Fast Monocular Visual-Inertial Initialization Leveraging Learned Single-View Depth

Nathaniel Merrill, Patrick Geneva, Saimouli Katragadda, Chuchu Chen, and Guoquan Huang

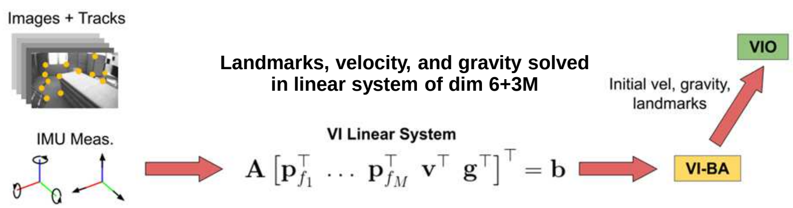
Robot Perception and Navigation Group (RPNG), University of Delaware, USA

Introduction

- Visual-Inertial Odometry (VIO) requires 3D landmarks, velocity, and gravity to initialize
- To decrease user's wait time, it is ideal to initialize as **fast as possible**
- State-of-the-art systems require 2 seconds, large parallax and many features to initialize
- **Contributions**
  - Propose a new initialization method for monocular VIO leveraging **learned monocular depth**
  - Shown to be faster, more accurate, and **more robust,** initializing with low parallax and only **15 features**

Baseline Monocular Visual-Inertial Initialization

- State-of-the-art monocular initialization methods [1] use image tracks and IMU measurements in a VI-SfM to solve for initial conditions
- Large number (M) features and large baseline between keyframes required to initialize



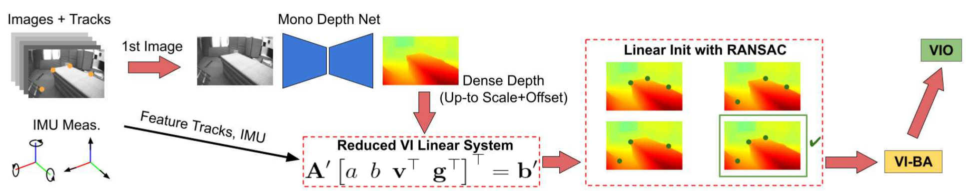
Proposed Initialization Method

- **Key idea:** Leverage learned monocular depth to *reduce* the linear system
  - Propose new model of 3D landmarks w.r.t. learned affine-invariant depth d_i

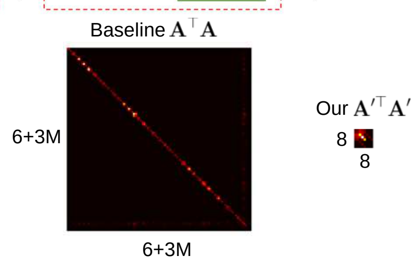
$$\mathbf{p}_{f_i} = z_i \, \boldsymbol{\theta}_{f_i} = (a d_i + b)\, \boldsymbol{\theta}_{f_i}$$

Only estimate a, b **to represent all landmarks**

- Because of the reduced system, RANSAC is practical
  - State size reduced from **6+3M** for M landmarks to just **8**
  - 3 views and 2 features is proved to be the minimal problem, but 4 features used to increase robustness



- **Linear System Structure**
  - Our linear system is considerably smaller than the baseline one
  - System size is 8 regardless of the number of features
  - **Right:** The structure of the normal equations is shown



EuRoC Results

- Tested on EuRoC Vicon room datasets with 5 KFs over a 0.5sec window (avg over hundreds of initializations) using MiDaS [2] v2.1 small network
- **Comparisons**
  - **DS 3D:** Baseline initialization [1] estimating 3D landmarks separately
  - **DS + DP:** Our reimplementation of [3] (mono depth priors in VI-BA only, linear system the same)

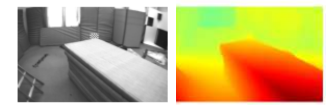
Table: Scale error (%)

Algorithm	Avg.
DS 3D	57.6
DS + DP	58.8
Ours w/o RANSAC	17.3
Ours	**5.8**

Table: ATE (deg / m)

Algorithm	Average
DS 3D	1.592 / 0.028
DS + DP	1.523 / 0.027
Zhou [3]	- / 0.024
Ours w/o RANSAC	1.467 / 0.026
Ours	**1.419 / 0.022**

Example input and depth



TUM-VI Results

- On the TUM-VI dataset, we tested initialization with 5 KFs and only a **300ms window**
- Found reasonable performance of MiDaS on fisheye images despite being trained on rectified

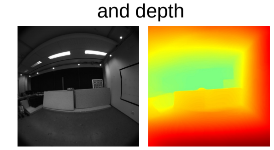
Table: Init window ATE and scale error (deg / m (%))

Algorithm	Average
DS 3D	**1.243** / 0.018 (9.20)
DS + DP	1.276 / 0.020 (8.73)
Ours	1.274 / **0.011 (6.47)**

Table: VIO ATE using init conditions (deg / m)

Algorithm	Average
DS 3D	1.381 / 0.133
DS + DP	1.384 / 0.122
Ours	**1.214 / 0.059**

Example input and depth



Robustness Experiments (TUM-VI)

- **Right:** We added large random noise to the measurements for different outlier percentages, showing our method's **robustness to outliers**
- **Below:** We simulated tracking failure by reducing the number of features available, showing our method is **robust to tracking failure**

Table: Init window ATE (deg / m) with varying artificial outliers

Outliers	Algorithm	Average
5%	DS 3D	1.257 / 0.017
	Ours w/o RANSAC	1.242 / **0.014**
	Ours	**1.047 / 0.014**
10%	DS 3D	1.280 / 0.016
	Ours w/o RANSAC	1.474 / 0.012
	Ours	**0.957 / 0.011**
25%	DS 3D	1.995 / 0.021
	Ours w/o RANSAC	2.413 / 0.025
	Ours	**1.409 / 0.014**
45%	DS 3D	2.929 / 0.035
	Ours w/o RANSAC	4.035 / 0.039
	Ours	**2.663 / 0.030**

Table: % of successful initializations

Algorithm	60 feats	45 feats	30 feats	15 feats
DS 3D	81.25	17.50	33.75	2.50
DS 3D + DP	78.75	16.25	32.50	2.50
Ours w/o RANSAC	**100.00**	**98.75**	**97.50**	**55.00**
Ours	**100.00**	95.00	96.25	47.50

References

[1] T.-C. Dong-Si and A. I. Mourikis, "Estimator initialization in vision-aided inertial navigation with unknown camera-imu calibration," in IROS 2012

[2] R. Ranftl et. al, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," in TPAMI 2022.

[3] Y. Zhou et. al, "Learned Monocular Depth Priors in Visual-Inertial Initialization," in ECCV 2022.

Link to paper