2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

A Robust Multi-Stereo Visual-Inertial Odometry Pipeline

Joshua Jaekel, Joshua G. Mangelson, Sebastian Scherer, and Michael Kaess

***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 propose a 1-point RANdom SAmple Consensus (RANSAC) algorithm which is able to perform outlier rejection across features from all stereo pairs. To handle the problem of noisy extrinsics, we account for uncertainty in the calibration of each stereo pair and model it in both our front-end and back-end. The result is a VIO system which is able to maintain an accurate state estimate under conditions that 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 with both simulated and real world data. We show that our proposed algorithm is able to maintain a state estimate in scenarios where traditional VIO algorithms fail.**

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 give 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.

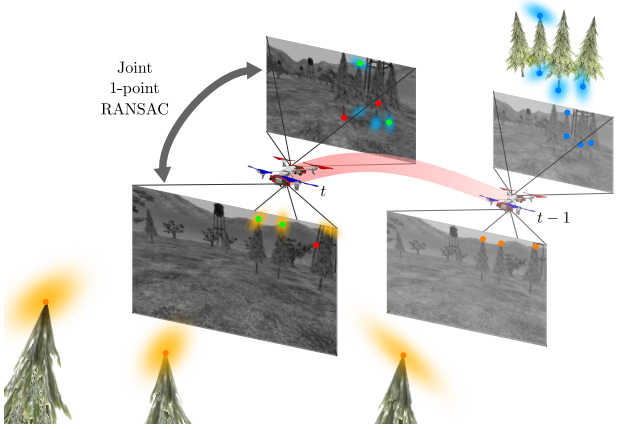

Fig. 1: Visualization of the proposed multi-stereo RANSAC algorithm. Sample feature points for the forward facing (orange) and backward facing (blue) cameras are shown. Also visualized are the uncertain 3D positions of the points in the world and their uncertain projections back into the image. In the proposed method we are able to jointly select inliers from the features observed in all cameras and account for the extrinsic uncertainty in the process.

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 axes 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 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. Our proposed system addresses the problem of having an unreliable extrinsic calibration by modeling the uncertainty in both the outlier rejection scheme and the back-end graph based optimization.

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA. {jjaekel, jmangels, basti, kaess}@andrew.cmu.edu
This work was partially supported by DARPA agreement #HR00111820044.

Our RANSAC scheme also models the 3D uncertainty of triangulated points, which is quadratic in depth. Since our algorithm only uses the RANSAC result to select inliers to insert in the back-end and does not use the result to calculate odometry directly, we have seen little adverse effects from the sometimes noisy results of DLT triangulation, which minimizes algebraic error instead of geometric error, but in a much more computationally efficient manor than iterative approaches [7].

To demonstrate the benefits of our multi-camera VIO algorithm, we evaluate it in simulation against VINS-Fusion [22], a current state-of-the-art VIO algorithm running on each stereo pair individually. We also provide a detailed comparison of our proposed outlier rejection scheme against a fundamental matrix RANSAC approach both in terms of the accuracy of the resulting state estimate and the computational load of each method. We show that the multi-camera approach is able to maintain a more accurate state estimate in several challenging situations. This paper is an extension of our previous work in [12]. Our main contributions are:

- The description of a multi-stereo front-end pipeline which can be used with a feature based back-end solver
- The design and evaluation of 1-point RANSAC scheme with a novel application to multi-stereo camera configurations
- A framework to model uncertainty in camera extrinsics in both the front-end and back-end of our algorithm

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 [2, 3, 4, 28] estimate temporal motion by continuously aligning consecutive camera frames as to minimize the photometric error between them. On the other hand, indirect methods [13, 19, 22, 25] 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 counterpart VINS-Fusion [22] 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 [25] implementation of stereo MSCKF [18], a 2-point RANSAC approach described in [27] is used. They first compensate for temporal rotation by integrating the IMU. 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 features 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. [20] 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. [10] explored using a multi-camera system in a graph based SLAM framework with their proposed extension of ORB-SLAM [19]. 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. [26] proposed a multi-camera extension of Parallel Tracking and Mapping [13] (PTAM) using a spherical camera model.

For joint multi-camera outlier rejection, most existing methods use the generalized camera model (GCM) and generalized epipolar constraint (GEC) introduced by Pless in [21]. 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. [14] propose a 4-point solution 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 [9] 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 must treat each camera in the stereo pair independently to ensure there will always be inter-camera correspondences.

In a separate work [8], Heng et al. describe a 1-point RANSAC scheme similar to ours in that it uses rotation measured from the IMU and estimates the relative translation between 3D features observed from a RGB-D camera as the RANSAC model. Our work extends theirs by formulating how this 1-point RANSAC scheme can be used for joint multi-camera outlier rejection. We also characterize uncertainty in both stereo triangulation and camera extrinsics as part of our RANSAC.

III. Problem Formulation

The goal of this paper is to develop a framework to robustly select features in multi-stereo systems, and to use those features effectively in an indirect VIO pipeline. For each frame, we define a set of camera measurements $\mathcal{O}_t$ that contains all the measurements across the K stereo pairs:

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

$\mathcal{O}_t^j$ is the subset of $\mathcal{O}_t$ containing the measurements observed in stereo pair j. We denote stereo pair i as S_i and the extrinsics of the left camera with reference to the body frame as $\mathbf{T}_{S_i}^B$ or equivalently $\mathbf{T}_{S_i}$. We denote $\mathbf{R}_t$ and $\mathbf{t}_t$

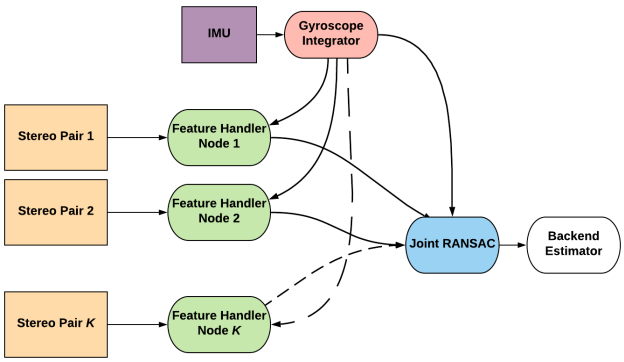

Fig. 2: The structure of the front-end for the proposed multi-stereo VIO. Features are tracked in each stereo frame by a feature handler instance and then passed to a joint RANSAC algorithm to select inliers to send to the back-end estimator.

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). 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 features, $\mathcal{O}_t$, and extract a smaller subset of features $\mathcal{C}_t$, to be added as stereo projection factors in the optimization such that each feature in $\mathcal{C}_t$ is consistent with the motion of the robot.

IV. Robust Multi-Stereo VIO

An indirect VIO system following our framework (see Figure 2) has three main steps:

1) Feature handling (temporal and stereo matching)
2) Outlier rejection
3) Back-end estimation

In this section we will briefly elaborate on Steps 1 and 3 while Section V is entirely dedicated to our proposed outlier rejection scheme.

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. During initialization we 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 [24] 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,

$$\mathcal{O}_t^j = \mathcal{C}_{t-1}^j \cup \mathcal{N}_{t-1}^j \tag{2}$$

where $\mathcal{N}_{t-1}^j$ represents the new features that were added at the previous iteration. Put simply, this states that the candidate features to be tracked at time step t are the features which were inliers at time step $t-1$ and the new features which were previously initialized. At each new image we:

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

For Step i, 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. The output of feature handler j is $\mathcal{F}_t^j$. We denote the set of temporally tracked stereo feature points across all frames at time t as $\mathcal{F}_t$:

$$\mathcal{F}_t = \bigcup_{j=1}^{K} \mathcal{F}_t^j \tag{3}$$

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 image coordinates of the features corresponding to $\mathbf{p}_L^t, \mathbf{p}_R^t$ at time $t-1$
S_i	The i-th stereo pair
$\mathbf{T}_{S_i}$	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 back-end at time step t
Multi-Camera RANSAC	
$\mathcal{F}_t$	Set of successfully temporally tracked image point pairs
$\mathbf{P}_B^t \in \mathbb{R}^3$	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
$\hat{\mathbf{\Upsilon}}_{\hat{\mathbf{p}}_L}$	Covariance matrix in image pixel space
$\hat{\mathbf{t}} \in \mathbb{R}^3$	Candidate temporal translation from RANSAC
δ	RANSAC threshold
π_j	Projection function into stereo pair j

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 $\mathbf{P}_{S_j}^{t-1} \leftarrow$ Triangulate$(\mathbf{p}_L^{t-1}, \mathbf{p}_R^{t-1})$
7 $\mathbf{P}_{S_j}^{t} \leftarrow$ Triangulate$(\mathbf{p}_L^{t}, \mathbf{p}_R^{t})$
8 $\mathbf{P}_B^t \leftarrow \mathbf{T}_{S_j}^B \mathbf{P}_{S_j}^t$
9 $\mathbf{P}_B^{t-1} \leftarrow \mathbf{T}_{S_j}^B \mathbf{P}_{S_j}^{t-1}$
10 $\mathbf{P}_B^{(t-1)'} \leftarrow \begin{bmatrix} \hat{\mathbf{R}}_t & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix} \mathbf{P}_B^{t-1}$
11 $\mathcal{X} \leftarrow \mathcal{X} \cup \{(\mathbf{P}_B^{(t-1)'}, \mathbf{P}_B^t, \mathbf{p}_L^t, \mathbf{p}_R^t, j)\}$
12 **end**
13 **end**
14 **for** $i := 1$ *to* N **do**
15 $\mathcal{I} \leftarrow \emptyset$
16 $(\hat{\mathbf{P}}_B^{(t-1)'}, \hat{\mathbf{P}}_B^t, \dots) \xleftarrow{Rand} \mathcal{X}$
17 $\hat{\mathbf{t}} \leftarrow \hat{\mathbf{P}}_B^t - \hat{\mathbf{P}}_B^{(t-1)'}$
18 **for** $(\mathbf{P}_B^{(t-1)'}, \mathbf{P}_B^t, \mathbf{p}_L^t, \mathbf{p}_R^t, j) \in \mathcal{X}$ **do**
19 $\tilde{\mathbf{P}}_B \leftarrow \mathbf{T}_{\hat{\mathbf{t}}} \mathbf{P}_B^{(t-1)'}$
20 $(\tilde{\mathbf{p}}_L^t, \boldsymbol{\Upsilon}_{\tilde{\mathbf{P}}_L}) \leftarrow \pi_j(\mathbf{T}_B^{S_j} \tilde{\mathbf{P}}_B)$
21 **if** $(||\tilde{\mathbf{p}}_L^t - \mathbf{p}_L^t||^2_{\boldsymbol{\Upsilon}_{\tilde{\mathbf{P}}_L}} < \delta)$ **then**
22 $\mathcal{I} \leftarrow \mathcal{I} \cup \{(\mathbf{p}_L^t, \mathbf{p}_R^t)\}$
23 **end**
24 **end**
25 **if** $(|\mathcal{I}| > |\mathcal{C}_t|)$ **then**
26 $\mathcal{C}_t \leftarrow \mathcal{I}$
27 **end**
28 **end**
29 **return** $\mathcal{C}_t$

B. 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:

$$x_t = [\xi_t^\top, \mathbf{v}_t^\top, \mathbf{b}_t^\top]^\top \tag{4}$$

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. Measurements associated with each frame consist of relative and marginalized IMU measurements [5, 11] between consecutive poses, as well as stereo projection factors which connect a pose and a landmark. Our back-end is based on [11] which uses a marginalization strategy that follows from Mazuran's Non-linear Factor Recovery [17] to maintain a sparse information matrix without discarding the information contained in the dense priors created by marginalization. Our back-end adds a modification to account for extrinsic uncertainty in the noise model of the projection factor. This is discussed in greater depth in Section VI.

V. Multi-Stereo RANSAC Algorithm

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 explain the method with reference to Algorithm 1.

TABLE II: RANSAC iterations for varying minimal solution points

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

We triangulate each candidate feature point in $\mathcal{F}_t$ in its respective stereo 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 estimate the temporal rotation, $\hat{\mathbf{R}}_t$ between consecutive camera frames by integrating measurements from the onboard gyroscope (line 3). Using this estimate for temporal rotation we rotate the triangulated points from the previous time step into the current time frame (line 10). At this point we 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 obtain an estimate for the temporal translation by randomly selecting a single feature correspondence and subtracting their 3D positions (line 16 and 17). Using this estimate for translation, we then project all the triangulated feature points in the previous temporal frame into the current image frame (line 20). In this step we also calculate a covariance in the pixel space of the image based on the uncertainty of the extrinsic parameters. This is described in more depth in Section VI. We perform outlier rejection by thresholding the Mahalanobis distance between the projected points in the left camera frame and the tracked points (line 21). 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:

$$N = \frac{\log(1-p)}{\log(1-(1-\epsilon)^s)} \tag{5}$$

Where N is the number of iterations needed, p is the desired probability of success, ϵ 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 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. In the proposed method, the RANSAC model only requires a single correspondence, which calls for the fewest possible number of iterations to satisfy a given confidence level.

VI. Extrinsic Uncertainty Compensation

Obtaining an accurate extrinsic calibration between multiple sensors is a significant challenge in robotics in general, and is especially difficult for systems with multiple cameras and inertial sensors. While there are several works which specifically aim to improve the quality of the calibration [6, 9, 23] some uncertainty will always remain. This remaining uncertainty can be attributed to the noisy sensor data used to perform calibration and the physical deformation

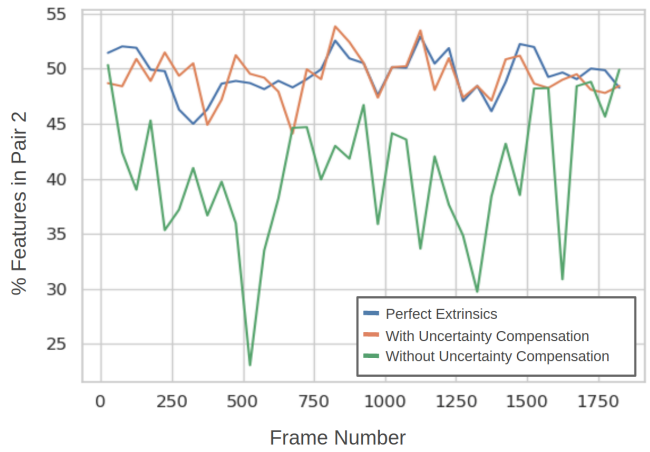

Fig. 3: This graph displays the percentage of selected inliers in *pair 2*, which has a noisy extrinsic calibration. Without compensating for uncertainty, our outlier rejection scheme has a bias towards features observed in cameras with less uncertain extrinsics. After compensating for uncertainty we see that the feature distribution more closely matches the original distribution, which are the results with perfect extrinsics.

of the camera rig that can vary with time and temperature. Knowing that it is impossible to obtain a perfect calibration, we decide to model and account for the uncertainty. Although this paper does not specifically focus on strategies to obtain a measurement of extrinsic uncertainty, it can generally be done by either extracting it from a tool which formulates calibration as the optimization of a nonlinear least squares problem or by estimating it based on the physical parameters of the camera rig.

We choose to represent the uncertainty in the extrinsics by modeling uncertainty in the transformation between the left camera of each stereo pair and the IMU. Using the same convention as [1], we represent each of these estimated transformations $\mathbf{T}_{S_i}^{B}$ as a member of the special Euclidean group $SE(3)$. We can model each transformation as some "true" transformation $\bar{\mathbf{T}}_{S_i}^{B}$ perturbed by some noise $\xi_i \in \mathbb{R}^6$ where $\xi_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{S_i}^{B})$.

$$\mathbf{T}_{S_i}^{B} = \exp(\xi_i^{\wedge})\bar{\mathbf{T}}_{S_i}^{B} \tag{6}$$

Where $\wedge$ is an overloaded operator which can either transform the noise perturbation vector $\xi_i \in \mathbb{R}^6$ to a member of the 4×4 Lie algebra $\xi_i^{\wedge} \in \mathfrak{se}(3)$ or transform a vector $\phi \in \mathbb{R}^3$ to a 3×3 member of the Lie algebra $\phi^{\wedge} \in \mathfrak{so}(3)$. Since we will be operating on the transformations in both directions it is useful to model uncertainty in both the camera-to-IMU and IMU-to-camera transformations. To do this, we follow the method described in [15]. In this section we make the conservative assumption that the uncertainty in camera-to-imu and imu-to-camera transformations are uncorrelated.

$$\bar{\mathbf{T}}_{B}^{S_i} = (\bar{\mathbf{T}}_{S_i}^{B})^{-1} \tag{7}$$

$$\mathbf{T}_{B}^{S_i} = \exp(\psi_i^{\wedge})\bar{\mathbf{T}}_{B}^{S_i} \tag{8}$$

Where $\psi_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{B}^{S_i})$ and $\mathbf{\Sigma}_{B}^{S_i} = \mathrm{Ad}_{\mathbf{T}_{B}^{S_i}} \mathbf{\Sigma}_{S_i}^{B} \mathrm{Ad}_{\mathbf{T}_{B}^{S_i}}^{\top}$

A. *Front-end*

To motivate the need for uncertainty compensation in the front-end of our proposed system we consider the example of a two stereo configuration, where *pair 1* has a very accurate camera-to-IMU calibration and *pair 2* a very noisy calibration. Points observed in each stereo frame are first triangulated in their respective camera frames and then transformed into the body frame. Inliers are determined by applying the RANSAC model and reprojecting candidate 3D feature points from the body frame back into their respective cameras. Features observed by *pair 2* will tend to have a higher reprojection error due to the two noisy transformations they underwent even if they are consistent with the motion of the robot. Without compensating for uncertainty our outlier rejection scheme will have an inherent bias toward features observed in stereo pairs with relatively stronger calibrations (illustrated in Figure 3). To address this issue, our proposed method propagates the uncertainty in the transformation through both the camera-to-IMU transformation as well as the reprojection to obtain an uncertainty in the pixel space of the image. With an uncertainty in the pixel space, we can determine inliers by setting a threshold on the Mahalanobis distance between the projected features and their actual observed locations.

We refer readers to [1] for a detailed derivation of uncertainty propagation used in this section. We start with a triangulated 3D point in one of the camera frames of the robot. It is well known that the error in triangulation is quadratic with respect to the depth of the point. We model an initial uncertainty on the triangulated 3D point by propagating the pixel noise in the image using the method described in [16]. We propagate the uncertain point through the uncertain camera-to-IMU transformation:

$$\bar{\mathbf{P}}_B = \bar{\mathbf{T}}_{S_i}^{B} \bar{\mathbf{P}}_{S_i} \tag{9}$$

If $\bar{\mathbf{P}}_B = [\mathbf{h}^{\top}, \lambda]^{\top}$ then the 4×9 Jacobian of the homogenous transformed point with respect to the parameters of both the transformation and the original point is:

$$\mathbf{J} = \begin{bmatrix} \lambda\mathbf{I}_3 & -\mathbf{h}^{\wedge} & \mathbf{R}_{S_i}^{B} \\ \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} & \mathbf{0}_{1\times3} \end{bmatrix} \tag{10}$$

We obtain a covariance matrix estimating the uncertainty of the point in the body frame using a first order approximation.

$$\mathbf{\Sigma}_{\mathbf{P}_B} = \mathbf{J} \begin{bmatrix} \mathbf{\Sigma}_{S_i}^{B} & \mathbf{0}_{6\times3} \\ \mathbf{0}_{3\times6} & \mathbf{\Sigma}_{\mathbf{P}_{S_i}} \end{bmatrix} \mathbf{J}^{\top} \tag{11}$$

We use the method described in [1] to propagate the uncertainty of the extrinsics and the uncertainty of the point in the body frame through projection into a nonlinear camera model. For each candidate motion model in RANSAC, we obtain a 3D point in the body frame, $\tilde{\mathbf{P}}_B$, which needs to be projected back into the original image to determine inliers. We model an uncertain 3D point as:

$$\tilde{\mathbf{P}}_B = \bar{\tilde{\mathbf{P}}}_B + \mathbf{D}\zeta \tag{12}$$

where $\zeta \in \mathbb{R}^3$ and $\zeta \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\tilde{\mathbf{P}}_B})$ and $\mathbf{D}$ is the 4×3 matrix defined by $\mathbf{D} = [\mathbf{I}_3, \mathbf{0}]^{\top}$.

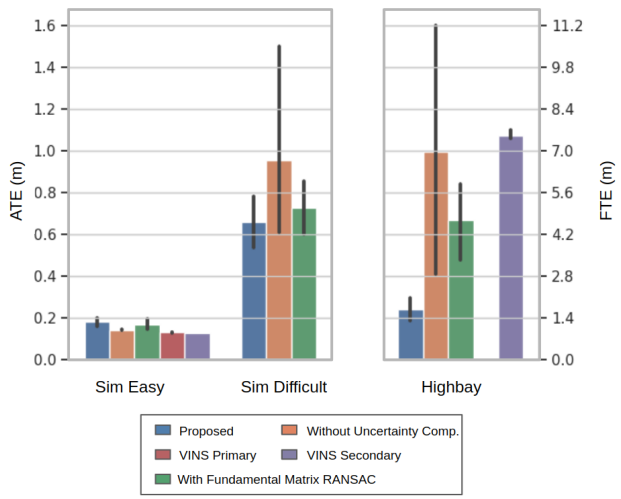

Fig. 4: Average Trajectory Error (ATE) on simulated data and Final Trajectory Error (FTE) on Highbay data. If bars are not shown it means the VIO algorithm failed to obtain a state estimate. Each trial was run 5 times. Median value is shown in graph with error bars representing the range of data.

The projection function $\pi : \mathbb{R}^4 \rightarrow \mathbb{R}^2$ is a nonlinear function which takes a homogeneous 3D point in the left camera frame and projects it to an image pixel. We define the Jacobian of the projection function with respect to the homogenous points in the camera frame as

$$\mathbf{\Pi} = \frac{\partial \pi}{\partial \mathbf{w}}\bigg|_{w} = \begin{bmatrix} \frac{f_x}{w_3} & 0 & -\frac{f_x w_1}{{w_3}^2} & 0 \\ 0 & \frac{f_y}{w_3} & -\frac{f_y w_2}{{w_3}^2} & 0 \end{bmatrix} \tag{13}$$

Similarly to Eq. 10 we define $\mathbf{G}$ as the Jacobian of the transformed homogenous point with respect to the parameters of the transformation and original point. In this case the transformation we are considering is from the body frame back to the camera frame.

$$\mathbf{G} = \begin{bmatrix} \lambda \mathbf{I}_3 & -\mathbf{d}^{\wedge} & \mathbf{R}_B^{S_i} \\ \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} & \mathbf{0}_{1\times 3} \end{bmatrix} \tag{14}$$

where $\tilde{\mathbf{P}}'_{S_i} = \mathbf{T}_B^{S_i} \tilde{\mathbf{P}}_B = [\mathbf{d}^\top, \lambda]^\top$. We make the important distinction that $\mathbf{P}_{S_i}$ is the original triangulated point in the camera frame, while $\mathbf{P}'_{S_i}$ is the point after it has been transformed back into the camera frame as part of the RANSAC scheme. From here we define the Jacobian of the entire reprojection function, from the point in the body frame to a pixel in the image as:

$$\mathbf{H} = \mathbf{\Pi}\mathbf{G} \tag{15}$$

and the covariance in the image space as:

$$\mathbf{\Upsilon}_{\tilde{\mathbf{p}}_L} = \mathbf{H}\mathbf{\Xi}\mathbf{H}^\top, \quad \mathbf{\Xi} = \begin{bmatrix} \mathbf{\Sigma}_B^{S_i} & \mathbf{0}_{6\times 3} \\ \mathbf{0}_{3\times 6} & \mathbf{\Sigma}_{\tilde{\mathbf{P}}_B} \end{bmatrix} \tag{16}$$

B. Back-end

By compensating for uncertainty in our outlier rejection scheme we end up selecting a greater number of features observed in stereo pairs with relatively weaker extrinsic calibrations. Although these features still contain valuable information that can constrain the state estimate of a robot,

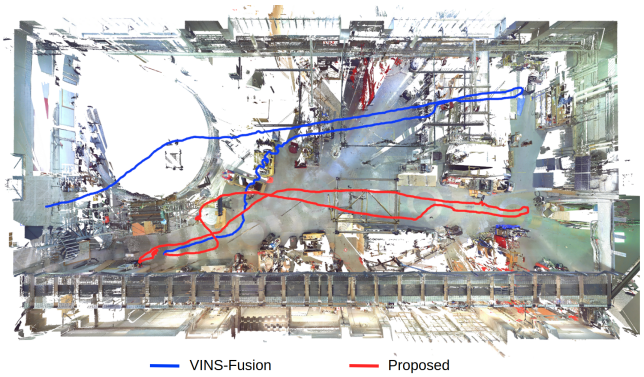

Fig. 5: Visualized trajectories of the proposed algorithm and VINS-Fusion running on the forward facing stereo pair. The trajectory was manually registered against the reconstructed scene to aid in visualization. We can see that the trajectory produced by our proposed algorithm is able to return to the original starting location while VINS-Fusion drifts significantly.

TABLE III: Average Computational Time for Outlier Rejection

	Proposed	Proposed without Uncertainty Compensation
Sim Easy	0.0138 s	0.0435 s
Sim Difficult	0.0118 s	0.0418 s
Highbay	0.00774 s	0.0418 s

the uncertainty in the extrinsics adds noise to the measurements which needs to be accounted for. In our proposed method, the noise model of each stereo projection factor in the back-end factor graph is modified to model the noise in the extrinsics. Our final noise model accounts for both the uncertainty in the camera model as well as uncertainty in extrinsic calibration of each stereo pair. Specifically, we modify Eq. 4 in [11] such that the residual associated with each projection factor is weighted by $\mathbf{\Sigma}_m$ rather than $\mathbf{\Sigma}_{c_{ij}}$:

$$\mathbf{\Sigma}_m = \mathbf{\Sigma}_{c_{ij}} + \mathbf{\Upsilon}_{\mathbf{p}_L^*} \tag{17}$$

where $\mathbf{\Sigma}_{c_{ij}}$ is the noise model of the camera and $\mathbf{\Upsilon}_{\mathbf{p}_L^*}$ is the covariance of the given landmark projected from the current linearization point of the optimization into the left camera. Since we are projecting an estimate of the 3D landmark position we have $\mathbf{\Sigma}_{\mathbf{P}_B^*} = \mathbf{0}$.

VII. Experimental Results

The proposed multi-camera VIO pipeline was evaluated using a MAV in a simulated Gazebo environment as well as on real world data collected in the Robotics Institute's Highbay. The simulated environment allowed us to obtain ground truth extrinsics and manually add noise to test our system, while the Highbay data allowed us to verify the robustness of our algorithm on real data. For both types of experiments we evaluate the accuracy of the trajectory against VINS-Fusion running on both stereo pairs individually. VINS-Fusion was run with default parameters and the option to optimize for extrinsics enabled. As a comparison we also compare against a version of our full VIO pipeline which uses a fundamental matrix RANSAC for outlier rejection on each stereo pair individually. For the simulated data the primary stereo pair was facing forwards and the secondary pair was facing backwards. For the Highbay data, the primary stereo pair was

TABLE IV: ATE in Simulated Environments

	Proposed	Proposed without Uncertainty Compensation	Proposed with Fundamental Matrix RANSAC	VINS-Fusion Primary	VINS-Fusion Secondary
Easy	0.179 m	0.144 m	0.167 m	0.131 m	**0.128 m**
Difficult	**0.791 m**	0.954 m	0.810 m	Failed	7.510 m

TABLE V: FTE in Field Robotics Center Highbay

	Proposed	Proposed without Uncertainty Compensation	Proposed with Fundamental Matrix RANSAC	VINS-Fusion Primary	VINS-Fusion Secondary
Highbay	**1.694 m**	6.980 m	5.217 m	Failed	Failed

facing forwards the secondary pair was facing downwards. We also compare against our proposed algorithm without uncertainty compensation in order to precisely observe the effects of uncertainty compensation. Each reported result is the median over 5 trials. Errors are plotted in Figure 4 and reported in Tables IV and V. Failure is defined as an error over 10% of the length of the trajectory. To demonstrate the computational benefit of our 1-point RANSAC formulation, a comparison of computational time required to run the proposed outlier rejection scheme and the fundamental matrix RANSAC is shown in Table III.

A. Simulated Results

The simulated data was recorded on a MAV in an outdoor *Gazebo* environment. Noise was randomly added to the extrinsics of each stereo pair. The results of two simulated flights were recorded. One flight was a relatively easy trajectory with no aggressive motion or sudden scene occlusions. Another simulated flights included aggressive turns and flight very close to obstacles which could partially occlude the fields of view of the cameras on board. Images and IMU measurements from the simulator were taken as inputs to the VIO pipeline. All methods are able to achieve a similarly low ATE on the easy flight. The main benefit of our proposed system is apparent in the difficult dataset.

B. Highbay Data

Our real world data was collected using a two stereo camera rig with time synchronized images and IMU data, seen in Figure 6. The multi-camera rig was moved around the Highbay and returned precisely to its original starting position. To test the robustness of our VIO the data was intentionally made to be extremely challenging, with several points of sudden occlusion occurring during the run. We evaluated each algorithm by returning to the same starting location and measuring Final Trajectory Error (FTE), which is the absolute drift of the final position. A visualization of two trajectories is shown in Figure 5 for a reference of scale.

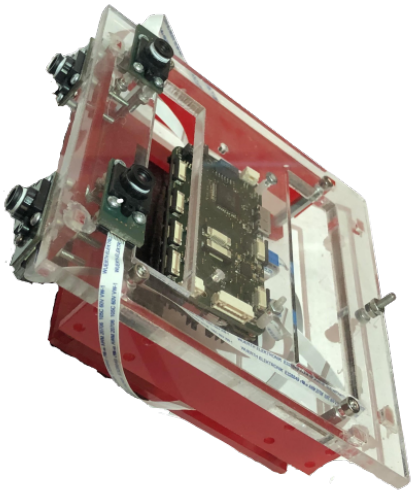

Fig. 6: The multi-stereo camera rig used to collect experimental results. Images were captured at 25 Hz and inertial data was collected at 200 Hz. All cameras were synchronized using the on-board FPGA.

VIII. Conclusion

In this paper we have introduced a novel application of a 1-point RANSAC algorithm which is used as part of a multi-stereo VIO pipeline. Our outlier rejection scheme operates on stereo triangulated points, which allows us to formulate a 1-point minimal solution. 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. We address the issue of noisy calibration by compensating for extrinsic uncertainty in both the front-end and back-end of our proposed algorithm. 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 stereo pairs individually in a simulated environment. We also demonstrate that compensating for extrinsic uncertainty improves the accuracy and robustness of the VIO's state estimate.

References

[1] T. G. Barfoot and P. T. Furgale, "Associating uncertainty with three-dimensional poses for use in estimation problems," *IEEE Trans. on Robotics (TRO)*, vol. 30, no. 3, pp. 679–693, 2014.

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

[3] 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.

[4] 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.

[5] 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.

[6] P. Furgale, T. D. Barfoot, and G. Sibley, "Unified temporal and spatial calibration for multi-sensor systems," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013, pp. 1280–1286.

[7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. New York, NY, USA: Cambridge University Press, 2003.

[8] L. Heng, G. H. Lee, F. Fraundorfer, and M. Pollefeys, "Real-time photo-realistic 3d mapping for micro aerial vehicles," in *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, San Francisco, CA, USA, Sep. 2011, pp. 4012–4019.

[9] 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.

[10] 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.

[11] 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.

[12] J. Jaekel and M. Kaess, "Robust multi-stereo visual-inertial odometry," *"IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS). Workshop on Visual-Inertial Navigation: Challenges and Applications"*, Nov. 2019.

[13] 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.

[14] 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.

[15] J. G. Mangelson, M. Ghaffari, R. Vasudevan, and R. M. Eustice, "Characterizing the uncertainty of jointly distributed poses in the lie algebra," vol. abs/1906.07795, 2019. [Online]. Available: http://arxiv.org/abs/1906.07795

[16] L. Matthies and S. A. Shafer, "Error modeling in stereo navigation," *IEEE Journal of Robotics and Automation*, vol. 3, pp. 239 – 248, Jun. 1987.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Stockholm, Sweden, 2013, pp. 4304–4311.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.