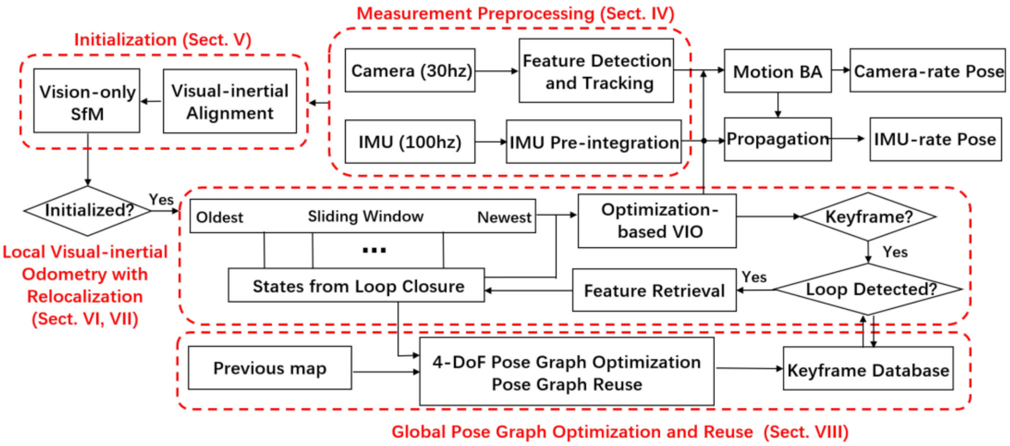


Figure 3.4: Block diagram illustrating the full pipeline of monocular VINS

Existing features are tracked by the KLT sparse optical flow algorithm [20]. Corner features are detected [21] to maintain a minimum number (100– 300) of features in each image. Outlier rejection is performed using RANSAC with a fundamental matrix model [22].

Keyframe selection is based on two criteria :

- the average parallax apart from the previous keyframe. If the average parallax of tracked features is between the current frame and the latest keyframe is beyond a certain threshold, the frame is treated as a new keyframe.
- Tracking quality : If the number of tracked features goes below a certain threshold, The frame is treated as a new keyframe.

The initialization procedure starts with a vision-only SfM to estimate a graph of up-to-scale camera poses and feature positions. First, feature correspondences between the latest frame and all previous frames are checked. The relative rotation and up-to-scale translation between these two frames is recovered using the five-point algorithm [23]. Subsequently, the scale is set arbitrarily and all features observed in these two frames are triangulated. Based on these triangulated features, a perspective-n-point (PnP) method [24] is performed to estimate poses of all other frames in the window. Finally, a global full bundle adjustment [25] is applied to minimize the total reprojection error of all feature observations.

In order to bound the computational complexity of the optimization-based VIO, marginalization is incorporated. The marginalization is carried out using the Schur complement [26].

Place Recognition using DBoW2 [27]. 500 more corners are detected and described by the BRIEF descriptor [28]. All BRIEF descriptors are kept for feature retrieving, but the raw image is discarded to reduce the memory consumption.

Benefiting from the inertial measurement of the gravity, the roll and pitch angles are fully observable in the VINS. To take full advantage of valid information and correct drift efficiently, we fix the drift-free roll and pitch, and only perform pose graph optimization in 4-DOF.

Meaning that because the gravity is known the 2 axes roll and pitch are known from the gravity measurement.

Keyframes are added into the pose graph after the VIO process. Every keyframe serves as a **vertex** in the pose graph, and it connects with other vertexes by two types of edges.

- **Sequential Edge** : A keyframe establishes several sequential edges to its previous keyframes. A sequential edge represents the relative transformation between two keyframes, which is taken directly from VIO.

- **Loop-Closure Edge** : If the keyframe has a loop connec- tion, it connects the loop-closure frame by a loop-closure edge in the pose graph.

Although the tightly coupled relocalization already helps with eliminating wrong loop closures, another **Huber norm** [29] is added to further reduce the impact of any possible wrong loops. The pose graph optimization and relocalization run asynchronously in two separate threads.

Every keyframe is a vertex in the pose graph.

3.4 SVO: Fast Semi-Direct Monocular Visual Odometry

3.4.1 Abstract

We propose a semi-direct monocular visual odometry algorithm that is precise, robust, and faster than current state-of-the-art methods. The semi-direct approach eliminates the need of costly feature extraction and robust matching techniques for motion estimation. Our algorithm operates directly on pixel intensities, which results in subpixel precision at high framerates. A probabilistic mapping method that explicitly models outlier measurements is used to estimate 3D points, which results in fewer outliers and more reliable points. Precise and high frame-rate motion estimation brings increased robustness in scenes of little, repetitive, and high-frequency texture. The algorithm is applied to micro-aerial-vehicle state estimation in GPS-denied environments and runs at 55 frames per second on the onboard embedded computer and at more than 300 frames per second on a consumer laptop. We call our approach SVO (Semi-direct Visual Odometry) and release our implementation as open-source software [30].

3.4.2 Key Points

A semi-direct VO that combines the success-factors of feature-based methods (tracking many features, parallel tracking and mapping, keyframe selection) with the accuracy and speed of direct methods.

Methods that simultaneously recover camera pose and scene structure from video can be divided into two classes:

- Feature-Based Methods : Extract a sparse set of salient image features in each image; match them in successive frames using invariant feature descriptors; robustly recover both camera motion and structure using **epipolar geometry**; finally, refine the pose and structure through **reprojection error minimization**. The disadvantage of feature-based approaches is the reliance on detection and matching thresholds, the necessity for robust estimation techniques to deal with wrong correspondences, and the fact that most feature detectors are optimized for speed rather than precision, such that drift in the motion estimate must be compensated by averaging over many feature-measurements.

- Direct Methods : Direct methods [31] estimate structure and motion directly from intensity values in the image. The local intensity gradient magnitude and direction is used in the optimisation compared to feature-based methods that consider only the distance to some feature-location. The computation of the **photometric error** is more intensive than the reprojection error, as it involves warping and integrating large image regions.

Semi-Direct Visual Odometry (SVO) algorithm uses **feature-correspondence** which is an implicit result of direct motion estimation rather than of explicit feature extraction and matching. The sparse model-based image alignment algorithm for motion estimation is related to **model-based dense image alignment** [32], [33], [34], [35]. As soon as feature correspondences and an initial estimate of the camera pose are established, the algorithm continues using only point-features; hence, the name “semi-direct”. **A Bayesian filter** that explicitly models outlier measurements is used to estimate the depth at feature locations.

3.4.3 System Overview

The algorithm uses two parallel threads (as in [16]), one for estimating the camera motion, and a second one for mapping as the environment is being explored.

The motion estimation thread implements the proposed semi-direct approach to relative-pose estimation. The first step is pose initialisation through sparse model-based image alignment: the camera pose relative to the previous frame is found through minimizing the photometric error between pixels corresponding to the projected location of the same 3D points.

The 2D coordinates corresponding to the reprojected points are refined in the next step through alignment of the corresponding feature-patches. Motion estimation concludes by refining the pose and the structure through minimizing the reprojection error introduced in the previous feature-alignment step.

In the mapping thread, a probabilistic depth-filter is initialized for each 2D feature for which the corresponding 3D point is to be estimated. New depth-filters are initialised whenever a new keyframe is selected in regions of the image where few 3D-to-2D correspondences are found. The filters are initialised with a large uncertainty in depth. At every subsequent frame the depth estimate is updated in a Bayesian fashion (see Figure 5). When a depth filter's uncertainty becomes small enough, a new 3D point is inserted in the map and is immediately used for motion estimation.

3.5 Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping

3.5.1 Abstract

We provide an open-source C++ library for realtime metric-semantic visual-inertial Simultaneous Localization And Mapping (SLAM). The library goes beyond existing visual and visual-inertial SLAM libraries (e.g., ORB-SLAM, VINSMono, OKVIS, ROVIO) by enabling mesh reconstruction and semantic labeling in 3D. Kimera is designed with modularity in mind and has four key components: a visual-inertial odometry (VIO) module for fast and accurate state estimation, a robust pose graph optimizer for global trajectory estimation, a lightweight 3D mesher module for fast mesh reconstruction, and a dense 3D metric-semantic reconstruction module. The modules can be run in isolation or in combination, hence Kimera can easily fall back to a state-of-the-art VIO or a full SLAM system. Kimera runs in real-time on a CPU and produces a 3D metric-semantic mesh from semantically labeled images, which can be obtained by modern deep learning methods. We hope that the flexibility, computational efficiency, robustness, and accuracy afforded by Kimera will build a solid basis for future metric-semantic SLAM and perception research, and will allow researchers across multiple areas (e.g., VIO, SLAM, 3D reconstruction, segmentation) to benchmark and prototype their own efforts without having to start from scratch [36].

3.5.2 Key Points

- **Kimera-VIO** : a VIO module for fast and accurate IMU-rate state estimation. At its core, Kimera-VIO features a GTSAM-based VIO approach [44], using IMUpreintegration and structureless vision factors [45], and achieves top performance on the EuRoC dataset [19];

- **Kimera-RPGO** : a robust pose graph optimization (RPGO) method that capitalizes on modern techniques for outlier rejection [46]. Kimera-RPGO adds a robustness layer that avoids SLAM failures due to perceptual aliasing, and relieves the user from time-consuming parameter tuning;

- **Kimera-Mesher** : a module that computes a fast per-frame and multi-frame regularized 3D mesh to support obstacle avoidance. The mesher builds on previous algorithms by the authors and other groups [42], [47]–[49];

- **Kimera-Semantics** : a module that builds a slower-butmore- accurate global 3D mesh using a volumetric approach [27], and semantically annotates the 3D mesh using 2D pixel-wise semantic segmentation.

Kimera takes stereo frames and high-rate inertial measurements as input and returns (i) a highly accurate state estimate at IMU rate, (ii) a globally-consistent trajectory estimate, and (iii) multiple meshes of the environment, including a fast local mesh and a global semantically annotated mesh.

The vision front-end detects Shi-Tomasi corners [21], tracks them across frames using the Lukas-Kanade tracker [37], finds left-right stereo matches, and performs geometric verification . We perform both mono(cular) verification using 5-point RANSAC [38] and stereo verification using 3-point RANSAC [39]; the code also offers the option to use the IMU rotation and perform mono and stereo verification using 2-point [40] and 1-point RANSAC, respectively. Feature detection, stereo matching, and geometric verification are executed at each keyframe, while we only track features at intermediate frames.

VIO Back-end : At each keyframe, preintegrated IMU and visual measurements are added to a fixed-lag smoother (a factor graph) which constitutes our VIO back-end. We use the preintegrated IMU model and the structureless vision model of [41]. The factor graph is solved using iSAM2

[42] in GTSAM [43]. At each iSAM2 iteration, the structureless vision model estimates the 3D position of the observed features using DLT [44] and analytically eliminates the corresponding 3D points from the VIO state [45]. Before elimination, degenerate points (i.e., points behind the camera or without enough parallax for triangulation) and outliers (i.e., points with large reprojection error) are removed, providing an extra robustness layer. Finally, states that fall out of the smoothing horizon are marginalized out using GTSAM.

Robust PGO : This module is implemented in GTSAM, and includes a modern outlier rejection method, Incremental Consistent Measurement Set Maximization (PCM) [46], that we tailor to a single-robot and online setup. We store separately the odometry edges (produced by Kimera-VIO) and the loop closures (produced by the loop closure detection); each time the PGO is executed, we first select the largest set of consistent loop closures using a modified version of PCM,and then execute GTSAM on the pose graph including the odometry and the consistent loop closures.

3.6 A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors

3.6.1 Abstract

4 Feature Detection

Feature detection can be divided into four categories namely edge, corner, blobs and ridges.

4.1 Interest Point Detection

The interest points detection technique has been recognized as a popular method in visual SLAM (vSLAM) since many literatures has been published using this techniques. Point features have a clear mathematical definition with well defined position in image area. The high local information content has simplified further processing in the vision system. It is not susceptible to disturbance such as deformation (i.e., orientation or scale changes). Corner and blob detection is classified under interest point detection since it has those properties. The difference between these two techniques is only substantial when image is small. Corner looks for sharp image features while blobs look for smooth image features. Blob detectors compliment corner detectors by detecting regions that are too smooth. Harris detection, Shi-Tomasi, SIFT, SURF and FAST are popular method categorized in interest point detection technique. Some detection though is a combination of edge and interest point detection.

4.1.1 Harris detection

Harris and Stephens (1988) has proposed a combined technique of corner and edge detector to cater image regions with texture and isolated feature by improving Moravac's corner detector (Frintrop et al., 2007). According to their research (Harris and Stephens, 1988), for an explicit tracking of image features, the image features must be discrete and not from continuum like texture or edge pixels. This is because curved lines and texture edges can be fragment differently. They used the local autocorrelation function of a signal to measure the local changes of the signal with patches shifted by a small amount in different directions.

`https://scialert.net/fulltext/?doi=itj.2009.250.262`

5 Ideas

- Use 2 zed cameras one on head one on belly pointing the same way (distance known from each other), after feature matching on both compare the save features to have an even more accurate position of the pointcloud.

- Create a Voronoi diagram of the features in each room and recognise places when robot finds features that match the Voronoi diagram (Voronoi Diagram : on a 3D surface with points the Voronoi Diagram computes the least distance of all points creates a diagram of shapes that represent the least distance of each point).

- Euclidean Clustering on sets of features to create layers of objects.

- Relative probability of distance from feature matching according to 40 times baseline paper. according to the distance of the feature take into account how accurate it is, like close far points.

- IMU state propagation (IMU higher rate than camera), IMU preintegration (to avoid the repeated IMU reintegration) (Paragraph 3.3.3).
- IMU Noise and Bias equations are shown in [13] page 4.
- Choose sufficient number of features through research and probability. (papers etc.).
- extrinsic parameters (camera to imu) are shown in [13] page 5.
- check the gravity. It always has value of 9.81 but can be on any axis due to camera movement and it should maove the location of the object at most times. It is addressed in [13] page 5.
- create equations for the translation and rotation of the robot. Some equations are presented in [13] page 6.
- account for different types of cameras [13] page 7.
- Apart from features, shapes and lines can be detected.
- improve map using more computation which runs in another thread. (map updates on Xhz, map improvement runs on ¡Xhz)
- Good outlier detection (ακραίες τιμές)
- floor/wall detection
- ROS provides the ability to be sure about the timestamp of the topic you subscribe, so create a msg with 2 cv::Mat Images to be sure that both have the same timestamp.
- use time synchronizer (ROS) to sync both camera messages.
- **Feature Matching by taking into account IMU data (quicker processing-wise)**

To be researched :

- nonlinear least squares optimization.
- sliding widow.

- PTAM [47] is a feature-based SLAM algorithm that achieves robustness through tracking and mapping many (hundreds) of features.
- SVO PAPER page 4 : This alignment is solved using the inverse compositional Lucas-Kanade algorithm
- SVO PAPER page 4 : Since Equation (7) we solve it in an iterative Gauss-Newton procedure.
- SVO PAPER page 4 : This is the well known problem of motion-only BA [17] and can efficiently be solved using an iterative non-linear least squares minimization algorithm such as Gauss Newton.
- SVO PAPER page 5 : The recursive Bayesian update step for this model is described in detail in [48].
- for loop closure detection DBoW2.

Feature Descriptors :

- we care for scalability and rotation
- change descriptors according to imu data (if imu says there is movement forward, scale the descriptors accordingly(more pixels) to match more keypoints)

Feature Matching:

- Feature Matching by taking into account IMU data (quicker processing-wise)

6 TO DO

- Create Feature Descriptors
- Feature Matching
- Sort Matches by close-far and take them into account accordingly
- Integrate IMU messages to help localization
- Research Loop Detection

- Integrate Loop Detection
- Map optimization (if loop is detected update map correctly)
- Row Detection Using Deep Learing
- Create train-test Datasets

List of Figures

3.1 ORB-SLAM2 Diagram of Parallel Threads 6
3.2 Map Dilation for Room Detection 10
3.3 Hydra's Functional Blocks 12
3.4 Block diagram illustrating the full pipeline of monocular VINS 14

References

[1] Mark Maimone, Yang Cheng, and Larry Matthies. Two years of visual odometry on the mars exploration rovers. *Journal of Field Robotics, Special Issue on Space Robotics*, 24:2007, 2007.

[2] Raúl Mur-Artal and Juan D. Tardós. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017.

[3] Lina M. Paz, Pedro PiniÉs, Juan D. TardÓs, and JosÉ Neira. Large-scale 6-dof slam with stereo-in-hand. *IEEE Transactions on Robotics*, 24(5):946–957, 2008.

[4] Rainer Kümmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. G2o: A general framework for graph optimization. pages 3607 – 3613, 06 2011.

[5] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception engine for 3d scene graph construction and optimization, 2022.

[6] Helen Oleynikova, Zachary Taylor, Marius Fehr, Roland Siegwart, and Juan Nieto. Voxblox: Incremental 3d euclidean signed distance fields for on-board MAV planning. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, sep 2017.

[7] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans, 2020.

[8] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021.

[9] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping, 2019.

[10] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. 2020.

[11] Jose Luis Blanco and Pranjal Kumar Rai. nanoflann: a C++ header-only fork of FLANN, a library for nearest neighbor (NN) with kd-trees. `https://github.com/jlblancoc/nanoflann`, 2014.

[12] Robert Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 26, 07 2007.

[13] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[14] Stephan Weiss, Markus W. Achtelik, Simon Lynen, Margarita Chli, and Roland Siegwart. Real-time onboard visual-inertial state estimation and self-calibration of mavs in unknown environments. In *2012 IEEE International Conference on Robotics and Automation*, pages 957–964, 2012.

[15] Anastasios I. Mourikis and Stergios I. Roumeliotis. A multi-state constraint kalman filter for vision-aided inertial navigation. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3565–3572, 2007.

[16] Todd Lupton and Salah Sukkarieh. Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions. *IEEE Transactions on Robotics*, 28(1):61–76, 2012.

[17] Shaojie Shen, Nathan Michael, and Vijay Kumar. Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs. *Proceedings - IEEE International Conference on Robotics and Automation*, 2015:5303–5310, 06 2015.

[18] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. On-manifold preintegration for real-time visual–inertial odometry. *IEEE Transactions on Robotics*, 33(1):1–21, feb 2017.

[19] Christian Forster, Luca Carlone, Frank Dellaert, and Davide Scaramuzza. Imu preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation. 07 2015.

[20] Bruce Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). volume 81, 04 1981.

[21] Jianbo Shi and Tomasi. Good features to track. In *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.

[22] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[23] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.

[24] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *International Journal of Computer Vision*, 81, 02 2009.

[25] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Workshop on Vision Algorithms*, 1999.

[26] Gabe Sibley, Larry H. Matthies, and Gaurav S. Sukhatme. Sliding window filter with application to planetary landing. *Journal of Field Robotics*, 27:587–608, 2010.

[27] Dorian Galvez-López and Juan D. Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012.

[28] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. volume 6314, pages 778–792, 09 2010.

[29] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73 – 101, 1964.

[30] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast semi-direct monocular visual odometry. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[31] Michal Irani and P. Anandan. All about direct methods. 1999.

[32] A.I. Comport, E. Malis, and P. Rives. Real-time quadrifocal visual odometry. *The International Journal of Robotics Research*, 29(2-3):245–266, 2010.

[33] Tommy Tykkala, Cedric Audras, and Andrew I. Comport. Direct Iterative Closest Point for Real-time Visual Odometry. In *The Second international Workshop on Computer Vision in Vehicle Technology: From Earth to Mars in conjunction with the International Conference on Computer Vision*, Barcelona, Spain, 2011.

[34] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Robust odometry estimation for rgb-d cameras. In *2013 IEEE International Conference on Robotics and Automation*, pages 3748–3754, 2013.

[35] Selim Benhimane and Ezio Malis. Integration of euclidean constraints in template based visual tracking of piecewise-planar scenes. pages 1218 – 1223, 11 2006.

[36] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2020.

[37] Jean yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.

[38] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):756–770, 2004.

[39] Berthold Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society A*, 4:629–642, 04 1987.

[40] Laurent Kneip, Margarita Chli, and Roland Siegwart. Robust real-time visual odometry with a single camera and an imu. 08 2011.

[41] Jinyao Zhu, Chao Yao, and Klaus Janschek. Stereo visual-inertial fusion for uav state estimation. *IFAC-PapersOnLine*, 53(2):9420–9425, 2020. 21st IFAC World Congress.

[42] Michael Kaess, Hordur Johannsson, Richard Roberts, Viorela Ila, John Leonard, and Frank Dellaert. isam2: Incremental smoothing and mapping using the bayes tree. *International Journal of Robotic Research - IJRR*, 31:216–235, 05 2012.

[43] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. 2012.

[44] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2004.

[45] Luca Carlone, Zsolt Kira, Chris Beall, Vadim Indelman, and Frank Dellaert. Eliminating conditionally independent sets in factor graphs: A unifying perspective based on smart factors.

[46] R. M. Eustice J. G. Mangelson, D. Dominic and R. Vasudevan. Pairwise consistent measurement set maximization for robust multirobot map merging. *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 2916–2923, 2018.

[47] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 225–234, 2007.

[48] George Vogiatzis and Carlos Hernandez. Withdrawn: Video-based, realtime multi-view stereo. *Image and Vision Computing*, 08 2012.