Visual-Inertial Odometry based SLAM for Drones Survey

Timothy J. Wroge
Swanson School of Engineering
University of Pittsburgh
Pittsburgh, PA, USA
timothy.wroge@pitt.edu

Solomon A. Heisey
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
soh22@pitt.edu

***Abstract*—State estimation is a challenging problem. The problem is compounded since the sensors involved in the production of this system may be expensive and heavy. Since the weight of a traditional unmanned air vehicle (UAV) is tightly constrained, minimizing the weight of that payload can help tremendously. Monocular Visual Inertial (VINS-Mono) solves these engineering challenges by generating a solution to the state estimation problem using only an inertial measurement unit (IMU) and one monocular camera. The system is able to accurately generate poses of robotic systems though only visual and internal inputs. This provides a great leap in performance for UAV's due to these potential improvements in weight and complexity of the sensor payload.**

I. Introduction

VINS-Mono [1] [2] is a method of simultaneous localization and mapping (SLAM) that incorporates the sensor input from inertial inputs (through an IMU) and visual inputs (a monocular camera). These inputs are integrated together to generate an accurate pose of the robotic system. This is accomplished through a complex state model that involves a Kalman Filter which integrates these sensor modalities to a coherent state. This is a state estimation model, where the goal is to generate the state and differentials of state from one time point to the next.

If $x(t)$ is defined as a the state of the system at a certain time (e.g. x, y, z, roll, pitch, yaw):

$$\begin{bmatrix} x(t) \\ \dot{x}(t) \\ \ddot{x}(t) \\ \vdots \end{bmatrix}$$

the goal of a state estimator is to generate the best model of the state of a system given the previous state of a system. Since traditional mechanics of robotic bodies are tightly linked to external forces, the external forces have to be factored in which increases the complexity of the system. The Kalman Filter solves this problem by providing a optimal state estimator. The Kalman Filter is optimal in that it minimizes the mean squared error of a signal in the presence of Gaussian noise.

A traditional use of a Kalman Filter usually requires sensors that indicate position (such as a GPS), velocity (encoders for wheels) and acceleration (IMU). Calibrating and using these systems can be both expensive and heavy for UAVs. VINS-mono offers a solution to these issues by relying solely on the presence of a monocular camera and an IMU.

In the past, visual based state estimation typically required binocular cameras, which are expensive, difficult to calibrate, and large. Monocular cameras, on the other hand, are lightweight, inexpensive, and have little to no hardware setup issues. The current state of the art of monocular based state estimation and simultaneous localization and mapping (SLAM) is currently limited from widespread adoption because current systems do not allow for resolving scale in an image. By incorporating an IMU in combination with a camera, VINS-Mono is able to fuse this information to get an accurate estimation on a vehicle's state and size in the world.

Fusing inertial information with camera sensor input poses some key challenges. The previous VINS systems all required slow movement in order to properly localize. Since the localization problem is highly nonlinear, this slow speed is required to track points and features in the environment, there are some critical challenges to use traditional linear state estimation techniques in practice.

VINS-mono is a large advancement compared to previous VINS based techniques since it has key optimizations to flaws in earlier systems. The the system uses a sophisticated initialization feature, so it is able to localize itself even from unknown starting locations in a map. One key downside to other systems that used traditional Kalman filtering techniques were that they depended on a internal bias (some small acceleration bias in one direction or another), the robot may correct its pose and integrate its uncertainty leading to very inaccurate estimations of the robots position. To avoid this, the VINS-mono system uses bias correction and information that is derived from the extremities of the robot to fix these accumulating errors. In addition, the system is also able to accurately localize its position in a mapped out environment based on the geometry of the scene.

II. Related Work

Visual based simultaneous localization and mapping (SLAM) has slowly been optimized over previous years and now rivals the performance of LIDAR based mapping techniques. Previous techniques for visual-based SLAM focused primarily on sensor fusion through extended Kalman filtering. An extended Kalman filter (sometimes referred to as an EKF) is a form of the traditional Kalman filter where the functions needed to predict the next state need not be linear, only differentiable. This provides some major advancements over traditional Kalman filtering because it allows for straight forward geometric estimation. A rotation of a robot or vehicle can be approximated as some rotation of the feature space around the vehicle. The features can be approximated as traveling along some sphere where, the model is able to generate an accurate estimation of feature trajectories and of a robot's position. The visual feature input can allow for another constraint of the system which allows it to be neatly optimized. Such models include work by Mourikis et al [3] and Bloesch et al [4]. One way to optimize the model is to take results for a few time steps and work to sort out signal from noise through graphical approaches. These, however, typically require a significant amount of online processing which precludes this technique from being used on embedded and low computation devices.

Other techniques require information about the odometry of the robots. These techniques include encoders which measure speed of the robotic platforms as well as a number of other key features. These approaches also have serious issues with long term drift for position (x, y, z) and rotation (roll, pitch, yaw) information about the robotic platform. The main ways to resolve this issue is through loop closure. Some main optimizations have been found to alleviate the issue posed by this problem. Most notably, work by Newman and Ho [5] solved this through a visual array of feature matching from the environment. Loop closure has great implications for slam performance since it can vastly reduce the constraints posed by strange path plans in a region and greatly reduces drift issues.

III. Method

VINS-Mono is fundamentally built around a state graph which addresses the limitations of other inertial and visual-based SLAM systems. The system begins with a measurement step. This step takes the sensory measurements such as camera frames and IMU inputs and generates optical feature tracking as well as IMU pre-integration. These go on to produce a IMU pose and a camera pose. The robotic system then works to generate an accurate visual-inertial alignment which feeds into the Structure from Motion (SfM) step. After this phase is complete, the robotic platform is then said to be "initialized." The system then generates a sophisticated graph optimization and map reuse decision which take advantage of previous states that the robot may have been in as well as storing key visual information.

A. Measurement Preprocessing

The measurement prepossessing step takes advantage of the visual frame information. The visual frame information is captured in at a certain frame rate and the system extracts visual features such as scale invariant feature transform (SIFT) proposed by David Lowe [6] or KAZE Features proposed by Alcantarilla et al. [7]. During the first frame all new optical features are extracted. At the next time step, the system extracts another set of optical features and matches these features against the previous frame. The system then extracts new features for this frame.

This can be generalized through the equation shown in equation 1 where F_t represents the optical features for a frame at time t and M_t are the matches from the features from F_{t-1} and the the current features. $\vec{N_t}$ represents the new features that are not matches from the previous frame. A local pose homography is computed with the matched features M_t so that a transformation of the camera can be approximated for the pose of the robot.

$$\vec{F_t} = \begin{bmatrix} \vec{M_t} \\ \vec{N_t} \end{bmatrix} \tag{1}$$

In order to avoid outliers adding noise in the estimation of the local homography, the VINS-Mono model uses RANSAC [8] to filter out outliers and generate an accurate estimate of the true camera rotation and translation (pose).

The inertial preprocessing uses "preintegration" to generate smoothed IMU pose. This approach requires a noise model and an IMU offset. The acceleration is modeled as a time varying quantity with a random noise component. Every IMU has a defined gyroscopic bias and acceleration bias which needs to be known in order to not cause uncertainty of the robot's positon over time. These errors can accumulate since the acceleration is double integrated to generate an estimate of position. VINS-Mono first uses preintegration which is a technique described in [9]. After the preintegration step, the acceleration and gyroscopic bias is found and corrected for using the values derived from the preintegration step.

B. Visual Inertial Initialization procedure

After the measurement preprocessing step, the system needs to work out a vision based structure from motion (SfM). This SfM generates an estimation of the robot's relative position with visually recognized features (denoted as F_t in equation 1). This system requires that the matched features from the previous frames (M_t from equation 1) are sufficiently consistent and have enough deviation in pixel positions to generate a estimation of pose through parallax. If these criteria are satisfied, the rotation and translation of the robot can be resolved.

In order to find the camera transform (rotation and position of the camera with respect to observed points) the detected image points (c_x, c_y) are transformed into a 3D pose (x, y, z) through solving the perspective-n-point problem (PnP). Since this step is computed over a window, there are a series of these perspective points. These poses can be corrected to generate the smallest re-projection error (the inverse process, translating the 3D pose into the image points) through a process called bundle adjustment. After this process, the IMU output from pre-integration is fused with the output of the visual feature tracking pose. This allows the scale and relative position of

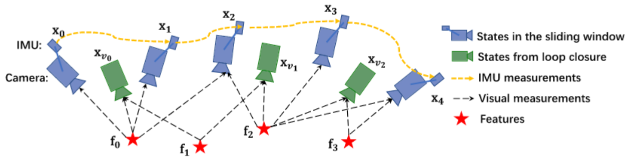

Fig. 1. Visual feature and inertial measurement integration from [2].

the robot to be resolved. Qin et al [2] call this process visual inertial alignment. This concept is illustrated in figure 1.

Once this visual feature tracking procedure is finished, the system is initialized and can enter into the graph optimization and mapping reuse state.

C. Local VINS Odometry

After the initialization procedure, the system must generate a notion of visual odometry that is integrated with the IMU sensor payload for an estimate of the robot's state. The model then takes a sliding window of camera frames and incorporates the inertial measurements such that the maximum likelihood solution to the state of the system is estimated. This is done by considering the noise model for the IMU measurements and visual inputs as a Gaussian distribution. The residual error of the state estimate can then be fit of the to this Gaussian distribution. Qin et al [2] offer two versions of the error residuals for the visual re-projection error and the inertial state estimate. These residual definitions offer close form solutions for these problems such that they can be optimized and minimized.

D. Graph Optimization and Map Reuse

This section focuses on how previously built maps can generate accurate external validation for the pose output from the measurement step and initialization steps. A visual feature map is generated when the robot takes a path through the world with locally similar features. The system can then take advantage of comparisons of the different visual maps and can combine these maps as shown in figure 2.

E. Review

In conclusion, the system begins with a global measurement step which collects the monocular camera frames as well as the inertial measurements from the IMU. The information collected is then smoothed with Gaussian noise models and passed to the visual inertial initialization step. This step generates camera-based positioning and estimates the pose of the robotic platform. After the system is initialized, then the system takes a sliding window of camera frames and conducts visual inertial odometry to minimize the global measurement error of the constructed pose. Finally, the system incorporates information from previous runs to minimize the state error of the system.

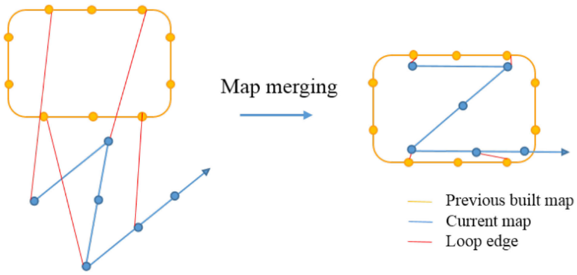

Fig. 2. Example of Map Merging for VINS Mono from [2].

IV. Discussion

The VINS Mono system is a sophisticated algorithm that offers significant advancements for state of the art robotic simultaneous localization and mapping systems. The main advantages of the technique include the simplicity of the sensors, the internal calibration and its online processing potential. Because the system is built around two relatively inexpensive sensors: a monocular camera and an inertial measurement unit, the system is orders of magnitude less expensive than other options that require LIDAR sensors or expensive GPS and Real Time Kinematic (RTK). In addition, since cameras and IMUs do not tend to be have a significant weight, the system does not introduce a payload risk for unmanned aerial vehicles.

Since VINS-Mono also produces a self initialization procedure, the system is able to localize the UAV or robot in any position with rich visual features. In addition, since the system has an internal noise model and bias offset for the IMU, the system is able to account for issues with unreliable sensors. This greatly reduces issues with localization and finds an optimal smoothing that accounts for this accumulation of noise.

Since, the system is able to fuse the visual features with inertial measurements, the system yields an increase in performance for each of these sensory modalities compared to if they were used independently. Also, since the model is capable of using a purely visual system with a monocular camera, it shows the movement of the robotic research toward the use of inexpensive and lightweight monocular cameras with minimal inertial sensors.

Further, since the system is open-source, the system can be readily upgraded and improved by the greater research community. Currently researchers from all over the world have already generated potential improvements to the algorithm and have integrated it into the current state estimators. This gives hope for future advancements in visual inertial SLAM based systems.

In addition to these great improvements over the current state of the art in robotics, the system is able to outperform other Visual Inertial systems as per figure 3 and figure 4 such as VINS and VINS_loop.

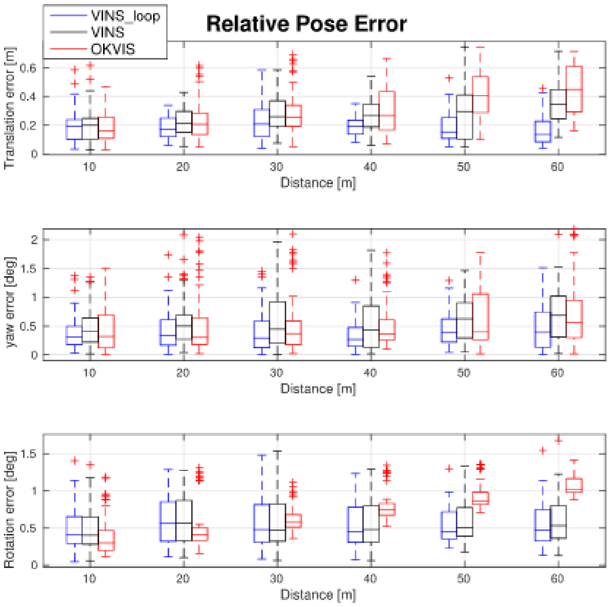

Fig. 3. Relative Pose Error from [2] based on the EuRoC MH_03_medium dataset

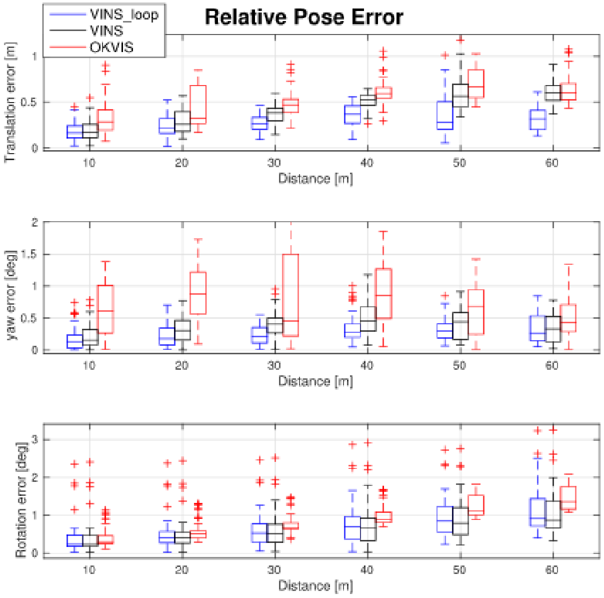

Fig. 4. Relative Pose Error from [2] based on the EuRoC MH_05_difficult dataset

Fig
loc
h 7
10

V. Conclusion

The Monocular Visual Inertial System (VINS-Mono) shows remarkable improvements for the state of the art in robotic state estimation. Due to the complicated nature of traditional state estimation and the simplicity of these sensors, the system has potential use cases including robotics, navigation, and artificial reality. In addition to all of these potential use cases, the system is able to work with a remarkably low error rate compared to other state of the art visual based SLAM algorithms. In addition, the performance of this system can be used as a model for other research in robotics and applied mathematics because of its elegant noise model and sensor smoothing methods. There are still a number of improvements that can be made to the system but this is a positive direction for the robotics research community. Due to the performance of VINS-Mono coupled with the cheap hardware and software costs, the system rivals other, more expensive sensor systems. In spite of its advantages, there will be much more research to be done to improve this system in the future.

References

[1] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, “Orb-slam: a versatile and accurate monocular slam system,” *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[2] T. Qin, P. Li, and S. Shen, “Vins-mono: A robust and versatile monocular visual-inertial state estimator,” *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[3] A. I. Mourikis and S. I. Roumeliotis, “A multi-state constraint kalman filter for vision-aided inertial navigation,” in *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pp. 3565–3572, IEEE, 2007.

[4] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, “Robust visual inertial odometry using a direct ekf-based approach,” in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 298–304, IEEE, 2015.

[5] P. Newman and K. Ho, “Slam-loop closing with visually salient features,” in *proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pp. 635–642, IEEE, 2005.

[6] D. G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, pp. 1150–1157, Ieee, 1999.

[7] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, “Kaze features,” in *European Conference on Computer Vision*, pp. 214–227, Springer, 2012.

[8] O. Chum, J. Matas, and J. Kittler, “Locally optimized ransac,” in *Joint Pattern Recognition Symposium*, pp. 236–243, Springer, 2003.

[9] S. Shen, N. Michael, and V. Kumar, “Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft mavs,” in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5303–5310, IEEE, 2015.