Comparison of Two Image and Inertial Sensor Fusion Techniques for Navigation in Unmapped Environments

CLARK N. TAYLOR, Senior Member, IEEE
Brigham Young University

MICHAEL J. VETH, Senior Member, IEEE
JOHN F. RAQUET, Member, IEEE
Air Force Institute of Technology

MIKEL M. MILLER, Member, IEEE
Air Force Research Laboratory

To enable navigation of miniature aerial vehicles (MAVs) with a low-quality inertial measurement unit (IMU), external sensors are typically fused with the information generated by the low-quality IMU. Most commercial systems for MAVs currently fuse GPS measurements with IMU information to navigate the MAV. However there are many scenarios in which an MAV might prove useful, but GPS is not available (e.g., indoors, urban terrain, etc.). Therefore several approaches have recently been introduced that couple information from an IMU with visual information (usually captured by an electro-optical camera). In general the methods for fusing visual information with an IMU utilizes one of two techniques: 1) applying rigid body constraints on where landmarks should appear in a set of two images (constraint-based fusion) or 2) simultaneously estimating the location of features that are observed by the camera (mapping) and the location of the camera (simultaneous localization and mapping—SLAM-based fusion). While each technique has some nuances associated with its implementation in a true MAV environment (i.e., computational requirements, real-time implementation, feature tracking, etc.), this paper focuses solely on answering the question "Which fusion technique (constraint- or SLAM-based) enables more accurate long-term MAV navigation?" To answer this question, specific implementations of a constraint- and SLAM-based fusion technique, with novel modifications for improved results on MAVs, are described. A basic simulation environment is used to perform a comparison of the constraint- and SLAM-based fusion methods. We demonstrate the superiority of SLAM-based techniques in specific MAV flight scenarios and discuss the relative weaknesses and strengths of each fusion approach.

Manuscript received April 15, 2008; revised January 15, 2009; released for publication November 13, 2009.

IEEE Log No. T-AES/47/2/940823.

Refereeing of this contribution was handled by D. Gebre-Egziabher.

This work is funded by an AFOSR Young Investigator Award FA9550-07-1-0167 and an ASEE/Air Force Summer Faculty Fellowship.

Authors' addresses: C. N. Taylor, Sensors Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, OH 45433, E-mail: (clark.n.taylor@gmail.com); M. J. Veth and J. F. Raquet, Air Force Institute of Technology, Wright-Patterson Air Force Base, OH; M. M. Miller, Air Force Research Laboratory, Munitions Directorate, Eglin Air Force Base, FL.

I. INTRODUCTION

Recently unmanned aerial vehicles (UAVs) have seen a dramatic increase in utilization for military applications. In addition, UAVs are being investigated for multiple civilian uses, including rural search and rescue, forest fire monitoring, and agricultural information gathering [1–3]. Due to their small size, miniature UAVs (MAVs) are an attractive platform for executing many of these missions. Some of the primary advantages of MAVs include 1) they are significantly less expensive to purchase than the large UAVs typically used by the military; 2) their small size simplifies transport, launch, and retrieval; and 3) they are less expensive to operate than large UAVs.

Accurate navigation state (location, attitude, and velocity) estimation is essential for many MAV missions. For example, if an MAV is being utilized as part of a search and rescue operation, knowing the correct location and attitude of the MAV is critical to geolocate an observed object. Navigation methods implemented in MAVs today are primarily based on the fusion of measurements from the Global Positioning System (GPS) and the inertial measurement unit (IMU) [4–7]. However there are many scenarios in which an MAV might prove useful, but GPS is not available (e.g., indoors, urban terrain, etc.). When flying without GPS, large UAVs can perform navigation using a high-quality IMU. However, the size and weight constraints of an MAV dictate the use of lightweight and therefore inaccurate sensors on their IMU. These low-quality sensors cause significant drift in navigation state estimates if used alone in a navigation system. Therefore accurate MAV navigation requires fusion of low-quality IMU data with other sensors to decrease the long-term navigation error.

Several methods have been proposed for fusing visual information with IMU measurements to enable the navigation of MAVs without GPS [8–15]. Visual sensors (i.e., electro-optical cameras) possess a number of advantages for use on an MAV: the sensors themselves are lightweight and low power, most MAV systems already have an onboard camera, and a single vision sensor is capable of detecting motion in five of the navigation state parameters simultaneously (i.e., 3 velocity values and 3 attitude angles—less a scale ambiguity in the velocity). In addition, cameras can be used in areas that may not receive the GPS radio signals required for GPS-based navigation (e.g., urban terrain, indoors, GPS-jamming environment, etc.).

One approach to MAV navigation using visual information is to find and track prestored landmarks (e.g., [16, 17]). Prior to the flight of the MAVs, landmarks and location in the world can be found through the projected path of the MAV and can be stored in a database. During flight of the MAV, features observed by the camera onboard the MAV are

compared with the landmarks in the database. When matches are found, the position and attitude of the MAV with respect to the landmarks can be estimated, yielding a long-term navigation solution.

Despite the advantages and high performance of the prestored landmarks approach, it requires that landmarks be found, localized, and placed in a database prior to flight of the MAV. However, many usage scenarios for MAVs involve the MAV exploring areas that have not been previously observed or where significant changes may have occurred. Therefore in this paper we focus on methods that can be used when the MAV is entering an unmapped area (i.e., one that does not have a prestored database of landmarks).

Methods for fusing visual information with IMU data that do not require prior knowledge of landmarks can be divided into two main categories. First, constraints on where a static object being imaged by a moving camera appears in two separate images can be used to refine the navigation estimates returned by the IMU. Second the location and attitude of the camera and the location of objects observed by the camera can be estimated simultaneously, leading to the simultaneous localization and mapping (SLAM) problem. While SLAM-based methods are highly effective, there are two bottlenecks to SLAM that make it difficult to implement in the computationally limited environments that characterize MAVs. First visual SLAM requires that objects in the video be tracked for an extended period of time. Second the size of the state grows with the number of landmarks that SLAM is attempting to find the location for, dramatically increasing the computation time required. The primary contribution of this paper is an understanding of the benefits incurred from this increased cost in complexity for SLAM-based methods compared to constraint-based methods. Note that there has been a significant amount of work describing the theoretical and practical limitations of SLAM-based algorithms [18–20]. As far as we know, however, this is the first work that directly compares these two fundamentally distinct approaches to visual/IMU sensor fusion.

While a significant amount of research has been published in both SLAM-based and constraint-based IMU/vision fusion (including works demonstrating their efficacy with real video and IMU information), there are some significant difficulties in comparing these results. Some of the difficulties in comparing results include the following.

1) How to identify a feature across multiple images (the correspondence or data association problem) differs greatly from implementation to implementation. The results of any algorithm are greatly dependent on how well features are associated over time, making direct comparison of published results difficult.

2) Several different options exist for the "fusion engine" used in these approached. Possible options include 1) the extended Kalman filter (EKF) [21], 2) the unscented Kalman filter (UKF) [22], 3) the information filter (IF) [21], or 4) the particle filter (PF) [23]. The exact methodology used to implement each of these engines can also differ greatly from published work to published work.

3) The specific camera and IMU setup used to get final results differs greatly between implementations.

4) What noise is present in each of the sensors can vary greatly from implementation to implementation, yielding significantly different final results.

5) When fusing data from multiple sensors, the methods used to synchronize their timing and precisely align their geometric orientations with respect to each other can cause significantly different results to be reported from the same basic fusion techniques.

Because of these and several other small differences in implementations between published works, it is impossible to perform a direct comparison between previously published fusion techniques.

To perform a direct comparison between fusion techniques, we assumed three basic similarities between the environments for the techniques: 1) perfect data association across images, 2) an UKF framework [22] is used as the basic filtering technique in both approaches, and 3) the same noise is injected into the system with both scenarios. With these ambiguities removed from the environment, we perform a comparison between navigation approaches, rather than navigation implementations in this paper (the primary contribution of this paper). The secondary contributions of this paper are some proposed modifications to the constraint-based approach for visual/inertial fusion for MAV environments.

The remainder of this paper is organized as follows. Section II describes the constraint-based approach that we implemented with a discussion of the novel modifications made to the algorithm to enable MAV navigation. Section III describes our SLAM-based approach to navigation for MAVs. Section IV presents our comparison between different navigation approaches. Section V concludes the paper.

II. CONSTRAINT-BASED FUSION

When performing navigation of an MAV, the primary goal is to minimize the error in the navigation state estimates over time. The fundamental concept of IMU-based navigation approaches is to integrate the inertial accelerometer and gyroscope inputs to maintain an estimate of navigation state over time. The weakness of this method is that because the noise from the inertial sensors is integrated over time, the error growth in navigation estimates can be very rapid.

Therefore when additional information can be fused together with inertial sensors, the growth in error over time can be decreased.

One method for fusing visual and IMU data together is to use the "epipolar" constraint between two images. The epipolar constraint is used when a single object is observed by a camera at two locations (or two cameras at different locations). As any three points define a plane, the observed point and the two camera projection centers form a plane in the 3-D world. Similarly if a point is observed in two images, the two vectors representing where the point was observed in the image plane ($\vec{x}'$ and $\vec{x}$) and the translation vector between the two camera locations, should all be in the same plane. Mathematically this is represented by the formula

$$\vec{x}'[\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}\vec{x} = 0 \tag{1}$$

where $\mathbf{C}_{c1}^{c2}$ is the direction cosine matrix between the two cameras, and $[\vec{p}_{c1}^{c2}]_\times$ is the position of camera 1 in the coordinate frame of camera 2, put into a skew-symmetric matrix. (In other words, we calculate the cross product between $\vec{p}_{c1}^{c2}$ and $\mathbf{C}_{c1}^{c2}\vec{x}$.)

Using the epipolar constraint, it is possible to "measure" the rotation and the direction of translation between two camera poses from the feature locations in two images. Several approaches in the literature [8, 9, 14, 24] have utilized the epipolar constraint to fuse IMU and visual measurements together. To directly include the epipolar constraint in a Kalman filter, we use the method presented in [24] with some modifications for use in an MAV environment.

To utilize the epipolar constraint, the state must contain enough information to generate $\mathbf{C}_{c1}^{c2}$ and $\vec{p}_{c1}^{c2}$. To represent general motion with an MAV between two times, the UKF state

$$\vec{X} = \begin{bmatrix} p_t \\ q_t \\ p_{t-1} \\ q_{t-1} \\ v_t \\ b \end{bmatrix} \tag{2}$$

was used, where p_t is the position estimate at time t, including three elements (p_n, p_e, $-h$), q_t is a quaternion-based representation of the attitude at time t, v_t is the velocity of the MAV at time t, and b is the estimated biases on the accelerometers and gyros of the IMU. This method for setting up the UKF with two navigation state estimates was first introduced in [9] where its efficacy was demonstrated with real MAV video and IMU readings.

A. Performing the Time Update

The time update for this UKF implementation takes two different forms. The first form updates the current location, velocity, and attitude estimates every time an IMU measurement occurs. The second type of update occurs after each measurement update is as follows:

$$\vec{X}^+ = \mathbf{A}\vec{X}^- \qquad \text{where}$$
$$\mathbf{A} = \begin{bmatrix} I_7 & 0 & 0 \\ I_7 & 0 & 0 \\ 0 & 0 & I_{10} \end{bmatrix} \tag{3}$$

where I_n is an $n \times n$ sized identity matrix. This second update enables the Kalman filter to maintain the two most recent navigation state estimates at all times in its state vector.

B. Performing the Measurement Update

Assuming a feature has been detected in two images, the locations of the features are represented by $\vec{x}'$ and $\vec{x}$. In [24], Soatto et al. describes how the epipolar constraint can be used as a direct measurement in a Kalman Filter. Because the epipolar constraint should always be equal to zero, the measurement used by the UKF is a vector of zeros in length equal to the number of corresponding features found between the two images. The predicted measurement is $\vec{x}'[\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}\vec{x}$ for each set of features $\vec{x}'$ and $\vec{x}$, where $\vec{p}_{c1}^{c2}$ and $\mathbf{C}_{c1}^{c2}$ are derived from the current state $\vec{X}$. Following the measurement step, it is important that the quaternion be normalized to maintain a valid representation of rotation. This is performed using the method outlined in [25].

While this method for fusing epipolar information with the state yields good results in most scenarios, there are two underlying assumptions to this method that are often violated when using video typical of MAVs. These two assumptions are 1) that the translation of the camera causes a significant movement of the features in the image and 2) that the magnitude of translation is known precisely. In the sections below, we discuss how we overcame these problems to improve navigation on MAVs using the constraint-based approach.

C. Overcoming the Significant Translation Assumption

While including the epipolar constraint as a measurement in a UKF always improves the accuracy of the attitude estimates over time, we found that in many of our simulated cases, the location error was actually worse with the constraint-based approach than in the IMU-only case. To understand the source of this error, let us analyze the epipolar constraint expressed in (1). The value $x'[\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}x$ can be rewritten as a cross product of two vectors followed by a dot product of two vectors. The final magnitude of this computation will be $\|x'\| \, \|\vec{p}_{c1}^{c2}\| \, \|x\| \sin(\theta_{\vec{p}_{c1}^{c2} \to \mathbf{C}_{c1}^{c2}x}) \cos(\theta_{x' \to [\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}x})$, where $\theta_{\vec{p}_{c1}^{c2} \to \mathbf{C}_{c1}^{c2}x}$ is the angle between $\vec{p}_{c1}^{c2}$, and $\mathbf{C}_{c1}^{c2}x$ and $\theta_{x' \to [\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}x}$ is the angle between x' and $[\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2}x$.

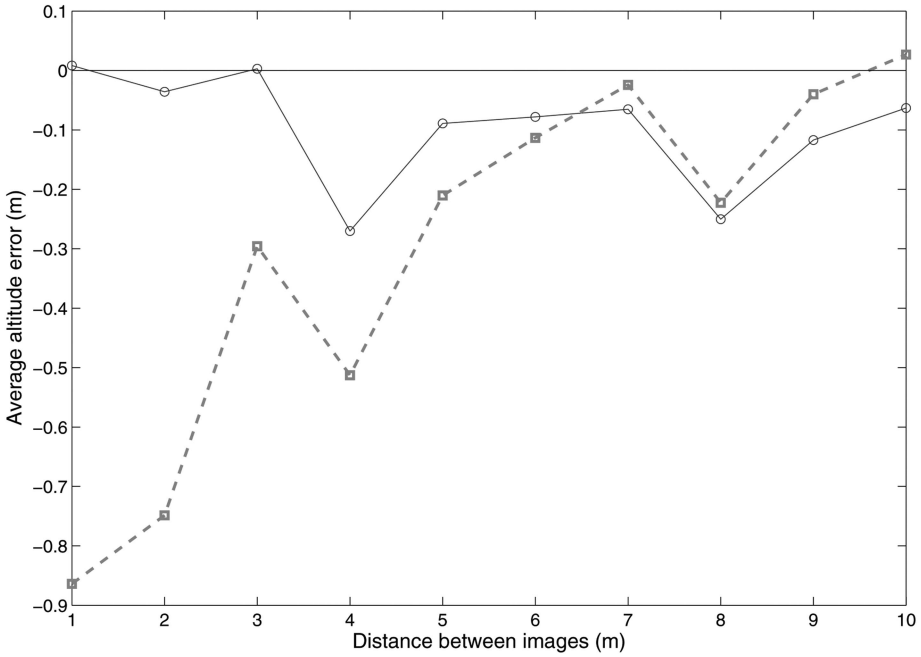

Fig. 1. Figure demonstrates bias inherent in traditional constraint-based fusion techniques. When distance to features is large (~ 100 m) compared to translation direction (x axis), error in estimated altitude (y axis) can be significantly biased. Proposed "sin out" method helps to reduce bias in traditional techniques (solid line with circles).

As mentioned previously, the correct magnitude for the total calculation is 0. However, there are two conditions under which $\|x'\| \, \|\vec{p}_{c1}^{c2}\| \, \|x\| \sin(\theta_{\vec{p}_{c1}^{c2} \rightarrow \mathbf{C}_{c1}^{c2} x}) \cos(\theta_{x' \rightarrow [\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2} x}) = 0$: 1) when $\theta_{x' \rightarrow [\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2} x} = 90^\circ$ or 2) when $\theta_{\vec{p}_{c1}^{c2} \rightarrow \mathbf{C}_{c1}^{c2} x} = 0$. When attempting to meet the epipolar constraint, the first condition corresponds with verifying that three distinct vectors lie on a plane, the basis for the epipolar constraint. The second condition, however, is a degeneracy when two of the vectors point in the same direction. Because of this second condition, the UKF measurement update may try to "push" the translation direction ($\vec{p}_{c1}^{c2}$) to be in the same direction as the observed points ($\mathbf{C}_{c1}^{c2} x$), an undesired effect. With multiple points in the image, this pushes the translation direction toward the center of the image points.

When the magnitude of the translation vector is sufficiently large, the first condition's effect far outweighs the second condition. When utilizing MAV video, however, the translation magnitude (the movement of the MAV in 1/10th of a second) will often be small in relation to the distance of the MAV from the ground. Therefore, this improper biasing of the translation direction to the middle of the image points causes a serious degradation in results.

To overcome this biasing the measurement step of the UKF was modified to eliminate the $\sin(\theta_{\vec{p}_{c1}^{c2} \rightarrow \mathbf{C}_{c1}^{c2} x})$ term from the measurement. To eliminate the effect of $\sin(\theta_{\vec{p}_{c1}^{c2} \rightarrow \mathbf{C}_{c1}^{c2} x})$, the term $[\vec{p}_{c1}^{c2}]_\times \mathbf{C}_{c1}^{c2} x$ is first computed and then normalized to be of length one. The inner product of this term with x' is then taken and returned as the predicted measurement. Note that this method may not work when the camera is aligned with the direction of translation.

In Fig. 1, we plot the average altitude error after 1 s of imaging/inertial fusion using both the traditional epipolar constraint and the proposed constraint, which removes the sin term. In these simulations, the MAV was flying at an altitude of 100 m with the camera pointed at features on the ground. Therefore, when a bias of direction toward the feature locations is present, the estimated altitude will decrease with time. In Fig. 1, we plot the results for different lengths of translation between camera locations (the x axis). The plotted results are an average of 100 runs of the simulation environment. Note that the traditional method (the dashed plot with squares) has a significant downward bias when the translation distance is small. However, the proposed "sin out" method does not have a downward bias even at small magnitudes of translation. This plot demonstrates the necessity of using the modified epipolar constraint proposed in this paper as many MAV flights will have a small translation distance between cameras when compared with the distance to observed objects.

D. Overcoming the Known Translation Magnitude Assumption

Another weakness of using the epipolar constraint as a measurement is that the magnitude of the translation has no effect on the final measurement. Therefore the velocity magnitude is unobservable, though the direction of velocity is measurable.

TABLE I
Results Demonstrating the Effects of Making our Modifications for MAV-Captured Video

	Truth	IMU-only	Unmodified	sin Removed	Translation Magnitude Corrected
T_x (μ,σ)	100	(115.8,318.9)	(−137.9,161.3)	(205.1,57.2)	(109.2,83.2)
T_y (μ,σ)	0	(−46.3,354.1)	(1.74,17.1)	(−0.63,9.3)	(−0.27,7.0)
T_z (μ,σ)	−100	(578.8,329.3)	(95.5,84.8)	(−72.0,14.4)	(−83.1,19.6)

Note: By correcting for the bias in both translation direction and magnitude in the epipolar constraint we achieve significantly more accurate results.

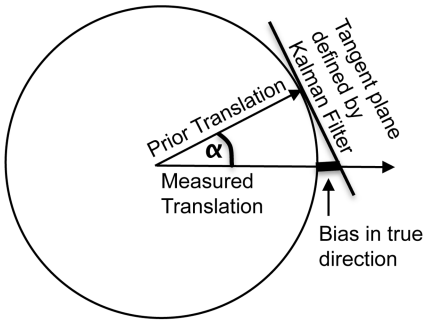

Fig. 2. Simple example of why translation magnitude is biased in direction of true translation when performing constraint-based fusion. Measurement step of Kalman Filter combines prior estimated direction with measured direction using a tangent plane approximation. The tangent plane, however, will always increase the magnitude of the resulting translation vector.

Unfortunately this unobservability of velocity magnitude leads to a positive bias in velocity magnitude, in the true direction of velocity. To understand this effect, consider Fig. 2 where a 2-D example of how the Kalman filter will merge translation directions together is shown. Due to the scale ambiguity, the measurement step can only correct the direction of translation, not the magnitude. However when a derivative is taken of the translation direction, it forms a tangent plane that will always lead to an increased magnitude of the resulting vector. Because the increase in translation leads to an increase in velocity and because the IMU only measures accelerations, these increases lead to an exponential increase in error in the true direction of travel.

To overcome this weakness in the epipolar constraint fusion method, we used a Kalman-filter type update on the magnitude of velocity after each measurement step. To determine the "update" that should be applied, a discussion of what errors in the velocity may occur is required. Errors in velocity can be broken into two components, errors in the direction of the true velocity, and errors in orthogonal directions (see Fig. 2). After the measurement step of the UKF is completed, we assume that the direction of the translation is correct. Therefore we would like to set the magnitude of the post-measurement translation equal to the magnitude of the pre-measurement magnitude in the direction of the post-measurement translation.

A simplistic approach to modifying the velocity would be to simply change the velocity after each measurement step (v^+) to the velocity magnitude that was present before the measurement step (v^-). However simply changing the velocity magnitude will not reverse errors introduced in the accelerometer bias estimates due to coupling between the velocity estimate and the biases. Therefore a "virtual measurement" is used to correct the change in velocity magnitude and its effects on IMU bias estimates. This virtual measurement Δv is found as

$$\Delta v = (k-1) * v^+ \qquad \text{where} \tag{4}$$

$$k = \frac{v^{-T} v^+}{v^{+T} v^+} \tag{5}$$

which sets the magnitude of v^+ equal to the magnitude of v^- projected in the direction of v^+. This virtual measurement is used as a quasi-Kalman measurement update as

$$\begin{aligned} \vec{X}^+ &= \vec{X}^- + K\Delta v \qquad \text{where} \\ K &= PH^T(HPH^T)^{-1} \\ H &= [0 \quad I_3 \quad 0]. \end{aligned} \tag{6}$$

Note that the covariance matrix is not modified in this step as no real measurement has occurred. Using this method the measurement step of the Kalman filter does not significantly alter the velocity magnitude or the accelerometer biases in the velocity direction.

E. Results of MAV-based Modifications

In Table I, we present some results validating the efficacy of our modifications of constraint-based navigation of MAV-based scenarios. We performed 100 simulations of a straight-line flight in the x direction (no movement in y or z) that lasts 15 s. The three rows of the table represent the final estimated location in the x, y, and z location. The columns represent 1) what the true location of the camera was, 2) the mean and standard deviation of the estimates using the IMU only, 3) results if the unmodified epipolar constraint was used, 4) results when the sin bias is removed (per Section IIC), and 5) results after correcting the velocity magnitude following each measurement update. The results after removing the sin bias are significantly improved results in the y and z direction, and the mean value in the x direction is

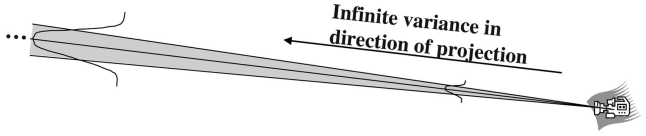

Fig. 3. Figure demonstrates extremely nonlinear variance resulting from single camera observing point. First there is infinite variance in direction of observation. Second as distance from camera increases, variance in orthogonal directions also increase.

much closer to the truth when the virtual measurement of velocity is applied.

III. SLAM-BASED FUSION

The second type of fusion technique that we evaluate is based on previous simultaneous localization and mapping (SLAM) work. The general concept of SLAM is that a set of world points should be observed (mapped) and localized over time while simultaneously determining the current location of the sensor (localization). By maintaining a full covariance matrix between the location of the sensor and the location of the world points, the SLAM algorithm has been shown to be convergent in specific circumstances [26]. Therefore, SLAM has been an area of intensive research for the past several years. Most SLAM methods, however, have concentrated on utilizing sonar- or laser-based sensor for measurements of the world as opposed to a camera. The subset of SLAM work that has focused on using a camera is generally called bearing-only or visual SLAM. In general, however, these methods have assumed that a camera is the only input, and have not attempted to fuse in IMU data while performing SLAM (although there have been exceptions, [10, 27]).

For our implementation of SLAM, a UKF was once again used, helping to make the comparison between SLAM-based and constraint-based techniques equitable. The state of our UKF includes the current navigation state of the camera (p, q, v), the biases of the IMU (b), followed by entries for the world points that are being tracked. One of the primary difficulties with visual SLAM is the initialization of new features. Because only the bearing of a world object from the camera is observed the first time, the 3-D location of the world point cannot be added directly to UKF state for two reasons. First there is essentially infinite variance in one direction if only one observation has occurred (see Fig. 3). Second because there may be error in the observed bearing, the variance in the directions orthogonal to the infinite variance direction grow with distance from the camera, causing the initial observation of the point to have an extremely nonlinear variance.

Several attempts to overcome this initialization problem have been utilized previously. In [11], a stereo camera system was used, which eliminates the infinite variance problem for world features that are "close enough" to the stereo camera pair. Because the small size of MAVs severely limits the baseline that can be placed between stereo cameras, however, this many not be a feasible option. Therefore the authors in [12] initialized world features observed from an aircraft by assuming points were on the ground and intersecting the ray from the camera with a model of the ground elevation. However with an MAV that may be used to fly through urban terrain, the assumption of points on the ground may not be valid. In [28], targets of known size were placed on the ground, and the size of the target was used to initialize the depth of the target from the camera. This method is not applicable to our work as we assume world points that are not known a priori.

To provide an initialization methodology that can be used with a single camera and does not presuppose information about the target, we use the method introduced in [29] to initialize feature locations. This method, the "inverse depth" method, places three new items in the Kalman Filter state for each new object that is observed, namely, 1) the projective center of the camera when the feature was first observed (represented by P_c), 2) the direction of the ray (in 3-space) of the first observation (represented by two angles, θ and ϕ), and 3) the inverse depth of the point along that line (represented by ξ). The inverse depth rather than the depth is used because a finite value between 0 and 1 covers the entire depth from 1 to infinity in a linear fashion. Therefore a variance of 0.5 on the inverse depth is equivalent to an infinite variance when using depth. In addition because the bearing of the original observation is included in the Kalman Filter state, the "increasing variance with distance" problem discussed above is properly represented in the Kalman Filter.

While the inverse depth methodology has many advantages, it has a significant failing when it comes to MAV video. In [29], the direction of the ray that first observed the point to be localized is represented by two angles, θ and ϕ. This works very well for the situation outlined in [29], where the camera is generally facing forward. However in an MAV where the camera may be facing down, this representation has a significant shortcoming. Assume two rays, both pointing almost straight down, but with one facing slightly forward and one slightly to the side. Using the two angles ϕ and θ to represent these locations, the rays will be very far apart (i.e., $\phi = 90°$, $\theta = 0°$, and $\theta = 90°$) Therefore there is a singularity in the two angle representations around the downward direction. To remove this singularity, we replaced the 2-angle representation with a 3-vector (r) that is normalized to 1. To keep the vector normalized during the measurement step, we use the approach outlined in [25].

With the state vector described, we can now describe the time and measurement updates utilized in our SLAM-based fusion technique. During the time update, only the current navigation state estimate of the camera needs to be updated. The updates occur according to the accelerometer and gyro measurements from the IMU. During the measurement step of the UKF, the location of each world point that is currently being observed in the image is used. The predicted measurement for each world location is

$$\vec{x} = \lambda \mathbf{C}_n^c \left(\vec{P}_c^n + \frac{1}{\xi}\vec{r} - \vec{p}_c^n \right) \tag{7}$$

where $\mathbf{C}_n^c$, $\vec{P}_c^n$, ξ, $\vec{r}$, and $\vec{p}_c^n$ are all derived from the current state estimate.

IV. COMPARISON OF METHODS

To compare the constraint and SLAM-based imaging and inertial fusion systems, we created a simulation environment that simulates both an IMU and camera for use in the imaging/inertial fusion algorithms. This simulator is described in the following subsection. The results obtained from adding noise and using each fusion method are described in Section IVB. A discussion of these results concludes this section.

A. Simulation Environment

To generate true navigation states of the MAV over time, a Bézier curve representing the true path of the MAV was created. A Bézier curve was chosen due to its inherent flexibility in representing many different types of curves in 3-D space. In addition Bézier curves are a polynomial function of a single scalar t, yielding two significant advantages. First the location at any time can be easily determined. Second by differentiating the polynomial with respect to t, the velocity and acceleration at any point on the curve can be computed in closed form. All quantities are assumed to be in a "navigation frame" which has North as its x axis, East as its y axis, and Down as the z axis. The origin of this frame was arbitrarily chosen as a location on the ground in Utah, near Brigham Young University (close to our MAV flight test area).

In addition to generating the location, velocity, and acceleration of the MAV, we also need to generate the angular orientation (attitude) of the MAV camera. We have used two basic approaches to generating the attitude of the MAV camera: 1) for a "fixed" camera, the angular orientation is always constant within the MAV body frame; 2) for a "gimballed" camera, the attitude of the camera is set so that the origin of the navigation frame is imaged in the center of the image.

Once the true location and attitude of the camera are known, the inputs to the fusion algorithm are generated. We assume the inputs from the IMU consist of 3-axis accelerometer and gyroscope (gyro) readings. To generate accelerometer readings, the acceleration of the camera is computed from the Bézier curve. The effects of gravity, Coriolis, and the rotation of the earth are then added to the accelerometer readings as described in [30], yielding noise-free accelerometer readings. To generate gyro readings, the attitude at two locations on the Bézier curve is computed. The locations on the curve are separated by the gyro sample time. The difference in attitude is then used to compute the angular rates of the camera, yielding noise-free gyro readings.

Once the noise-free readings have been computed, two types of noise are added to the sensor readings. First Gaussian, zero-mean white noise is added to the computed readings. The variance of the noise values were chosen to approximate measurement errors observed on a Kestrel autopilot [6]. Second a constant bias is added to the gyro and accelerometer readings. For each run of the simulator, biases were randomly selected from a Gaussian distribution with 10x the standard deviation of the white noise for that sensor.

To simulate inputs from the camera, a set of random world points to be imaged are created. These points are randomly distributed both in x-y location and about the zero altitude plane (i.e., the points are nonplanar.) Using the locations of the world points and the location and attitude of the MAV over time, a set of feature locations corresponding with time along its flight path are created. Feature locations for a specific MAV location and attitude are computed using the formula

$$\lambda \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} = \mathbf{K}\mathbf{C}_n^c(\vec{X}^n - \vec{p}^n) \tag{8}$$

where $\vec{X}^n$ is the location of the world point (in the navigation frame), $\vec{p}^n$ is the position of the camera in navigation frame coordinates (determined from its point on the Bézier curve), $\mathbf{C}_n^c$ is the direction cosine matrix from the navigation frame to the camera frame (also a function of location on the Bézier curve), $\mathbf{K}$ is the calibration matrix of the camera, mapping from Euclidean to pixel locations, λ is a scale factor used for normalizing the third element of the image frame vector to 1, and x_i and y_i are the image coordinates of the point.

After determining the location of the object in the image space, Gaussian white zero-mean noise is added to the image location. We set the standard deviation of the noise equal to a single pixel in the image plane. After adding noise, the pixel values are then "decalibrated" (multiplied by $\mathbf{K}^{-1}$) to obtain vectors in the same Euclidean space as the MAV navigation state.

B. Setup and Results of Comparison

Using the general setups described above for constraint-based and SLAM-based fusion of the IMU and visual information, we designed three different "flight scenarios" to test the relative merits of each fusion approach. In the first scenario, the camera moves in a straight line starting at 100 m above the ground, 100 m South of the navigation frame origin. The camera then moves in a straight line to 100 m North of the navigation frame origin, holding a constant altitude. In the East-West (y) direction, the MAV is always at 0. Along this path, 151 images were captured at a rate of 10 Hz, requiring 15 s to fly this path. These values were chosen to achieve an airspeed (13.3 m/s) typical of MAVs. In addition, 1501 samples of the gyro and accelerometer readings were collected. Note that while this flight scenario may seem like an overly simplistic maneuver (flying in a straight line), it was chosen because it actually exacerbates one of the fundamental problems of vision, the universal scale ambiguity. Therefore this scenario is one of the most difficult scenarios for vision-aided navigation. We refer to this scenario as the "sStraight-line flight" scenario. Results for this scenario, with a gimballed camera, are shown in Figs. 4 and 5.

To determine the overall accuracy of each technique, we ran each UKF filter setup 100 times. In Figs. 4, 5, 6, and 7, we use the "boxplot" command in MATLAB to display the median and spread of the errors present in the final location and attitude estimate across the 100 runs. In these plots, the line in the middle of the box represents the median of the data, the box represents the middle 50% of the data, and the crosses outside the lines represent outlier points.

(Note that in Fig. 4, it is apparent that both the constraint and SLAM-based fusion techniques achieve significant improvements over IMU-only fusion. Similar results were seen with all scenarios tested, so all other figures simply show comparative results between fusion techniques, allowing a more detailed display of results. Fig. 5 is the same as Fig. 4 except that IMU results have been removed.)

The second scenario tested represents a more generic flight of an MAV. It starts at −100 m North, 100 m in altitude. It then flies an "S" pattern, going East before turning back, passing directly over the navigation frame origin, headed in a northwest direction, and eventually turning back to arrive at −30 m East, 100 m North. During the course of the flight, the altitude drops from 100 m to 60 m. This entire flight takes 18 s. We refer to this scenario as "the S pattern," with results shown in Fig. 6.

In both scenarios described above, the world points being observed were distributed using a 3-D Gaussian distribution centered about the navigation frame origin. To keep the objects in view, the camera is continuously rotated to be "looking at" the origin, with the "top" of the camera facing in the direction of travel at the current point.

The final scenario used was to determine the response of the different fusion techniques to having a feature in view for only a short period of time. The flight pattern in this case was the same as the "straight-line flight" described above. However rather than distributing the world points about the navigation frame origin and always pointing the camera at the origin, 30 points were uniformly distributed in the North-South direction from 150 m North to 110 m South. The camera was kept at a fixed attitude of 80 degrees down. In the East-West and Down-Up direction, the points were distributed using a Gaussian distribution with a standard deviation of 20 m. This scenario is referred to as the "straight-line, fixed camera" scenario, and results are shown in Fig. 7.

For each of these flight scenarios, three different setups of our UKF environment were used. The most basic comparison of constraint-based and SLAM-based fusion entails running two different UKFs with the exact same IMU and feature location inputs. We refer to this as the "constraint1" and "SLAM" setups. For the straight-line flight and S pattern, 10 world points were used.

In addition to comparing the constraint and SLAM-based techniques using the exact same inputs, we also created a case that represents one of the fundamental differences between the two navigation approaches. One of the principal shortcomings of SLAM-based techniques is that, due to the inclusion of world points in the filter state, real-time SLAM filters are intrinsically limited in the number of features they can simultaneously localize. In addition it is far more difficult to reliably track features over a long time period as required by SLAM than to simply track features from one frame to the next. Therefore it is easier to reliably track a large number of features using the constraint-based technique then SLAM-based techniques. To represent this effect, we also simulated a constraint-based fusion technique (constraint2) that tracks 4 times as many features as the first two setups (i.e., 40 world points for the straight-line and S-pattern results, and 120 world points for the straight-line, fixed-camera results).

To tune the UKF, we first set the process noise matrix Q to exactly equal the sources of noise on the IMU. The Q matrix was then increased to compensate for nonlinearities in the process model. The values of Q were increased until the performance of the filter no longer increased, thereby tuning the Q matrix for use in these simulations.

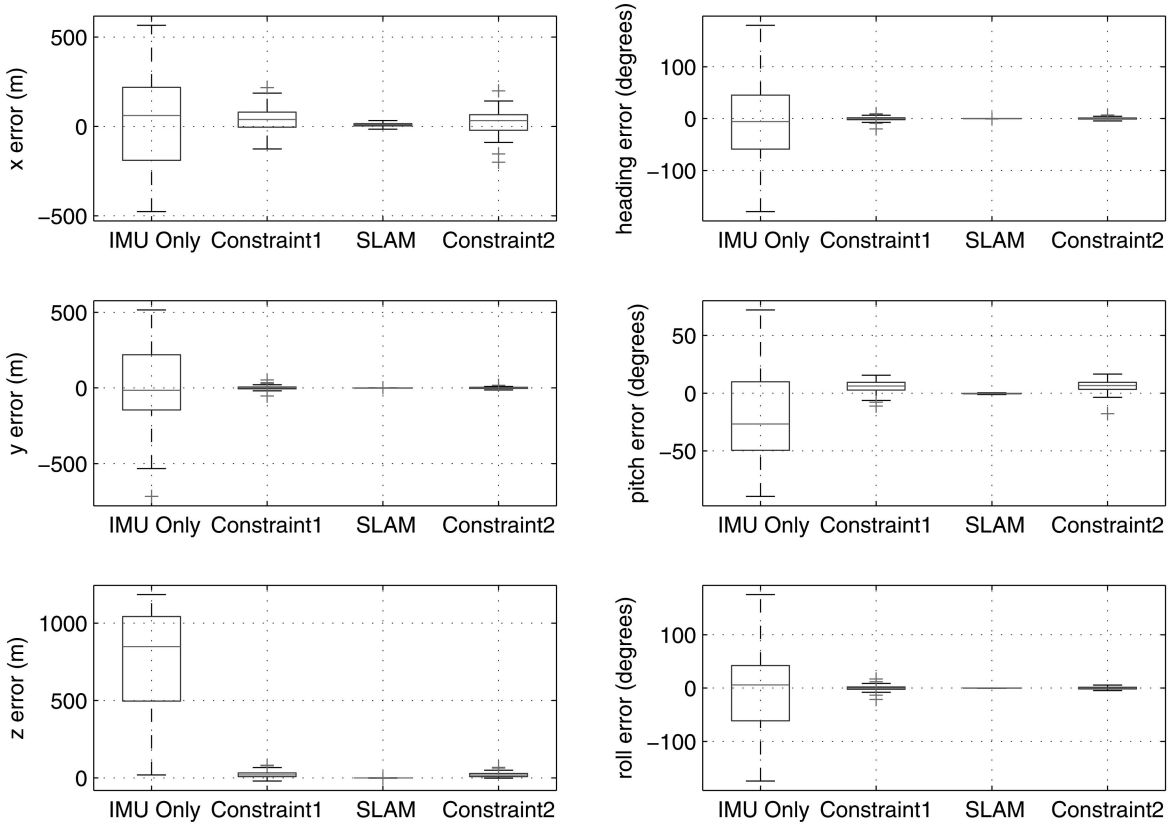

Fig. 4. Comparative results between fusion techniques and IMU-only—straight-line flight.

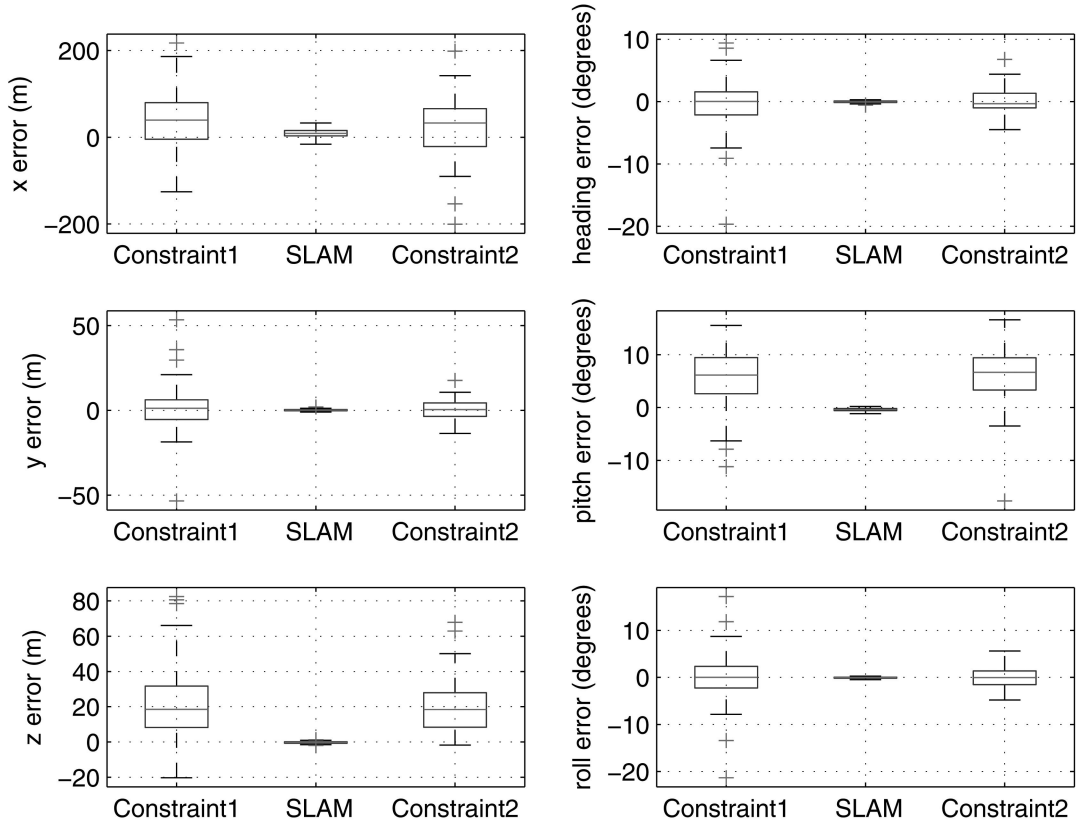

Fig. 5. Comparative results between fusion techniques—straight-line flight.

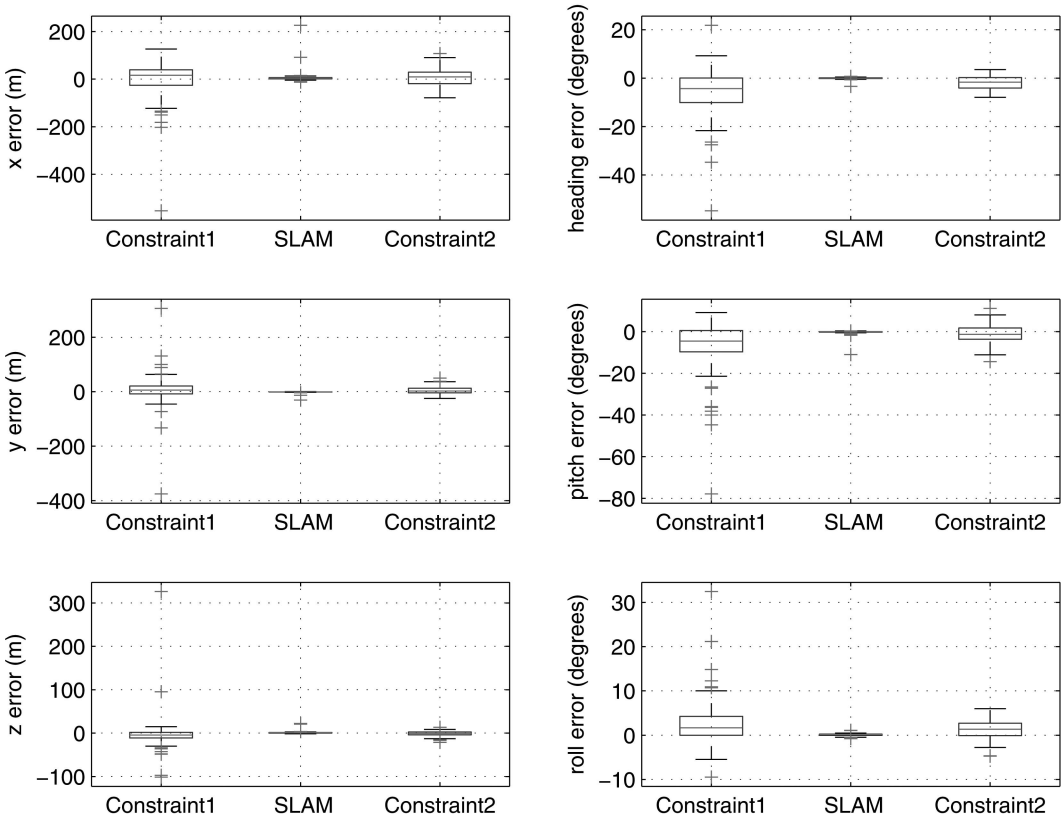

Fig. 6. Comparative results between fusion techniques—S-pattern.

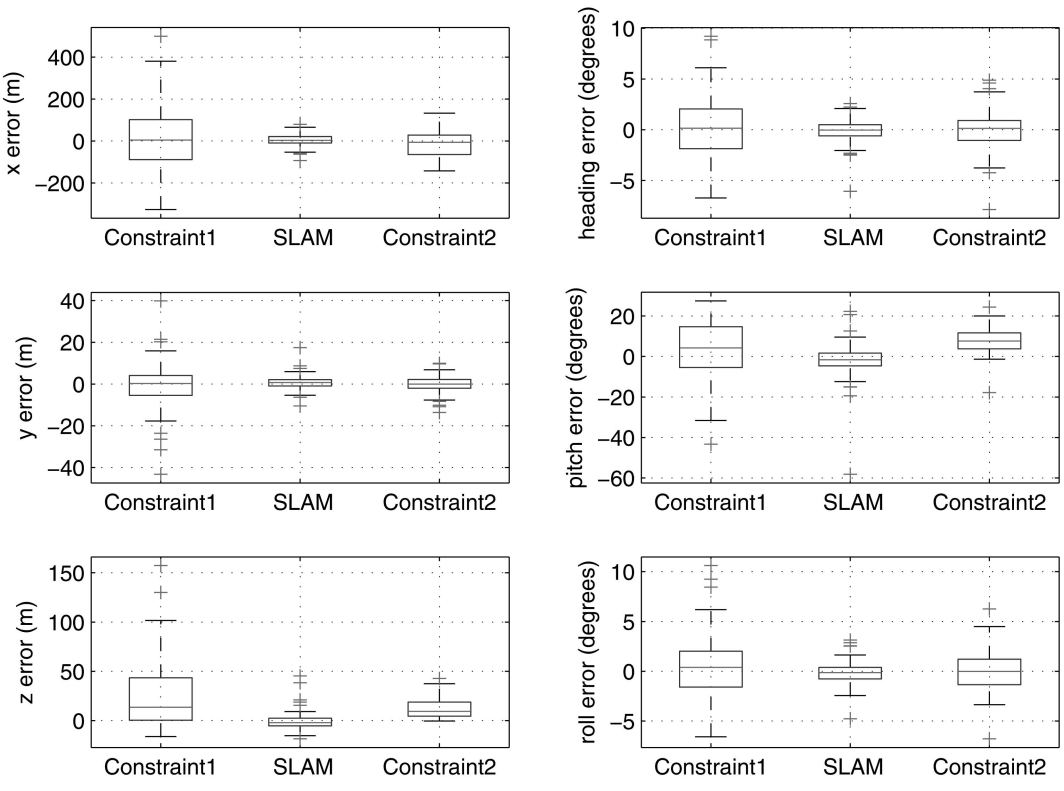

Fig. 7. Comparative results between fusion techniques—straight-line, fixed-camera flight.

C. Discussion of Results

From the data presented in Figs. 4–7, we can draw several conclusions about the relative strengths and weaknesses of the constraint- and SLAM-based techniques implemented in this paper for vision-assisted navigation. First let us consider

the case when the camera continuously tracks the same objects, whether flying a straight-line (Fig. 5) or an S-pattern (Fig. 6). In both these cases, the average error in the final navigation state estimated for SLAM-based techniques are at least an order of magnitude better than either constraint-based technique. From these results we conclude that if the MAV is flying in a pattern where the same general area is being observed over time (i.e., in an orbit pattern, perimeter surveillance, etc.), then SLAM-based navigation algorithms are required for accurate navigation.

However when the camera is fixed and the MAV camera is not continuously observing the same world points (Fig. 7), we observe a different pattern in the results. While the SLAM-based fusion technique is better than the constraint1 and constraint2 cases in position estimation, the constraint2 and SLAM results are much more comparable. Therefore when an area is quickly being observed by an MAV and will not be re-observed, then the constraint-based techniques may be useful in MAV flights.

In addition to the observation of what type of flight pattern favors SLAM-based techniques versus constraint-based techniques, our results also demonstrate some interesting attributes of each fusion technique. Note that in both straight-line flight scenarios (Figs.5 and 7), the standard deviation of the SLAM-based technique for T_x is significantly smaller than the constraint-based techniques. This demonstrates the significance of the global scale ambiguity in vision-based methods. With vision, there is always a scale ambiguity, but the SLAM-based technique attempts to find one scale ambiguity across the entire flight. Constraint-based techniques, on the other hand, have a scale ambiguity at every step of the fusion filter. Therefore the standard deviation in the direction of travel is much larger using constraint-based techniques than SLAM-based techniques.

V. CONCLUSIONS

In this paper, we have described two possible methods, constraint-based and SLAM-based, for fusing visual and IMU information together in an MAV environment. We have shown that the SLAM-based techniques we have implemented are significantly better at navigating when the camera can observe a single set of world locations for an extended period of time. However when consistently observing new points, the constraint-based technique is relatively equal to the SLAM-based technique. In all cases the SLAM-based technique will estimate with less variance the total distance traveled by the MAV. Both techniques are shown to be a significant improvement over IMU-only–based techniques.

REFERENCES

[1] Goodrich, M., Morse, B., Gerhardt, D., Cooper, J., Quigley, M., Adams, J., and Humphrey, C.
Supporting wilderness search and rescue using a camera-equipped mini UAV: Research articles.
Journal of Field Robotics, **25** (2008), 89–110.

[2] Bradley, J. and Taylor, C.
Particle filter based mosaic king for tracking forest fires.
In *Proceedings of the AIAA Conference on Guidance, Navigation, and Control*, Hilton Head, SC, Aug. 2007.

[3] Herwitz, S., Johnson, L., Arvesen, J., Higgins, R., Leung, J., and Dunagan, S.
Precision agriculture as a commercial application for solar-powered unmanned aerial vehicles.
Presented at the AIAA 1st Technical Conference and Workshop on Unmanned Aerospace Vehicles, 2002.

[4] Beard, R., Kingston, D., Quigley, M., Snyder, D., Christiansen, R., Johnson, W., McLain, T. and Goodrich, M.
Autonomous vehicle technologies for small fixed wing UAVs.
AIAA Journal of Aerospace Computing, Information, and Communication, **2**, 1 (Jan. 2005) 92–108.

[5] Kingston, D. B. and Beard, R. W.
Real-time attitude and position estimation for small UAVs using low-cost sensors.
In *Proceedings of the AIAA 3rd Unmanned Unlimited Systems Conference and Workshop*, Chicago, IL, Sept. 2004; Paper AIAA-2004-6488.

[6] Procerus Technologies
www.procerusuav.com, 2006. [Online].
Available: http://www.procerusuav.com/products.php.

[7] Micropilot
www.micropilot.com, 2006. [Online].
Available: http://www.micropilot.com/index.htm.

[8] Ready, B. B. and Taylor, C. N.
Improving accuracy of MAV pose estimation using visual odometry.
In *Proceedings of the American Control Conference* (ACC '07), 2007, 3721–3726.

[9] Andersen, E. D. and Taylor, C. N.
Improving MAV pose estimation using visual information.
In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS 2007), San Diego, CA, Oct. 29–Nov. 2, 2007, 3745–3750.

[10] Veth, M. and Raquet, J.
Fusing low-cost image and inertial sensors for passive navigation.
NAVIGATION: Journal of the Institute of Navigation, **54**, 1 (2007), 11–20.

[11] Veth, M. and Raquet, J.
Two-dimensional stochastic projections for tight integration of optical and inertial sensors for navigation.
In *Proceedings of the National Technical Meeting of the Institute of Navigation*, Air Force Institute of Technology, DTIC Research Report, 2006, 587–596.

[12] Veth, M. J., Raquet, J. F., and Pachter, M.
Stochastic constraints for efficient image correspondence search.
IEEE Transactions on Aerospace and Electronic Systems, **42**, 3 (2006), 973–982.

[13] Langelaan, J.
State estimation for autonomous flight in cluttered environments.
Ph.D. dissertation, Stanford University, 2006.

[14] Prazenica, R., Watkins, A., Kurdila, A., Ke, Q., and Kanade, T.
Vision-based Kalman filtering for aircraft state estimation and structure from motion.
In *Proceedings of the 2005 AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2005, 1–13.

[15] Strelow, D. and Singh, S.
Long-term motion estimation from images.
In *Proceedings of the International Symposium on Experimental Robotics*, July 2006.

[16] Brown, A., Bockius, B., Johnson, B., Holland, H., and Wetlesen, D.
Flight test results of a video-aided GPS/inertial navigation system.
In *Proceedings of the 2007 ION GNSS Conference*, 2007, 1111–1117.

[17] Kim, J. and Sukkarieh, S.
SLAM aided GPS/INS navigation in GPS denied and unknown environments.
Presented at the 2004 International Symposium on GNSS/GPS, Sydney, Australia, Dec. 6–8, 2004.

[18] Rees, G. and Alexander, R.
A framework for assessing and designing vision-based slam systems for autonomous vehicles.
In *Proceedings of the SEAS-DTC Technical Conference*, 2007.

[19] Huang, S. and Dissanayake, G.
Convergence and consistency analysis for extended Kalman filter based SLAM.
IEEE Transactions on Robotics, **23**, 5 (2007), 1036–1049.

[20] Gibbens, P., Dissanayake, G., and Durrant-Whyte, H.
A closed form solution to the single degree of freedom simultaneous localisation and map building (SLAM) problem.
In *Proceedings of the 39th IEEE Conference on Decision and Control*, vol. 1, 2000, 191–196.

[21] Maybeck, P.
Stochastic Models, Estimation and Control, vol. 1.
London: Academic Press London, 1982.

[22] Julier, S. J. and Uhlmann, J. K.
New extension of the Kalman filter to nonlinear systems.
In Ivan Kadar (Ed.), *Signal Processing, Sensor Fusion, and Target Recognition VI. Proceedings of the SPIE*, vol. 3068, July 1997, 182–193.

[23] Doucet, A. and De Freitas, N.
Sequential Monte Carlo Methods in Practice.
New York: Springer, 2001.

[24] Soatto, S., Frezza, R., and Perona, P.
Motion estimation via dynamic vision.
IEEE Transactions on Automatic Control, **41**, 3 (1996), 393–413.

[25] Julier, S. and LaViola, J.
On Kalman filtering with nonlinear equality constraints.
IEEE Journal of Signal Processing, **55**, 6 (2007), 2774–2784.

[26] Durrant-Whyte, H., Rye, D., and Nebot, E.
Localization of autonomous guided vehicles.
In G. Hirzinger and G. Giralt (Eds.), *Proceedings of the 8th International Symposium on Robotics Research*, New York: Springer, 1995, 613–625.

[27] Piniés, P., Lupton, T., Sukkarieh, S., and Tardós, J. D.
Inertial aiding of inverse depth SLAM using a monocular camera.
In *IEEE International Conference on Robotics and Automation*, 2007, 2797–2802.

[28] Kim, J-H. and Sukkarieh, S.
Airborne simultaneous localisation and map building.
In *Proceedings of the IEEE International Conference on Robotics and Automation* (ICRA '03), vol. 1, 2003, 406–411.

[29] Montiel, J. M. M., Civera, J., and Davison, J.
Unified inverse depth parametrization for monocular SLAM.
Robotics Science and Systems, **9** (2006), 1.

[30] Titterton, D. and Weston, J.
Strapdown Inertial Navigation Technology.
Lavenham, UK: Peregrinus Ltd., 1997.

Clark N. Taylor (M'00—SM'10) is currently an electronics engineer with the Sensors Directorate of the Air Force Research Laboratory. He received his Ph.D. degree in electrical and computer engineering from the University of California, San Diego, in 2004.

He has published over 30 papers in the fields of digital video processing for unmanned aerial vehicles (UAVs), vision-based navigation, video communication, and digital systems design. He is also an associate editor for the *IEEE Transactions on Circuits and Systems for Video Technology* and has served on the technical program committee for several conferences and workshops. He was awarded a Young Investigator Award from AFOSR for vision-aided navigation for MAVs and was selected as an Air Force summer faculty fellow in 2007 and 2009.

Michael J. Veth (SM'09) received his Ph.D. in electrical engineering from the Air Force Institute of Technology and a B.S. in electrical engineering from Purdue University.

Lt. Col. Veth is currently an Assistant Professor of Electrical Engineering at the Air Force Institute of Technology where he serves as the Deputy Director of the Advanced Navigation Technology Center. His current research focus is understanding and implementing bio-inspired methods to fuse image and inertial systems for navigation, targeting, and control. He is a member of the Institute of Navigation, Tau Beta Pi, and Eta Kappa Nu. In addition, he is a graduate of the Air Force Test Pilot School.

John F. Raquet (M'05) received a B.S. in astronautical engineering from the U.S. Air Force Academy, an M.S. in aero/astro engineering from the Massachusetts Institute of Technology, and a Ph.D. in geomatics engineering from the University of Calgary.

He currently serves as an Associate Professor of Electrical Engineering at the Air Force Institute of Technology, where he is also the Director of the Advanced Navigation Technology (ANT) Center. He has been working in navigation-related research for over 19 years, and has authored or coauthored over 100 journal articles and conference papers relating to a wide variety of navigation-related technologies.

Mikel Miller (M'02) earned his Ph.D. in electrical engineering in 1998 from the Air Force Institute of Technology (AFIT).

He is currently serving as Chief Scientist of the Munitions Directorate, Air Force Research Laboratory, Eglin Air Force Base, FL. He is the Directorate's principal scientific and technical advisor and primary authority for the technical direction of a broad, multi-disciplinary research and development portfolio encompassing all aspects of munitions science and technology.

Dr. Miller has authored/coauthored more than 50 journal articles, technical papers, and documents and a NATO handbook on navigation technologies. He is a Fellow of the Institute of Navigation and the current president. In addition, he recently finished a 2-year term as the Chairman of the Joint Service Data Exchange. He is also a Fellow of the Royal Institute of Navigation, and a member of AIAA. He is also an Adjunct Professor of Electrical Engineering at AFIT and Miami University of Ohio.