# Robust Multi-Stereo Visual-Inertial Odometry

Joshua Jaekel[1] and Michael Kaess[1]

***Abstract*— In this paper we present a novel multi-stereo visual-inertial odometry (VIO) framework which aims to improve the robustness of a robot's state estimate during aggressive motion and in visually challenging environments. Our system uses a fixed-lag smoother which jointly optimizes for poses and landmarks across all stereo pairs. We also makes use of nonlinear factor recovery (NFR) in order to enforce a sparse information matrix after marginalization while also maintaining the information contained in the dense priors. We propose a novel 1-point RANdom SAmple Consensus (RANSAC) algorithm which is able to jointly perform outlier rejection across features from all stereo pairs. The result is a VIO system which is able to maintain an accurate state estimate under conditions which have typically proven to be challenging for traditional state-of-the-art VIO systems. We demonstrate the benefits of our proposed multi-stereo algorithm by evaluating it against VINS-Mono on three challenging simulated micro-aerial vehicle (MAV) flights. We show that our proposed algorithm is able to achieve a significantly lower average trajectory error on all three flights.**

## I. Introduction

State estimation is one of the most fundamental problems in robotics. In many cases core functionalities of a robot such as motion planning, mapping, and control all depend on a reliable state estimate. Cameras and inertial measurement units (IMUs) are two of the most popular sensors used to obtain a state estimate, especially on smaller platforms like MAVs due to their light weight and complementary nature. IMUs provide high frequency data which can provide useful information about short-term dynamics, while cameras provide useful exteroceptive information about the structure of the environment over longer periods of time.

Visual-inertial odometry (VIO) is a technique which uses visual information from one or more cameras, and inertial information from an IMU to estimate the state of a robot relative to some fixed world frame. Specifically, a VIO system aims to estimate the six degree of freedom rigid body transformation between a starting pose and the current pose of the robot. Although VIO frameworks are able to obtain accurate state estimates in many environments, improving the robustness of these algorithms remains a significant challenge. In certain environments, such as those with sparse visual features or inconsistent lighting, current VIO algorithms are prone to failure. Furthermore, certain types of fast or aggressive motions can lead to failures in state estimation. In indirect systems which track features in the scene, these failures can often be attributed to poor feature tracking which results in incorrect camera measurements being used in back-end optimization. In traditional frameworks, where information from only a single monocular camera or single stereo pair is used, a single point of failure is introduced. If the field of view of the camera were to become suddenly occluded or experience rapid exposure changes, the accuracy of the state estimate could drastically decrease or the VIO algorithm could fail all together.

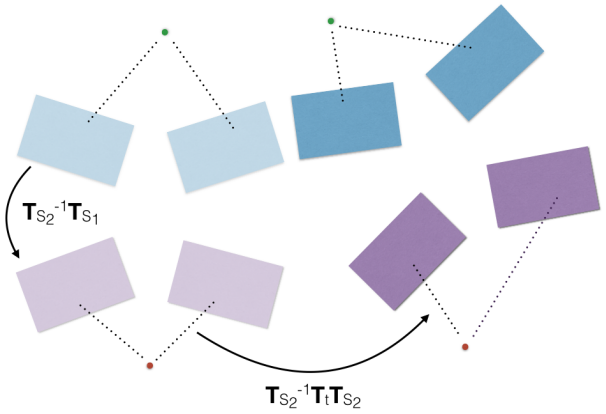


Fig. 1: Transformations between points in different coordinate frames. Stereo pair 1 is visualized in blue, stereo pair 2 is visualized in purple. The faded images represent the stereo frames at time step $t-1$, and the opaque frames represent the stereo frames at time $t$. In the proposed RANSAC algorithm, we triangulate landmarks in the IMU frame and estimate the temporal motion of the robot by comparing the triangulated position of features at consecutive time steps after compensating for the temporal rotation measured from an IMU.

Using information from multiple cameras with non-overlapping fields of view can drastically improve the robustness of a VIO system. If features from one of the cameras were suddenly lost, the VIO algorithm could continue to maintain a state estimate using only features from the other cameras and IMU. Furthermore, if the cameras are configured to have perpendicular optical axes, then when the robot undergoes fast rotation it is possible that at least one of the cameras' optical axis will be closely aligned with the axis of rotation and will be able to track features during the motion. Determining the correct set of features to use for optimization is a non-trivial task. Although several outlier rejection algorithms exist in state-of-the-art VIO pipelines, most of these cannot take advantage of the strong constraints

[1]Joshua Jaekel and Michael Kaess are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {jjaekel, kaess}@andrew.cmu.edu

provided by a calibrated multi-stereo system.

We propose a VIO system capable of incorporating an arbitrary number of stereo pairs with non-overlapping fields of view. Our paper introduces an outlier rejection scheme to jointly select features from all the stereo frames to be added as projection factors in a back-end solver. To demonstrate the benefits of our multi-camera VIO algorithm, we evaluate it in simulation against VINS-Mono, a current state-of-the VIO algorithm. We show that the multi-camera approach is able to maintain a more accurate state estimate over three difficult MAV flights. Our main contributions are:

- The description of a robust front-end pipeline used for multi-stereo VIO
- The design of a novel multi-stereo 1-point RANSAC scheme which operates on stereo triangulated features
- Simulated experiments which demonstrate the benefits of the proposed algorithm against state-of-the-art VIO

## II. Related Work

VIO and simultaneous localization and mapping (SLAM) algorithms can be roughly categorized into two main groups, *direct* and *indirect* methods. Direct methods [1, 2, 3, 21] estimate temporal motion by continuously aligning consecutive camera frames as to minimize the photometric error between them. On the other hand, indirect methods [8, 13, 16, 18] track landmarks in the scene and estimate motion by attempting to minimize the reprojection error between the observed location of features in an image and the projection of their 3D estimated locations.

RANSAC schemes are widely used in most indirect VIO algorithms. These algorithms can either be used to remove erroneous feature correspondences from being inserted into an optimization or to estimate the egomotion of the robot directly. VINS-Mono and its stereo counter part VINS-Fusion [16] both use a fundamental matrix RANSAC approach for outlier rejection. The minimal solution requires 7 correspondences and calculates inliers based on their distance from a candidate epipolar line. In Sun et al.'s [18] implementation of stereo MSCKF [12], a 2-point RANSAC approach described in [20] is used. Like our proposed method, they first compensate for temporal rotation by integrating the IMU, but instead of performing outlier detection on a triangulated 3D point, they apply an independent RANSAC to both the left and right image points and only accept the feature if it is an inlier in both images. Although both of these methods work well for detecting outliers observed from a single camera or stereo pair, neither generalize to feature points across multiple cameras.

There has been extensive work done in using multi-camera systems to improve the robustness of simultaneous localization and mapping (SLAM) systems. Oskiper et al. [14] proposed a multi-stereo VIO which extracts frame-to-frame motion constraints through a 3-point RANSAC and used an extended Kalman filter (EKF) to fuse those constraints with data from an IMU. Houben et al. [6] explored using a multi-camera system in a graph based SLAM framework with their proposed extension of ORB-SLAM [13]. Their system added a factor in the pose graph between key-frames observed from different cameras at the same time step based on the known extrinsic calibration of the multi-camera system. Tribou et al. [19] proposed a multi-camera extension of Parallel Tracking and Mapping [8] (PTAM) using a spherical camera model.

For joint multi-camera outlier rejection, most existing methods use the generalized camera mode (GCM) and generalized epipolar constraint (GEC) introduced by Pless in [15]. In this framework, feature points are parameterized by Plücker vectors which pass through the optical center of the camera in which the feature was observed and the normalized image point. Lee et al. propose a 4-point solution [9] based on the GEC for a multi-camera setup on board an autonomous vehicle. This system assumes the roll and pitch can be directly measured from the IMU but estimates the temporal yaw as part of the RANSAC formulation. In [5] Heng et al. propose a similar 3-point algorithm for a multi-stereo system on board a MAV. Like our proposed method, they also use an estimated rotation from IMU integration, but their algorithm is degenerate in the case of no temporal rotation and no inter-camera correspondences. Although their platform contains stereo cameras, they do not triangulate feature points and instead treat each camera in the stereo pair independently which ensures there will always be inter-camera correspondences.

Our novel RANSAC based outlier rejection scheme makes use of the known extrinsic calibration between cameras to jointly perform outlier rejection across features in all stereo frames. Since feature points are already triangulated as part of our back-end solver, we make use of this information to formulate our 1-point algorithm as opposed to the 3-point algorithm in [5]. Since our algorithm only uses the RANSAC result to select inliers to insert in the back-end and does not directly use the result to calculate odometry directly, we have seen little adverse effects from the sometimes noisy results of stereo triangulation.

## III. Problem Formulation

The goal of this paper is to develop a framework to robustly select a subset of features for use in an indirect VIO algorithm. Measurements associated with each frame consists of relative and marginalized IMU measurements [4, 7] between consecutive poses, as well as stereo projection factors which connect a pose and a landmark, $L \in \mathbb{R}^3$. For each frame, we define a set of camera measurements $\mathcal{C}_t$. This set contains all the measurements across the $K$ stereo cameras. We define $\mathcal{C}_t$ as:

$$\mathcal{C}_t = \bigcup_{j=1}^{K} \mathcal{C}_t^j \tag{1}$$

where $\mathcal{C}_t^j$ is the subset of $\mathcal{C}_t$ containing the measurements observed in stereo pair $j$. We define each individual stereo measurement as:

$$c = \begin{pmatrix} u_l \\ v_l \\ u_r \\ v_l \end{pmatrix} = \begin{pmatrix} \mathbf{p}_L \\ \mathbf{p}_R \end{pmatrix} \in \mathbb{R}^4 \tag{2}$$

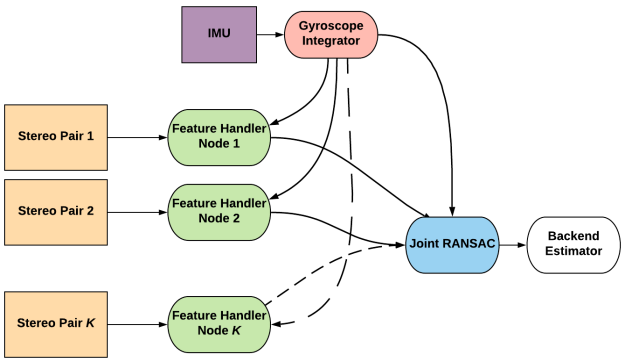


Fig. 2: The structure of the the front-end for the proposed multi-stereo VIO. Features are tracked in each stereo frame by a feature handler instance and then passed to a joint RANSAC algorithm to select inliers to send to the back-end estimator.

This measurement defines the pixel location of the feature point in the left and right images of the stereo camera. We denote each stereo camera as $S_i$ and the extrinsics of the left camera with reference to the body frame as $[\mathbf{R}_{S_i}, \mathbf{t}_{S_i}]$. We denote $\mathbf{R}_t$ and $\mathbf{t}_t$ as the rotation matrix and translation vector which can take a point from the previous body frame $(t-1)$ to the current body frame $(t)$. These transformations are visualized in Figure 1. We define the body frame of the robot as being aligned with the IMU. The goal of the proposed outlier rejection algorithm is to filter the set of candidate landmarks, $\mathcal{F}_t$, and extract a smaller subset of features $\mathcal{C}_t$, to be added as stereo projection factors in the optimization such that each feature is $\mathcal{C}_t$ is consistent with the temporal motion of the robot.

## IV. Robust Multi-Stereo VIO

An indirect VIO system following our framework has three main steps:

1) Feature handling (temporal and stereo matching)
2) Outlier rejection
3) Backend estimation

In this section we elaborate on each of these steps.

### A. Front-end Feature Handler

The role of the front-end is to provide the back-end with valid observations of landmarks in the scene over time. For each stereo pair on the robot we initialize a feature handler. To initialize each feature handler we first uniformly divide our images into a fixed number of buckets and enforce a maximum number of features in each bucket. Bucketing the image ensures we obtain an even distribution of features across the entire image and also avoid landmarks which would give redundant constraints on the optimization. We fill our buckets by detecting Shi-Tomasi features [17] in the left image of the stereo pair and use Kanade-Lucas-Tomasi (KLT) tracking to match features between the left and right images. We define $\mathcal{O}_t^j$ to be the set of features in the previous frame of stereo pair $j$ that are candidates to be tracked.

TABLE I: Notation Summary

| | |
|---|---|
| **Problem Formulation** | |
| $\mathbf{p}_L^t, \mathbf{p}_R^t \in \mathbb{R}^2$ | Left and right candidate feature image coordinates for stereo pair at time step $t$ |
| $\mathbf{p}_L^{t-1}, \mathbf{p}_R^{t-1} \in \mathbb{R}^2$ | The time step $t-1$ image coordinates corresponding to $\mathbf{p}_L^t, \mathbf{p}_R^t$ |
| $S_i$ | The $i$-th stereo stereo pair |
| $[\mathbf{R}_{S_i}, \mathbf{t}_{S_i}]$ | Extrinsics of the left camera of $S_i$ with respect to the body frame |
| $\mathcal{O}_t^j$ | Set of potential features to track at time step $t$ in stereo pair $j$ |
| $\mathcal{N}_t^j$ | Set of new feature points added at time step $t$ in stereo pair $j$ |
| $\mathcal{C}_t$ | Final set of image points to be used for VIO at time step $t$ |
| $\mathcal{C}_t^j$ | The subset of $\mathcal{C}_t$ corresponding to stereo pair $j$ |
| **Multi-Camera RANSAC** | |
| $\mathcal{F}_t$ | Set of successfully temporally tracked image point pairs |
| $\mathcal{F}_t^j$ | The subset of $\mathcal{F}_t$ corresponding to stereo pair $j$ |
| $\mathbf{P}_B^t \in \mathbb{R}^3$ | Triangulated 3D coordinate of a feature in the body frame at time step $t$ |
| $\mathbf{P}_B^{t-1} \in \mathbb{R}^3$ | The time step $t-1$ triangulated 3D feature coordinate corresponding with $\mathbf{P}_B^t$ represented in the body frame |
| $\mathbf{P}_B^{(t-1)'} \in \mathbb{R}^3$ | The 3D feature coordinate $\mathbf{P}_B^{t-1}$ after being rotated into the current (time $t$) frame via $\hat{\mathbf{R}}_t$ |
| $\mathcal{I}$ | Set of candidate inliers for a given iteration of RANSAC |
| $\mathcal{X}$ | Set of triangulated feature points |
| $\hat{\mathbf{R}}_t \in \mathrm{SO}(3)$ | Estimated temporal rotation matrix produced via IMU integration |
| $\hat{\mathbf{t}} \in \mathbb{R}^3$ | Candidate temporal translation from RANSAC |
| $\delta$ | RANSAC threshold |
| $\pi_j$ | Projection function into stereo pair $j$ |

$$\mathcal{O}_t^j = \mathcal{C}_{t-1}^j \cup \mathcal{N}_{t-1}^j \tag{3}$$

where $\mathcal{N}_{t-1}^j$ represents the new features that were added at the previous iteration. At each new image we do the following:

1) Perform KLT tracking from features in previous left image ($\mathcal{O}_t^j$) to the current left image.
2) Perform KLT tracking from the successfully tracked features in the current left image to the current right image. The result is $\mathcal{F}_t^j$.
3) Replenish the buckets which lost features during Steps 1 and 2 by adding new Shi-Tomasi features ($\mathcal{N}_t^j$).

For Step 1, we initialize the tracker with features warped by the temporal rotation, estimated from the IMU. Our algorithm also supports the ability to initialize the temporal tracker by compensating for the translation between frames. This can be done relatively inexpensively since each feature is already triangulated in the multi-camera RANSAC algorithm. We estimate the temporal translation by taking the most recent velocity estimate from the back-end and applying a constant velocity model. We initialize the tracker in Step 2 by compensating for the extrinsic rotation between the left and right cameras in the stereo pair. The output of feature handler $j$ is $\mathcal{F}_t^j$. We denote the set of temporally tracked stereo feature points across all camera frames at time $t$ as

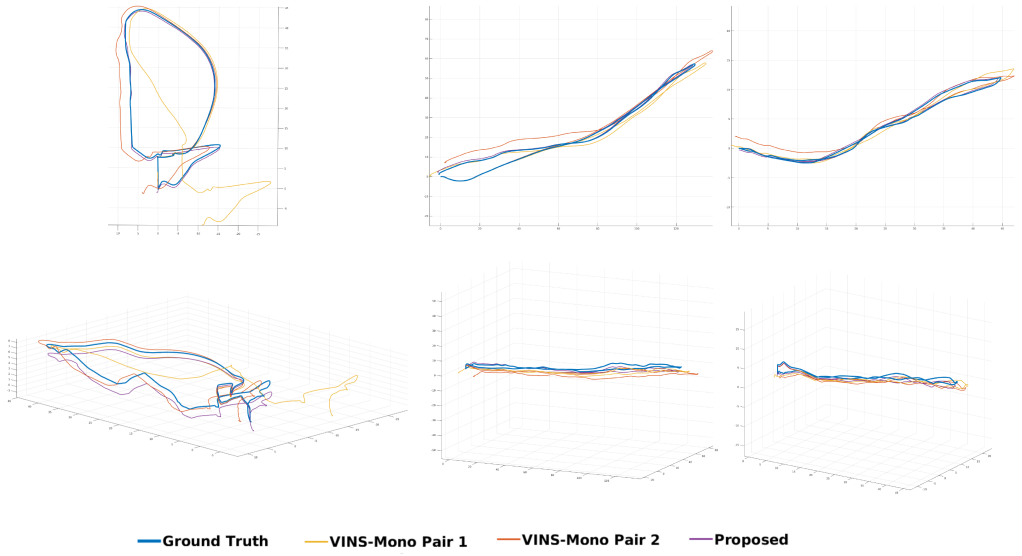


Fig. 3: An overhead view (top row) and side view (bottom row) of the 3 MAV simulated trajectories evaluated against ground truth. The proposed method achieves the lowest average trajectory error on all 3 simulated flights.

$\mathcal{F}_t$.

$$\mathcal{F}_t = \bigcup_{j=1}^{K} \mathcal{F}_t^j \tag{4}$$

Each measurement in $\mathcal{F}_t^j$ is defined as:

$$m = \begin{pmatrix} \mathbf{p}_L^{t-1} \\ \mathbf{p}_R^{t-1} \\ \mathbf{p}_L^{t} \\ \mathbf{p}_R^{t} \end{pmatrix} \in \mathbb{R}^8 \tag{5}$$

where $[\mathbf{p}_L^{t\top}, \mathbf{p}_R^{t\top}]^\top$ represents the stereo measurement in the current image and $[\mathbf{p}_L^{t-1\top}, \mathbf{p}_R^{t-1\top}]^\top$ represents the corresponding measurement in the previous image. Figure 2 shows an overview of the entire front-end structure.

### B. Multi-Stereo RANSAC

Although our outlier rejection scheme can be used to select features for any indirect back-end solver, we see a specific benefit in multi-stereo configurations since our algorithm operates on points triangulated in the body frame. The algorithm is explained in detail in Section V.

### C. Back-end

We use a fixed-lag smoother to optimize for the state of the $m$ previous key frames and $n$ most recent frames. We represent each state, $x_t \in \mathbb{R}^{15}$, as:

$$x_t = [\xi_t^\top, \mathbf{v}_t^\top, \mathbf{b}_t^\top]^\top \tag{6}$$

where $\xi \in \mathbb{R}^6$ is the 6 degree of freedom robot pose, $\mathbf{v} \in \mathbb{R}^3$ is the robot velocity, and $\mathbf{b} \in \mathbb{R}^6$ is the vector of biases of the accelerometer and gyroscope.

In order to bound computational complexity and maintain an optimization window of fixed size, fixed-lag smoothers continuously marginalize out selected states. This marginalization creates fill-in in the information matrix, and over time will destroy the sparsity of the information matrix and underlying factor graph. Without a sparse underlying factor graph, further optimization becomes computationally inefficient. To deal with this issue, several existing VIO algorithms [10, 16] selectively discard measurements to enforce sparsity. This strategy is not optimal from an information theoretic standpoint as those discarded measurements contain information which can no longer be recovered. Hsiung et. al [7] proposed a marginalization strategy which follows from Marzuran et al.'s Nonlinear Factor Recovery [11]. This strategy maintains sparsity by minimizing the Kullback-Leibler divergence (KLD) between a specified sparse topology and the original dense structure induced by marginalization. We adopt the same marginalization strategy.

## V. Multi-Stereo RANSAC Aglorithm

In this section we present our multi-stereo RANSAC algorithm. For the sake of clarity we have included Table I which summarizes the notation to be used in this section. We will also explain the method with reference to Algorithm 1.

We triangulate each candidate feature point in $\mathcal{F}_t$ in the IMU frame. This is done for the features in the current image (line 7) as well as the corresponding features from the previous image (line 6). We can estimate the temporal rotation, $\hat{\mathbf{R}}_t$ between consecutive camera frames by integrating measurements from the on board gyroscope (line

**Algorithm 1:** Multi-Stereo RANSAC

1 $\mathcal{X} \leftarrow \emptyset$
2 $\mathcal{C}_t \leftarrow \emptyset$
3 $\hat{\mathbf{R}}_t \leftarrow$ IMUIntegration()
4 **for** $j := 1$ *to* $K$ **do**
5  **for** $(\mathbf{p}_L^{t-1}, \mathbf{p}_R^{t-1}, \mathbf{p}_L^t, \mathbf{p}_R^t) \in \mathcal{F}_t^j$ **do**
6  $\mathbf{P}_B^{t-1} \leftarrow$ Triangulate$(\mathbf{p}_L^{t-1}, \mathbf{p}_R^{t-1})$
7  $\mathbf{P}_B^t \leftarrow$ Triangulate$(\mathbf{p}_L^t, \mathbf{p}_R^t)$
8  $\mathbf{P}_B^{(t-1)'} \leftarrow \hat{\mathbf{R}}_t \mathbf{P}_B^{t-1}$
9  $\mathcal{X} \leftarrow \mathcal{X} \cup \{(\mathbf{P}_B^{(t-1)'}, \mathbf{P}_B^t, \mathbf{p}_L^t, \mathbf{p}_R^t, j)\}$
10  **end**
11 **end**
12 **for** $i := 1$ *to* $N$ **do**
13  $\mathcal{I} \leftarrow \emptyset$
14  $(\hat{\mathbf{P}}_B^{(t-1)'}, \hat{\mathbf{P}}_B^i, \dots) \xleftarrow{Rand} \mathcal{X}$
15  $\hat{\mathbf{t}} \leftarrow \hat{\mathbf{P}}_B^i - \hat{\mathbf{P}}_B^{(i-1)'}$
16  **for** $(\mathbf{P}_B^{(t-1)'}, \mathbf{P}_B^t, \mathbf{p}_L^t, j) \in \mathcal{X}$ **do**
17  $(\hat{\mathbf{p}}_L^t, \hat{\mathbf{p}}_R^t) \leftarrow \pi_j(\mathbf{P}_B^{t-1}, \hat{\mathbf{R}}_t, \hat{\mathbf{t}}, \mathbf{R}_{C_j}, \mathbf{t}_{C_j})$
18  **if** $(||\hat{\mathbf{p}}_L^t - \mathbf{p}_L^t||_2^2 < \delta)$ **then**
19  $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\mathbf{p}_L^t, \mathbf{p}_R^t)\}$
20  **end**
21  **end**
22  **if** $(|\mathcal{I}| > |\mathcal{C}_t|)$ **then**
23  $\mathcal{C}_t \leftarrow \mathcal{I}$
24  **end**
25 **end**
26 **return** $\mathcal{C}_t$

TABLE II: Number of iterations needed for RANSAC

| $s$ | 1 | 2 | 3 | $\cdots$ | 7 |
|---|---|---|---|---|---|
| $N$ | 7 | 16 | 35 | $\cdots$ | 588 |

3). Using this estimate for temporal rotation we can rotate the triangulated points from the previous time step into the current time frame (line 8). At this point we can expect that the landmarks in $\mathbf{P}_B^{(t-1)'}$ and $\mathbf{P}_B^t$ only differ by the temporal translation of the robot. We can obtain an estimate for the temporal translation by randomly selecting a single feature correspondence and subtracting their 3D positions (line 14 and 15). Using this estimate for translation, we then reproject all the triangulated feature points in the previous temporal frame into the current image frame (line 17). We perform outlier rejection by thresholding the Euclidean distance between the reprojected points in the left camera frame and the observed points (line 18). Our RANSAC based outlier rejection scheme iteratively repeats this process and selects the largest set of inliers to insert as measurements in the factor graph. A main benefit of the 1-point algorithm is that it only requires a small number of iterations to provide strong probabilistic guarantees. This relationship is expressed as:

$$N = \frac{\log(1-p)}{\log(1-(1-\epsilon)^s)} \tag{7}$$

Where $N$ is the number of iterations needed, $p$ is the desired probability of success, $\epsilon$ is the estimated percentage of outliers and $s$ is the number of points required for a minimal solution. Table II shows the number of RANSAC iterations required to get to find a set of inliers with probability of success $p = 0.99$ and a conservative estimate of the percentage of outliers of $\epsilon = 0.5$. We can see that the number of iterations required grows exponentially with the number of points required for a minimal solution.

(a)

(b)

(c)

Fig. 4: Simulated multi-stereo dataset. Top: overview of the scene. Bottom: images from the left and right camera of the front facing stereo pair.

## VI. Experimental Results

The proposed multi-camera VIO pipeline was evaluated in a simulated *Gazebo* environment (shown in Figure 4). We compare the accuracy of the trajectory against VINS-Mono running on the front left camera and back left camera respectively. Three trajectories flown by a small MAV are used for evaluation. The camera configuration includes one stereo pair facing forwards and another facing backwards. All three of these flights include aggressive motion, fast rotation, and sudden obstructions in front of the cameras. The average trajectory error (ATE) is reported in Table III, and the trajectories are visualized against ground truth in Figure 3. We can see that our proposed method achieves the lowest ATE for all 3 of the proposed trajectories.

## VII. Conclusion

In this paper we have introduced a novel RANSAC algorithm which is used as part of a multi-stereo VIO pipeline

TABLE III: ATE in Simulated Environments

| | Traj. 1 ATE (m) | Traj. 2 ATE (m) | Traj. 3 ATE (m) |
|---|---|---|---|
| Proposed | **0.700** | **0.780** | **0.199** |
| VINS-Mono Front Left | 1.561 | 3.362 | 1.533 |
| VINS-Mono Back Left | 7.048 | 4.181 | 1.265 |

on MAVs. Our outlier rejection scheme operates on stereo triangulated points, which are already calculated as part of the back-end optimization. Our algorithm leverages the known extrinsics between cameras as well as the multi-view observation of each feature point from the stereo pair to be able to jointly perform outlier rejection with features observed across an arbitrary number of camera frames within reasonable computational constraints. We demonstrate that a multi-stereo VIO framework using this outlier rejection scheme is able to beat state-of-the-art VIO algorithms running on any of the cameras individually in a simulated environment. We have formulated our RANSAC in such a way that the minimal solution can be solved with a single correspondence, resulting in a more computationally efficient algorithm compared to other multi-camera outlier rejection schemes.

In the future we plan to incorporate the uncertainty of the extrinsic calibration in the RANSAC formulation as well as in the back-end optimization. We also plan to collect real-world data and evaluate the algorithm against several other state-of-the-art algorithms.

## ACKNOWLEDGMENT

The authors would like to thank members of the Robot Perception Lab for their support, and especially Joshua Mangelson and Eric Dexheimer for their theoretical and technical discussions which helped in the development of this work.

## REFERENCES

[1] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Eur. Conf. on Computer Vision (ECCV)*, September 2014.

[2] J. Engel, J. Stueckler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, September 2015.

[3] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, March 2018.

[4] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Trans. on Robotics (TRO)*, vol. 33, no. 1, pp. 1–21, 2017.

[5] L. Heng, G. H. Lee, and M. Pollefeys, "Self-calibration and visual slam with a multi-camera system on a micro aerial vehicle," *Autonomous Robots (AURO)*, vol. 39, pp. 259–277, 2015.

[6] S. Houben, J. Quenzel, N. Krombach, and S. Behnke, "Efficient multi-camera visual-interial slam for micro aerial vehicles," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, Oct. 2016, pp. 1616–1622.

[7] J. Hsiung, M. Hsiao, E. Westman, R. Valencia, and M. Kaess, "Information sparsification in visual-inertial odometry," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Madrid, Spain, 2018, pp. 1146–1153.

[8] G. Klein and D. Murray, "Parallel tracking and mapping on a camera phone," *International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 83–86, 2009.

[9] G. H. Lee, M. Pollefeyes, and F. Fraundorfer, "Relative pose estimation for a multi-camera system with known vertical direction," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[10] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visualinertial odometry using nonlinear optimization," *Intl. J. of Robotics Research (IJRR)*, vol. 34, no. 3, pp. 314–334, 2015.

[11] M. Mazuran, W. Burgard, and G. D. Tipaldi, "Nonlinear factor recovery for long-term SLAM," *Intl. J. of Robotics Research (IJRR)*, vol. 35, no. 1-3, pp. 50–72, 2016.

[12] A. I. Mourikis and S. I. Roumeliotis, "A multi-state Kalman filter for vision-aided inertial navigation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, no. April, 2007, pp. 10–14.

[13] R. Mur-Artal, J. Montiel, and J. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Trans. on Robotics (TRO)*, vol. 31, no. 5, pp. 1147–1163, 2015.

[14] T. Oskiper, Z. Zhu, S. Samarasekera, and R. Kumar, "Visual odometry system using multiple stereo cameras and inertial measurement unit," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2007.

[15] R. Pless, "Using many cameras as one," in *IEEE Computing Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2003.

[16] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Trans. on Robotics (TRO)*, vol. 34, no. 4, pp. 1004–1020, 2018.

[17] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.

[18] K. Sun, K. Mohta, B. Pfrommer, M. Watterson, S. Liu, Y. Mulgaonkar, C. J. Taylor, and V. Kumar, "Robust stereo visual inertial odometry for fast autonomous flight," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 2, pp. 965–972, 2018.

[19] M. J. Tribou, A. Harmat, D. W. L. Wang, I. Sharf, and S. L. Waslander, "Multi-camera parallel tracking and mapping with non-overlapping fields of view," *Intl. J. of Robotics Research (IJRR)*, vol. 34, no. 12, pp. 1480–1500, 2015.

[20] C. Troiani, A. Martinelli, C. Laugier, and D. Scaramuzza, "2-point-based outlier rejection for camera-imu systems with applications to micro aerial vehicles," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Hong Kong, China, Jun. 2014.

[21] V. Usenko, J. Engel, J. Stückler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2016, pp. 1885–1892.