

Multi Modal Pose Fusion for Monocular Flight with Unmanned Aerial Vehicles

Arbaaz Khan
GRASP Laboratory
University of Pennsylvania
Philadelphia, PA 19104
(412)-961-3258
arbaazk@seas.upenn.edu

Martial Hebert
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
(412)-268-5704
mhebert@andrew.cmu.edu

Abstract—pose estimation using a monocular camera is an area of active research in visual inertial unmanned aerial vehicles. Some of the common methods involve fusing pose from two sources using a kalman filter or a lie algebraic representation for solving the global motion constraints. Some systems even couple inertial systems (IMU) with the visual pose estimation from the camera for robust pose estimation. This work extends the idea of a tightly coupled visual inertial system to a photometric error based method of semi dense visual odometry. The estimated pose from the camera is prone to error spikes due to loss in tracking keyframes. With extensive experimentation we prove that our visual inertial system for semi dense visual odometry is better than using visual odometry alone. We also show robustness of our method and compare our performance to the existing state of the art tightly coupled visual inertial systems that exist in an outdoor environment.

TABLE OF CONTENTS

1. INTRODUCTION.....	1
2. RELATED WORK	2
3. SYSTEM OVERVIEW	2
4. POSE ESTIMATION	2
5. IMU PRE-INTEGRATION	4
6. EXPERIMENTS	5
7. CONCLUSION	6
ACKNOWLEDGMENTS	6
REFERENCES	6
BIOGRAPHY	7

1. INTRODUCTION

Achieving autonomous flight for unmanned aerial vehicles (UAV) with only visual sensors is a field that has been explored extensively in research in both Robotics and computer vision [2, 4, 5]. Some key problems in such autonomous flight systems are: 1) the difficulty associated with sensing the unknown environment, 2) avoiding obstacles in the flight path and 3) reaching a pre assigned goal position in an optimal manner. Flying an autonomous UAV can be broken down into several steps. The first step involves perception to build a representation of the UAV's environment. The next step is to localize the UAV's current position in its environment. This problem is called Simultaneous Localization and Mapping and has been explored in depth in robotics [3]. Once such a representation has been constructed, the current state of the UAV can be estimated. The parameters in the state representation are pose, linear velocities and angular velocities. These are then used to compute an optimal trajectory which the UAV executes to reach its goal location.

In this work, the idea of coupling the perception and the pose estimation systems for a unmanned aerial vehicle with an Inertial Measurement Unit (IMU) is explored. In the context of UAV flight, it is commonplace to use LIDARs [9, 10] and stereo cameras to build feature rich representations of the environment such as depth maps and 3D point clouds. Instead of this stereo approach, we propose the use a monocular camera to build a similar feature rich representation of the environment and localize the UAV's position in this representation. Using a monocular camera reduces the cost of building up the system since it is reasonably cheap to buy off the shelf monocular cameras as compared to LIDARs and stereo cameras. In addition it also reduces the computation complexity associated with feature mapping for generating depth maps from stereo images. The monocular method to estimate depth builds on previous work in computer vision [7, 8] where the intensity between keyframes is tracked to build a depth map of the environment. The monocular system uses the semi dense visual odometry (SDVO) method [8] (described in detail in Section 4) to track points of interest from key frame to key frame. Naturally, this tracking is prone to losses in tracking if the keyframes change drastically. Within the context of an unmanned aerial vehicle flying in an outdoor environment, such losses in tracking are fairly common due to fast rotations or disturbances from external conditions such as wind. If one were to estimate the state of the UAV from the SDVO, such losses would result in catastrophic failures in planning optimal trajectories. In addition, the mismatch between the UAV's estimated position and actual position would grow over time leading to large drift.

Since the pose estimation from the SDVO is highly unreliable, a downward facing camera for estimating pose by computing optical flow [13] is used. Using optical flow on a downward facing camera tends to be tricky. Previous studies on the performance of the optical flow algorithm for different textures and scenarios have shown the performance to vary hugely under varied scenarios [14]. If the texture of the ground is varied, results obtained are generally very close to ground truth. But if the texture on the ground is featureless as is the case for a lot of forest environments, state estimates from optical flow are noisy. Similarly, directly using a IMU to estimate position and velocities by integrating over accelerations tends to drift over time. Thus, ideally we would like to use the front facing camera for depth estimation and additionally pose estimation. Given that these visual odometry methods generate unreliable pose estimates, we formulate our problem as follows :

Problem Statement: *Given a semi dense method of generating visual odometry and estimating depth, an optical flow setup for estimating pose and an IMU, find a optimal data fusion paradigm that produces high quality depth maps and*

robust pose estimates.

To solve this, we propose a novel mechanism to couple the pose estimation from the SDVO to the IMU. This serves two advantages; it reduces the propagated errors in pose estimation from keyframe to keyframe since the pose is now injected into the SDVO loop from the IMU at every keyframe. Additionally, since the depth map construction also depends on the current pose, the depth maps are more robust to losses in tracking.

Thus, the key contributions of this paper are as follows :

1. A novel mechanism to inject pose from the IMU into a monocular depth estimation process by pre integration on lie manifolds.
2. Extensive experimentation on an unmanned aerial vehicle in a GPS denied outdoor setting with dense clutter.

We test our proposed algorithm on a UAV platform in an outdoor setting with high density of clutter and analyze its performance both qualitatively and quantitatively.

2. RELATED WORK

In the past, ideas from Structure from Motion [22, 23] have been used to reconstruct 3D scene geometry from a monocular camera under motion. The representation developed by such methods however is sparse in nature and unsuitable for high speed flight in the presence of large amounts of clutter. Using stereo cameras for UAV flight has been explored in [24]. While stereo methods are capable of generating high quality depth maps and robust pose estimations, they often come at the price of increased power costs and computational complexity. In recent years there has been a focus on looking at learning based methods for estimating depth and pose for the flight platform [20, 21]. Another idea that has been pursued in the fields of machine learning is to directly map input state to control outputs thus foregoing, perception, pose estimation and planning [25]. However, these methods are highly susceptible to training data and do not generalize well to unseen data making them unsuitable for exploration tasks. These are in contrast to our method where we use a monocular camera that needs minimum amount of power increasing the flight time of the UAV and offers high dimensional information about the environment with low latency. Our method is able to generate robust pose estimates without compromising on the quality of depth maps generated.

In the next sections, we describe our UAV platform, preliminary experiments to motivate our method, our proposed method for pose estimation and finally, experimental verification of our method.

3. SYSTEM OVERVIEW

A modified version of the 3DR ArduCopter with an onboard quad-core ARM processor and a Microstrain 3DM-GX3-25 IMU is used. Onboard, there are two monocular cameras: one PlayStation Eye camera facing downward for real time pose estimation and one high-dynamic range PointGrey Chameleon camera for monocular navigation. The distributed processing framework is shown in Figure 1. An image stream from the front facing camera is streamed to the base station where the perception module produces a local 3D scene map; the planning module uses these maps to find the best

trajectory to follow and transmits it back to the onboard computer where the control module does trajectory tracking. The quadrotor is flown in outdoor forest environments cluttered with trees (Figure 7). The forest cover prevents us from relying on GPS for localization. Our UAV platform is shown in Figure 2.

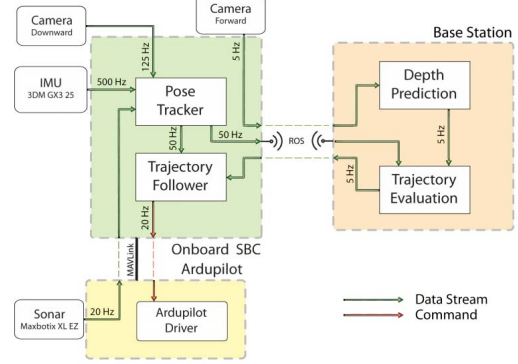


Figure 1: System Overview

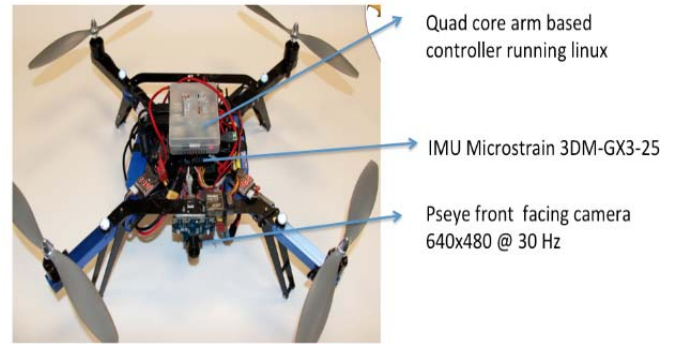


Figure 2: UAV platform

4. POSE ESTIMATION

In this section our approach for pose estimation and the problems that arise from our approach are described. Some of the standard practices for fusing pose estimation from two sources are applied and it is shown how these do not improve the overall results by a significant amount. Using this as motivation, a system for robust pose estimation using the on board inertial measurement unit while at the same time providing accurate semi dense depth maps is developed.

Pose Estimation - KLT Optical Flow

In our current system a downward facing camera for pose estimation that uses the famous Kanade-Lucas-Tomasi (KLT) optical flow (flow tracker) algorithm [13] is employed. Consider a pixel $I(x, y, t)$ where the t dimension represents time t . Let the motion between two frames be dx, dy in time dt . Since the pixel doesn't change itself, $I(x, y, t) = I(x + dx, y + dy, t + dt)$. This can be approximated by the Taylor series and with some simple mathematical manipulation it is easy to see:

$$I(x, y, t) = f_x(u) + f_y(v) + f_t = 0 \quad (1)$$

where

$$f_x = \frac{\partial f}{\partial x}; f_y = \frac{\partial f}{\partial y}u = \frac{dx}{dt}; v = \frac{dy}{dt} \quad (2)$$

f_x and f_y are easy to estimate since they are gradients of the image along the x and y axes. u and v are the unknowns we look to solve for. The KL tracker uses the assumption that the neighboring patch of 3×3 pixels around the pixel of interest moves at the same velocity as the pixel of interest. Instead of solving an over-determined system for 9 pixels, KLT tracker gives the velocity by using the least square fit method. This can be solved as :

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_i f_{x_i}^2 & \sum_i f_{x_i} f_{y_i} \\ \sum_i f_{x_i} f_{y_i} & \sum_i f_{y_i}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_i f_{x_i} f_{t_i} \\ -\sum_i f_{y_i} f_{t_i} \end{bmatrix} \quad (3)$$

Equation (3) is solved for each pixel in the images obtained from the downward facing camera and average over them to get the velocity of the UAV. It is clear that the first limitation of the flow tracker algorithm is that it is valid only if the displacements are small. Further, it is highly dependent on the texture of the ground. Previous studies on the performance of the optical flow algorithm for different textures and scenarios have shown the performance to vary hugely under varied scenarios [14]. If the texture of the ground is varied, results obtained are generally very close to ground truth. But if the texture on the ground is repetitive for example during fall, the ground is littered with mostly similar looking golden leaves and it is nearly impossible to detect change in textures from frame to frame. This makes the pose estimation from the optical flow noisy.

Pose Estimation - Semi Dense Visual Odometry (SDVO)

The front facing camera is primarily used for depth estimation. We build on the work of Engel *et al.* [8] to estimate pose from a inverse depth map estimated from a monocular image. They propose that, given an image $I_M : \Omega_{D_M} \rightarrow \mathbb{R}$, a semi-dense inverse depth map for the current image $D_M : \Omega_{D_M} \rightarrow \mathbb{R}^+$, where Ω_D contains all pixels that have a valid depth, the camera pose of the new frames can be estimated using direct image alignment. Given the current map $\{I_M, D_M\}$, the relative pose $\xi \in SE(3)$ of a new frame I is obtained by directly minimizing the photo metric error

$$E(\xi) := \sum_{x \in \Omega_{D_M}} \|I_M(x) - I(w(x, D_M(x), \xi))\|_\delta \quad (4)$$

where $w: \Omega_{D_M} \times \mathbb{R} \times SE(3) \rightarrow \omega$ projects a point from the reference frame image into the new frame. D and V denote mean and variance of the inverse depth, and $\|\cdot\|_\delta$ is the Huber norm to account for outliers. The minimum is computed using iteratively re-weighted Levenberg-Marquardt minimization. The warp function $w(x, D_M, T)$ as described in [18] is defined as

$$w(x, D_M, T) = \pi(T\pi^{-1}(x, D_M(x))). \quad (5)$$

where $\pi^{-1}(x, D_M)$ represents the inverse projection function and T represents the transformation matrix, for a rigid body motion g

$$T = \begin{bmatrix} \mathbf{R}_{3,3} & \mathbf{t}_{3,1} \\ 0 & 1 \end{bmatrix} \quad (6)$$

Since T is an over parametrized representation of g and has twelve parameters while g has only six degrees of freedom, we use ξ , the twist coordinate representation for T where

$$(\xi) = \log_{SE(3)}(T) \quad (7)$$

The photometric error is minimized iteratively and the pose update step can be represented as

$$(\xi)^{n+1} = \delta(\xi)^n \circ (\xi)^n \quad (8)$$

where $\delta(\xi)^n$ is calculated by solving for the minimum of a Gauss Newton second order approximation of E . The estimation of the depth map for our system can be seen in Figure 3.

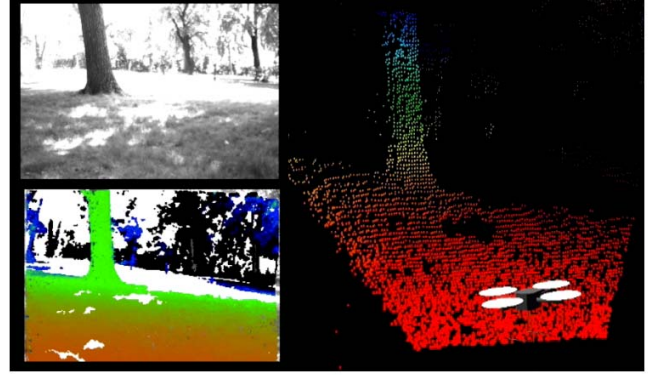


Figure 3: Semi Dense Depth Map Estimation. (Top Left) Example image frame. (Bottom Left) Depth Map. Colormap: Inverted Jet, Black pixels are for invalid pixels (Right) Corresponding 3D point cloud. Color based on height

A drawback of the SDVO is that it is highly sensitive to large inter frame rotations. Since it tracks points from keyframe to keyframe, if the rotations are large, there are jumps in tracking the projection function w fails to map a point from one image I_1 to another I_2 if the rotation is very large. This can be seen in Figure 4.

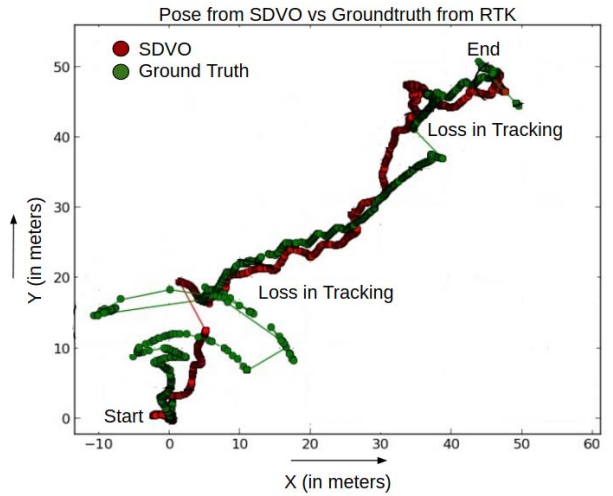


Figure 4: Loss in tracking pose from SDVO when there are large rotations.

Pose Fusion

To fix this problem of unreliable poses from both the downward facing camera and the front facing camera, standard pose fusion approaches are investigated in this section. We try two of the most standard approaches to fuse the pose estimates from both the cameras. The first approach uses lie algebraic averaging to produce globally consistent motion as

is described in the work of Govindu *et al.* [15]. Consider the 4x4 euclidean motion matrix $\mathbf{M} \in \text{SE3}$ defined as

$$\mathbf{M} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \quad (9)$$

where $\mathbf{R} \in \text{SO3}$ and $\mathbf{t} \in \mathbb{R}^3$. This $\mathbf{M}$ is converted to lie space by using the log operator. The result is another 4x4 matrix

$$\mathbf{m} = \begin{bmatrix} \Omega & u \\ 0 & 0 \end{bmatrix} \quad (10)$$

where $\mathbf{m} \in \mathfrak{se}(3)$. Algebraic averaging is then performed on this lie algebraic representation of motion. Consider two motions matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ with $\mathbf{M}_1$ being estimated from the projections of the downward facing camera and $\mathbf{M}_2$ being estimated from the projections of front facing camera. $\mathbf{M}_2$ is rotationally aligned and scaled to the same world as $\mathbf{M}_1$. These are then converted to lie space and then averaged out according to the algorithm described in [15]. We observe that such lie algebraic averaging results are still prone to losses in tracking. (see Fig 5). This can be attributed to the fact that the loss in tracking from the SDVO carries over to the lie space and the algebraic average in the lie space.

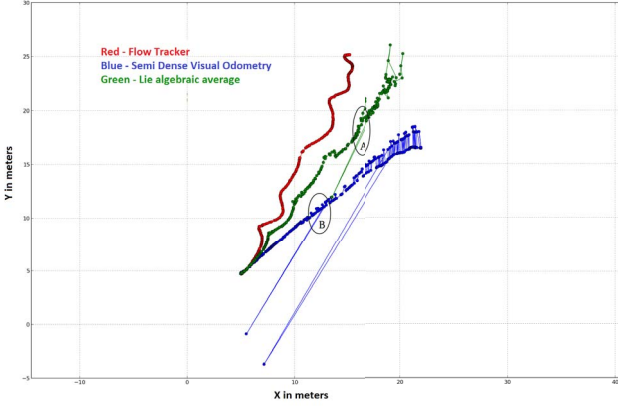


Figure 5: Loss of tracking in SDVO causes huge jumps in the pose which is also reflected in the lie algebraically averaged pose. This can be clearly seen at instances A and B

Another attempt to offset the reset problem and mitigate the unreliability in the pose from the flow tracker, is to use a standard variance weighted input averaging technique (a Kalman Filter based fusion) [6, 11, 12]. Treating the flow tracker as a known prior distribution $\mathcal{N}(d_p, \sigma_i^2)$ and the SDVO as a noisy observation $\mathcal{N}(d_q, \sigma_j^2)$, the posterior is given by

$$\mathcal{N}\left(\frac{\sigma_i^2 d_q + \sigma_j^2 d_p}{\sigma_i^2 + \sigma_j^2}, \frac{\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2}\right) \quad (11)$$

However, even with Kalman Filtering our results tend to be only marginally better. Evaluation of pose estimates was done by bench marking with a real time kinematic (RTK) GPS [16] with an accuracy up to 5 cm.

It is also noticed that at points where tracking is lost for the entire frame, there is a loss in depth map propagation. At such instances the entire SDVO pipeline needs to be reset and this is highly undesirable for fast monocular flight through a cluttered environment. Using this as motivation, in the next section we describe a mechanism where we couple our IMU with the SDVO to ensure robust pose and depth estimates.

Absolute error metrics



Figure 6: Pose benchmarking using a RTK GPS over a distance of 45 meters.

5. IMU PRE-INTEGRATION

Visual Odometry (VO) methods such as the flow tracker and the SDVO described in Section 4 are susceptible to noisy estimates. A possible solution is to increase the frame rate of the camera. But for a small agile platform such as ours, this does not scale well due to constraints on power and computational complexity. We build on some of the work explored in [2, 7, 8]. The key idea is to incorporate data provided by the on board Inertial Measurement Unit (IMU) to accurately estimate the motion between frames.

Let the IMU measurements recorded between two consecutive keyframes i and j be denoted by $\mathcal{I}_{i,j}$. Let the camera measurements at keyframes i be denoted by $\mathcal{C}_i$ and $\mathcal{K}_t$ denote all keyframes up to time t . The set of measurements collected up to time t can then be given as

$$\mathcal{Z}_k = \{\mathcal{C}_i, \mathcal{I}_{i,j}\}_{i,j \in \mathcal{K}_t} \quad (12)$$

We use this set of measurements to infer the motion of the system. Measurements between keyframes can be summarized into a single compound measurement called the preintegrated IMU measurement which constrains the inter-frame motion between consecutive frames. This was first proposed in [17] and used Euler angles for representing motion. [19] extended this to lie-manifolds. The resulting preintegrated rotation increment is given as :

$$\Delta \tilde{R}_{i,j} = \sum_{k=i}^{j-1} e^{(\tilde{w}_k - b_i) \Delta t} \quad (13)$$

where $\tilde{w}_k \in \mathbb{R}^3$ is the instantaneous angular velocity of the k^{th} frame relative to the world frame, b_i is a constant noise bias at time t_i and short hand $\Delta t = \sum_{k=i}^j \Delta t$. Equation (13) can be used to estimate the rotational motion between time t_i and t_j from the IMU for the SDVO. The warp function from equation (5) can now be calculated with the estimated motion being updated by using the calculated preintegrated inertial rotation from the previous step. Equation 6 then becomes :

$$\mathbf{T}_{imu} = \begin{bmatrix} \Delta \tilde{R}_{imu} & \mathbf{t} \\ 0 & 1 \end{bmatrix} \quad (14)$$

The update rule for the pose then comes from modifying Equation 8:

$$(\xi)^{n+1} = \delta(\xi)^n \circ (\xi)_{imu}^n \quad (15)$$

where $(\xi)_{imu}^n = \log_{\text{SE}(3)}(\mathbf{T}_{imu})$. $\delta(\xi)^n$ is then computed by solving for the minimum of a Gauss-Newton second order of the photometric error E as described in Equation (4).



Figure 7: (Top Left) Testing Area in Winter Conditions. (Bottom Left) Testing Conditions in Summer. (Right) Top down satellite view of testing area showing density of clutter.

Thus, we preintegrate the IMU measurements in the lie space to provide the system with accurate estimates of motion between frames. Since the optimizer is fed with more robust rotational estimates, the resulting pose after visual-inertial fusion is less susceptible to strong inter-frame rotations. The inverse depth map is propagated from frame to frame, once the pose of the frame has been determined and refined with new stereo depth measurements. Based on the inverse depth estimate d_0 for the pixel, the corresponding 3D point is calculated and projected into the new frame and assigned to the closest integer pixel position providing the new inverse depth estimate d_1 . The new inverse depth map can then be approximated as:

$$d_1(d_0) = (d_0^{-1} - t_z)^{-1} \quad (16)$$

where t_z is the camera translation along the optical axis. For each frame, after the depth map has been updated, a regularization step is performed by assigning each inverse depth value the average of the surrounding inverse depths, weighted by their respective inverse variance.

6. EXPERIMENTS

In this section we analyze our proposed IMU preintegration for SDVO by testing it on our UAV platform in an outdoor cluttered environment. All experiments are conducted while flying the UAV through dense clutter (Figure 7).

In order to collect groundtruth depth values, a Bumblebee color stereo camera pair (1024×768 at 20 Hz) was rigidly mounted with respect to the front camera using a custom 3D printed fiber plastic encasing. We calibrate the rigid body transform between the front camera and the left camera of the stereo pair. Stereo depth images and front camera images are recorded simultaneously while flying the drone. The depth images are then transformed to the front cameras coordinate system to provide groundtruth depth values for every pixel.

Accuracy of Depth Maps

The performance of the SDVO with preintegrated IMU is first evaluated for depth map accuracy. Figure 8 shows the average depth error against ground truth depth images obtained from stereo processing as explained above. To establish baselines, the performance is compared against state-of-the-art approaches for real-time monocular depth estimation using non-linear regression [20] and Depth CNN [21]. The testing area is split into 2 separate regions - one to collect data for the learning model and another region to test the learned model to ensure there is no overfitting. The corpus of train and test data consists of 50k and 10k images respectively. The error plots have been binned with respect to the ground-truth. It can be observed that SDVO with IMU intergration performs really well with low error values up to [15, 20] m. Please note that both the baseline methods are learning-based approaches and thus have been trained on the training data collected. This graph nevertheless serves to show the accuracy of our proposed perception method. We can observe that fusion of inertial data according to our proposed formulation improves the accuracy of the eventual depth maps. In particular, this can be directly attributed to two factors: First, better initial conditions to the optimizer leads to faster convergence and prevents it from being stuck in a local minima. In addition to this the preintegration reduces the number of tracking losses since now we have an estimate of motion for every instant of time and there is no loss in tracking pose.

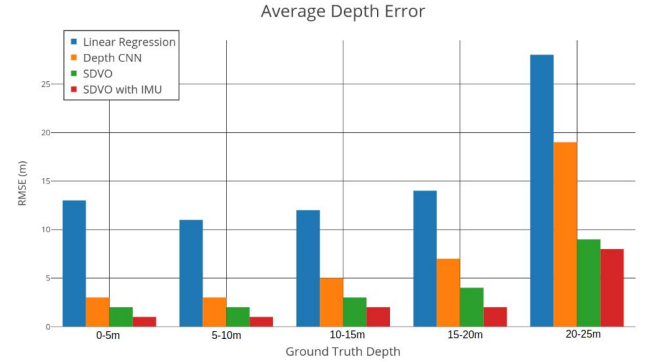


Figure 8: Average root-mean-squared-error (RMSE) binned over groundtruth depth buckets of [1, 5] m, [5, 10] m, etc. Groundtruth depth images are obtained from stereo image processing.

Pose Evaluation

With the IMU injection into the SDVO, the pose drift is considerably lesser. In addition to this, the instances where SDVO would crash are reduced to by 55% over a distance of 1 km. Further, we observe that the accuracy of the pose from the SDVO is very close to the groundtruth. In addition to this observe that the drift over time in the pose is also minimized. This can be attributed to the fact, that every time the SDVO loses tracking, its reference needs to be set to that of the world frame. This repeated loss of tracking induces errors in the long term pose estimation. Figure 9 shows that with just one reset, the pose drifts by as much as 16 meters. With the IMU preintegration, the drift is reduced to 7.2 meters.

Qualitative Evaluation of Depth Maps

Qualitatively, the estimated depth maps can be observed to be within the uncertainty range of stereo, as shown in Figure

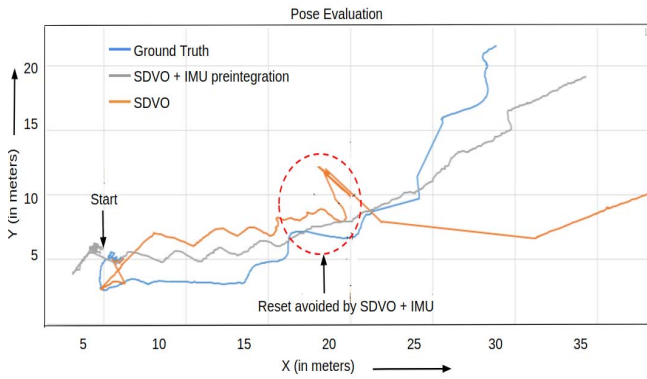


Figure 9: Trajectory comparison using SDVO with and without injection. The groundtruth is obtained by using a RTK GPS.

10. Its interesting to note from the qualitative results that our method is able to robustly handle even poor quality images that suffer from over- or under-exposure, shadows, etc. Since pose estimation in the SDVO is directly coupled with the depth map generation (Equation 16), it is safe to conclude that the high quality depth maps generated can be attributed to robust pose estimation. Robust depth estimation is important since the underlying planner plans optimal trajectories on the computed depth values. If the depth values for the pixels are uncertain, the planner computes trajectories that can be suboptimal. In addition to this higher uncertainty in depth can result in trajectories that could potentially result in catastrophic crashes.

7. CONCLUSION

Thus, in this paper we have presented a novel approach for achieving robust pose estimation as well as depth map estimation by adding pre integrated IMU terms into the optimizer for a semi dense visual odometry method with a monocular camera. The additional IMU integration on the lie manifold comes at a very low computational cost. We prove our proposed methodology works reliably when tested in an outdoor environment with dense clutter. Thus in an attempt to fix the problem of unreliable pose, we not only achieve robust pose estimation but are also able to produce high quality depth estimates from a monocular camera.

In future work, we hope to move towards using some of the recent advances in deep learning and machine learning to learn the instances at which the pose from the IMU should be integrated into the SDVO optimizer as opposed to integrating it at all timesteps.

ACKNOWLEDGMENTS

The authors would like to thank Shreyansh Daftry, Sam Zheng, Debadepta Dey and others at CMU RI for interesting discussions related to this paper. This work was sponsored by funding from the Office of Naval Research MURI grant for 'Provably Stable Vision-based Control of High-speed Flights through Forest and Urban Environments'.

REFERENCES

- [1] Nistr, David, Oleg Naroditsky, and James Bergen. "Visual odometry." *Computer Vision and Pattern Recognition*, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 1. Ieee, 2004.
- [2] Shen, Shaojie, Nathan Michael, and Vijay Kumar. "Tightly-coupled monocular visual-inertial fusion for autonomous flight of rotorcraft MAVs." *Robotics and Automation (ICRA)*, 2015 IEEE International Conference on. IEEE, 2015.
- [3] Dissanayake, Gamini, et al. "A review of recent developments in simultaneous localization and mapping." *Industrial and Information Systems (ICIIS)*, 2011 6th IEEE International Conference on. IEEE, 2011.
- [4] Shen, Shaojie, et al. "Vision-Based pose estimation and Trajectory Control Towards High-Speed Flight with a Quadrotor." *Robotics: Science and Systems*. Vol. 1. 2013.
- [5] Rasmussen, Christopher, Yan Lu, and Mehmet Kocamaz. "A trail-following UAV which uses appearance and structural cues." *Field and Service Robotics*. Springer Berlin Heidelberg, 2014. APA
- [6] Mourikis, Anastasios I., and Stergios I. Roumeliotis. "A multi-state constraint Kalman filter for vision-aided inertial navigation." *Robotics and automation*, 2007 IEEE international conference on. IEEE, 2007.
- [7] Engel, Jakob, Thomas Schps, and Daniel Cremers. "LSD-SLAM: Large-scale direct monocular SLAM." *European Conference on Computer Vision*. Springer, Cham, 2014.
- [8] Engel, Jakob, Jurgen Sturm, and Daniel Cremers. "Semi-dense visual odometry for a monocular camera." *Proceedings of the IEEE international conference on computer vision*. 2013.
- [9] Shen, Shaojie, Nathan Michael, and Vijay Kumar. "Autonomous multi-floor indoor navigation with a computationally constrained MAV." *Robotics and automation (ICRA)*, 2011 IEEE international conference on. IEEE, 2011.
- [10] Shen, Shaojie, Nathan Michael, and Vijay Kumar. "3D indoor exploration with a computationally constrained MAV." *Robotics: Science and Systems*. 2011.
- [11] Li, Mingyang, and Anastasios I. Mourikis. "High-precision, consistent EKF-based visual-inertial odometry." *The International Journal of Robotics Research* 32.6 (2013): 690-711.
- [12] Tanskanen, Petri, et al. "Semi-direct EKF-based monocular visual-inertial odometry." *Intelligent UAVs and Systems (IROS)*, 2015 IEEE/RSJ International Conference on. IEEE, 2015.
- [13] Bouguet, Jean-Yves. "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm." *Intel Corporation* 5.1-10 (2001): 4.
- [14] Barron, John L., David J. Fleet, and Steven S. Beauchemin. "Performance of optical flow techniques." *International journal of computer vision* 12.1 (1994): 43-77.
- [15] Govindu, Venu Madhav. "Lie-algebraic averaging for globally consistent motion estimation." *Computer Vision and Pattern Recognition*, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. Vol. 1. IEEE, 2004.
- [16] Keller, Russell J., Mark E. Nichols, and Arthur F. Lange. "Methods and apparatus for precision agriculture operations utilizing real time kinematic global positioning system systems." *U.S. Patent No. 6,199,000*. 6 Mar. 2001.

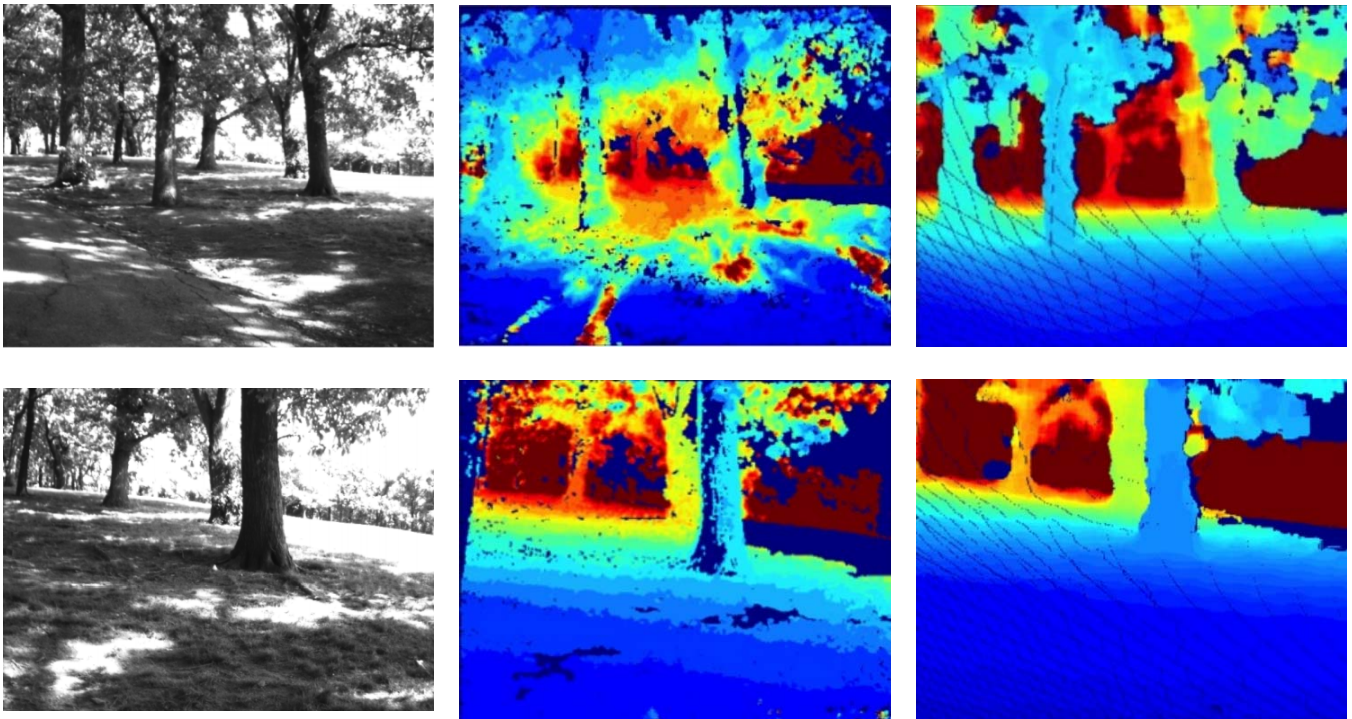


Figure 10: Qualitative analysis of depth maps produced by proposed SDVO + pre integrated IMU with respect to ground truth. (Left) Image from the front facing camera of the UAV. (Middle) Estimated Depth Map. (Right) Ground-truth depth map, Note:jet colormap

- [17] Lupton, Todd, and Salah Sukkarieh. "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions." *IEEE Transactions on Robotics* 28.1 (2012): 61-76.
- [18] Kerl, Christian. "Odometry from rgb-d cameras for autonomous quadcopters." Master's Thesis, Technical University (2012).
- [19] Forster, Christian, et al. "IMU preintegration on manifold for efficient visual-inertial maximum-a-posteriori estimation." Georgia Institute of Technology, 2015.
- [20] Dey, Debadeepta, et al. "Vision and learning for deliberative monocular cluttered flight." *Field and Service Robotics*. Springer International Publishing, 2016.
- [21] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." *Advances in neural information processing systems*. 2014.
- [22] Hoppe, Christof, et al. "Incremental Surface Extraction from Sparse Structure-from-Motion Point Clouds." *BMVC*. 2013.
- [23] Wu, Changchang. "Towards linear-time incremental structure from motion." *3DTV-Conference, 2013 International Conference on*. IEEE, 2013.
- [24] Hrabar, Stefan, et al. "Combined optic-flow and stereo-based navigation of urban canyons for a UAV." *Intelligent UAVs and Systems, 2005.(IROS 2005)*. 2005 IEEE/RSJ International Conference on. IEEE, 2005.
- [25] Levine, Sergey, et al. "End-to-end training of deep visuomotor policies." *Journal of Machine Learning Research* 17.39 (2016): 1-40.

BIOGRAPHY



Arbaaz Khan received his B.Tech from Manipal University, India in 2016 and is currently pursuing a M.S in Robotics at UPenn. He was a visiting research assistant at Carnegie Mellon University's Robotics Institute where this work was carried out. His current research activities include investigating deep learning, deep reinforcement learning and model based learning methods for Robotics.



Martial Hebert Martial Hebert is a Professor and Director for the Robotics Institute at Carnegie Mellon University. He has a large body work in the areas of computer vision and perception for autonomous systems. His current research interests are in the interpretation of perception data (both 2-D and 3-D), including building models of environments.