Precise Vision-Aided Aerial Navigation

Han-Pang Chiu, Aveek Das, Phillip Miller, Supun Samarasekera, Rakesh (Teddy) Kumar

***Abstract*— This paper proposes a novel vision-aided navigation approach that continuously estimates precise 3D absolute pose for aerial vehicles, using only inertial measurements and monocular camera observations. Our approach is able to provide accurate navigation solutions under long-term GPS outage, by tightly incorporating absolute geo-registered information into two kinds of visual measurements: 2D-3D tie-points, and geo-registered feature tracks. 2D-3D tie-points are established by finding feature correspondences to align an aerial video frame to a 2D geo-referenced image rendered from the 3D terrain database. These measurements provide global information to correct accumulated error in navigation estimation. Geo-registered feature tracks are generated by associating features across consecutive frames. They enable the propagation of 3D geo-referenced values to further improve the pose estimation. All sensor measurements are fully optimized in a smoother-based inference framework, which achieves efficient relinearization and real-time estimation of navigation states and their covariances over a constant-length of sliding window. Experimental results demonstrate that our approach provides accurate and consistent aerial navigation solutions on several large-scale GPS-denied scenarios.**

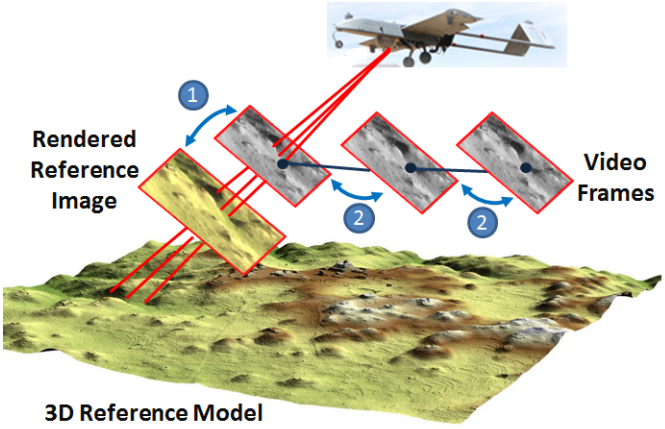

Fig. 1: The concept of two kinds of visual measurements: (1) 2D-3D tie-points, which are established by aligning a video frame to a 2D rendered reference image; (2) geo-registered feature tracks, which are generated by associating features across video frames with 3D information from (1).

I. Introduction

Precise navigation systems for aerial vehicles typically rely on GPS signals coupled with inertial measurement unit (IMU) data. The absolute position information from GPS can be introduced to eliminate the error accumulated by IMU propagation. However, GPS is unreliable in complicated environments. It is also vulnerable to other electromagnetic signals due to malicious attacks.

Augmenting aerial navigation with a monocular video camera is rapidly emerging as a feasible and low-cost solution under GPS outage. To avoid the estimated position from eventual drift, recent vision-aided navigation systems [2], [4], [6] aim to obtain a single position measurement by aligning images from the camera to reference aerial imagery. However, these methods compute only 2D absolute position of the aerial vehicle on the horizontal plane from the alignment, and require additional altitude information from an on-board barometer to update the height estimation.

In this paper, we propose a novel vision-aided aerial navigation approach which combines inertial measurements with two kinds of visual measurements (Figure 1): 2D-3D tie-points and geo-registered feature tracks. Our work focuses on how to fully integrate camera observations with geo-referenced information into a probabilistic sensor fusion framework. Unlike previous vision-aided aerial navigation methods, we treat each observation of a feature as a single measurement instead of computing only one pose measurement from all observations at a given time. This way tightly incorporates absolute geo-registered information into measurements, and is capable of precise 3D pose estimation in GPS-denied settings with no additional sensors other than IMU and cameras.

The 2D-3D tie-points are established by finding feature correspondences to align an aerial video frame to a 2D reference image. This 2D reference image is rendered from a 3D terrain model, which is built offline prior to navigation, using the pose predicted from the inference engine. The aligning process, which is called geo-registration, registers aerial video to a geo-referenced coordinate system at 1Hz rate. It provides 3D absolute information to 2D-3D tie-points through feature correspondences, and allows subsequent calculation of accurate geo-referenced 3D coordinates for any given 2D point on the video frame.

The geo-referenced 3D coordinates for 2D features are also utilized in the feature track measurements, generated by associating these features across consecutive frames from a monocular video camera. The measurement formulation enables the propagation of 3D absolute information from one frame to another through geometric constraints, by observing the same feature from multiple frames. It improves the accuracy of pose estimation involved with these frames.

To optimally incorporate all constraints from these visual measurements, we use a sliding-window factor graph framework [9] which is based on recent incremental graph-based smoothing techniques [1] for real-time estimation. This framework maintains only the portion of the total

H. Chiu, A. Dass, P. Miller, S. Samarasekera, and R. Kumar are with Center for Vision Technologies, SRI International, Princeton, NJ 08540, USA {han-pang.chiu,aveek.das,phillip.miller,supun.samarasekera,rakesh.kumar}@sri.com

factor graph [7] that exists inside a sliding time window, and supports efficient relinearization on stored factors. We extend this framework to handle measurement latency due to geo-registration process, and periodically propagate updates within the sliding window to current estimation.

The remainder of this paper begins with a discussion of related work in Section II. Section III introduces our sliding-window factor graph framework, and illustrates how it is extended to accommodate delayed measurements from geo-registration. Section IV describes our two kinds of visual measurements in detail based on this factor graph framework, while Section V focuses on the integration of inertial measurements and the maintenance of estimation consistency in our system. We demonstrate our approach provides accurate and consistent aerial navigation solutions on several large-scale GPS-denied scenarios in Section VI followed by our conclusions in Section VII.

II. Related Work

There have been many efforts to fuse vision data from an on-board camera with inertial measurements for aerial navigation systems. Most of these works focused on feature tracking and optical flow methods to estimate relative motion of the platform. However, without periodical updates using absolute information such as GPS [3], they suffer from the accumulated error over a long period.

Utilizing geo-referenced aerial imagery to provide absolute information becomes a feasible solution in recent vision-aided aerial navigation systems [2], [4]–[6]. These works used different ways to match a video frame to reference imagery, and computed only 2D absolute position of the vehicle from this video frame. For example, Sim et al. [5] aligned the video frame to a reference image using Hausdorff distance measure. However, this work mainly focuses on image processing issues and lacks the sensor fusion scheme.

Conte and Doherty [6] proposed a two-layer framework to fuse camera data with other sensors for aerial navigation. The first layer integrates both relative odometry and absolute position measured from a single camera. The video frame is registered to the reference imagery using normalized cross correlation to compute the absolute position of the aerial vehicle on the ground plane. The second layer then fuses the computed position from the first layer with inertial measurements. Note this framework also needs a barometer sensor which provides height estimation to complement the 2D updates from camera system.

Lindsten et al. [2] also obtained both related visual odometry and absolute position from a single camera. They segmented all video frames and reference images into regions, which are classified as grass, asphalt, or house. The alignment of a video frame to the reference imagery, which is based on these classified regions, becomes more robust against environment changes such as different lighting conditions. It computes a horizontal position and constrains the estimation in 2D coordinates.

Patterson et al. [4] extended the work in [6] by performing registration based on features, such as roads, paths, and water. The registration process computes a single 2D position measurement based on the correlation between two maps formed by these features. This way reduces storage requirements on geo-referenced imagery, and increases registration robustness to lighting variations. However, it still needs an additional barometer sensor to measure height information.

Absolute information from registration of geo-referenced imagery has also been used for other aerial applications, such as ground target localization [13] and planetary landing [8]. For these applications, geo-registration is performed opportunely and does not support long-term navigation. For example, Han et al. [13] used prediction from GPS and IMU data as initial pose guess for transforming a video frame to register reference imagery. The pose is then refined by Iterative Closest Point algorithm to improve feature-based registration for ground target localization. Mourikis et al. [8] fuses both absolute 2D-3D tie-points and relative geometric constraints from a camera, with IMU data for spacecraft landing. The 2D-3D tie-points are only established during landing, between 2D imaged features and 3D geo-referenced locations on the map of the landing site.

Compared to previous works in vision-aided aerial navigation, our approach is able to estimate precise 3D absolute pose using only IMU and cameras. We utilize two kinds of visual measurements to fully incorporate absolute geo-registered information. The 2D-3D tie-point model provides absolute information to update navigation estimation, while the feature track formulation propagates the influence of 3D geo-referenced information across consecutive video frames. Unlike [8], we share geo-registered information between two types of visual measurements to improve the estimation specifically for navigation accuracy. We also support efficient geo-registration process in a continuous manner for long-term navigation.

Instead of traditional filtering methods [4], [6], [8] for aerial navigation, we utilize a sliding-window factor graph framework to estimate states over a constant-length sliding window with fixed computational cost. This framework is designed to handle highly nonlinear measurements such as monocular camera observations, by iteratively relinearizing measurements within the window. We extend this framework to accommodate measurement latencies from geo-registration process in an optimal way, by associating new factors with the correct navigation states when the measurement arrives within the smoothing window. It concurrently processes all measurements inside the window at a specified rate, and corrections from delayed geo-registered measurements are automatically propagate to current estimation.

III. Sliding-Window Factor Graphs

In this section we introduce our sliding-window factor graph framework [9], which maintains only the portion of measurement factors that were received within a sliding time window. It achieves real-time estimation, and supports efficient relinearization on stored factors. Based on this methodology, we extend the framework to naturally incorporate delayed geo-registered visual measurements.

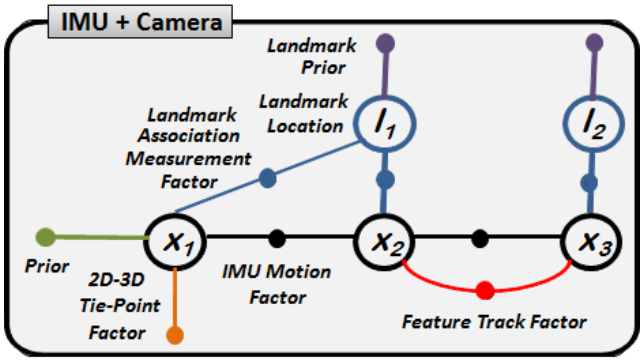

Fig. 2: The concept of all kinds of measurement factors in our framework. Each navigation state receives a camera frame at a given time. An IMU motion factor is used to connect consecutive navigation states. Each observation of a feature is treated as one measurement. There are three types of factors (unary: 2D-3D tie-point factor, binary: feature track factor, extrinsic: landmark association measurement factor) used to model our two kinds of visual measurements (2D-3D tie-points, and geo-registered feature tracks).

A. Factor Graphs

A factor graph [7] represents the navigation estimation problem at all times as a bipartite graph model $G = (\mathcal{F}, \Theta, \mathcal{E})$ with two node types: *factor nodes* $f_i \in \mathcal{F}$ and *state variable nodes* $\theta_j \in \Theta$. An edge $e_{ij} \in \mathcal{E}$ exists if and only if factor f_i involves state variables θ_j. The factor graph G defines the factorization of a function $f(\Theta)$ as

$$f(\Theta) = \prod_i f_i(\Theta_i) \tag{1}$$

where Θ_i is the set of all state variables θ_j involved in factor f_i, and independent relationships are encoded by edges e_{ij}.

A generative model

$$z_i = h_i(\Theta_i) + n_i \tag{2}$$

predicts a sensor measurement z_i using a function $h_i(\Theta_i)$ with measurement noise n_i. The difference between measurement function $h_i(\Theta_i)$ and the actual measurement $\tilde{z}_i$ is encoded into a factor. Assuming the underlying noise process is Gaussian with covariance Σ, the resulting factor is

$$f_i(\Theta_i) = ||h_i(\Theta_i) - \tilde{z}_i||^2_\Sigma \tag{3}$$

where $||\cdot||^2_\Sigma$ is the Mahalanobis distance.

The factor graph representation on the full non-linear optimization problem has led to the recent incremental solution, iSAM2 [1]. Using a Bayes tree data structure, iSAM2 keeps all past information and only updates variables influenced by each new measurement. It therefore obtains same result as a batch solution to the full non-linear optimization.

Figure 2 shows all kinds of measurement factors used in our framework. There are three different factor classes: unary, binary, and extrinsic. The unary factor only connects a state at a single time, while a binary factor involves two navigation states at different times. Each extrinsic factor involves a navigation state x and an unknown extrinsic entity. The landmark association measurement derived from cameras is the most popular example. Each landmark association factor involves both navigation state at time i and the state of unknown landmark position l.

B. Sliding-Window Extension

The sliding-window factor graph framework [9] utilizes a parallel architecture to split the estimation into a fast short-term smoother and a a slower global smoother. To maintain constant-time updates, the short-term smoother extends the iSAM2 algorithm to support inference over a sliding constant-length window and efficiently removes states that are outside of the window. The fully global smoother, which keeps all past states, processes only expensive loop closures.

We utilize the short-term smoother for better optimization of highly nonlinear factors. Traditional filtering methods only keep the current state, and linearize measurements only once at the time of arrival. However, some states require several measurements before a good estimate can be obtained. Using the original linearization point for these states may lead to poor estimation and convergence performance. This is particularly true of 3D landmark positions estimated from tracked visual features. The short-term smoother relinearizes factors inside the window at a particular frequency, and achieves a more optimal solution than filtering methods by checking consistency across a larger collection of sensor measurements.

C. Handling of Delayed Geo-Registered Measurements

We further extend this sliding-window factor graph framework to handle out-of-order measurements due to latencies from geo-registration process. The idea is to associate new factors with correct correspondent navigation states when delayed measurements arrive within the smoothing window. For example, when 2D-3D tie-points for time j arrive, the system first locates the IMU motion factor between x_i and x_k which covers the actual measurement time j. It then creates a new navigation state x_j for these measurements, and properly divides the located IMU motion factor into two new IMU motion factors which connect these three states (x_i,x_j,x_k). Finally it adds new unary factors for these tie-points to navigation state x_j. Since we currently set the smoothing window length as 4 seconds to cover the longest time feature gets tracked, all delayed measurements from geo-registration process can be received and processed.

This extended framework also generates the navigation estimation at a particular rate, which is set according to application requirement. It collects factors during the specified time interval, and adds them into the factor graph for inference periodically. It concurrently processes all measurement factors inside the window at the specified frequency. This way avoids performing updates every time when a new measurement is received, and decreases computational overhead if there are high-frequency sensors. It also naturally propagates the influence from delayed geo-registered measurements to current estimation within the window. Currently we set the inference rate as 1Hz, which is the same as the frequency of the geo-registration process in our system.

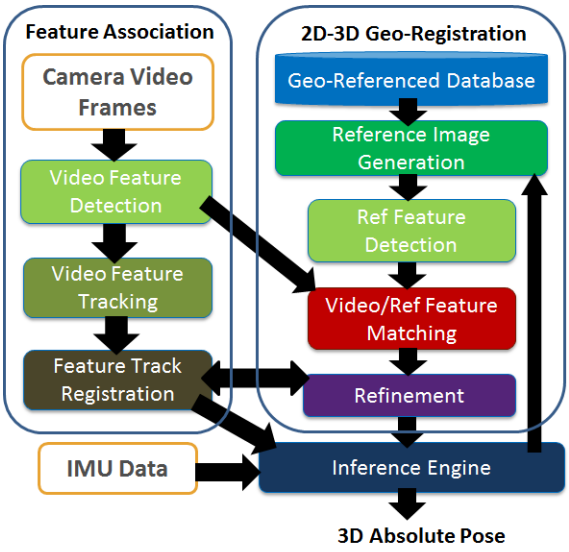

Fig. 3: The pre-processing procedures in our system. Note our inference engine (extended sliding-window factor graph) receives three kinds of measurements: IMU data, 2D-3D tie-points from 2D-3D geo-registration process, and geo-registered feature tracks from feature association process.

IV. Visual Measurement Model

In this section we introduce the formulations of our two kinds of visual measurements (Figure 2), which tightly incorporates absolute geo-referenced information into monocular camera observations of point features, within our extended sliding-window factor graph framework. The 2D-3D tie-points are encoded as unary factors, and geo-registered feature tracks are modeled as two types of factors: binary (feature track factor) and extrinsic (landmark association measurement factor).

A. Pre-Processing Procedures

While data from simple sensors such as IMU can be directly converted into factors in our system, data from complex sensors such as cameras must be pre-processed to extract meaningful measurements. Our pre-processing procedures (Figure 3), which include 2D-3D geo-registration and feature association, derive two kinds of visual measurements from image-based sensors.

1) 2D-3D Geo-Registration : The 2D-3D geo-registration process matches features extracted from both a single video frame and a rendered scene image to obtain 2D-3D tie-points after alignment. The database consists of the reference imagery of previously collected satellite images and terrain maps, that have been precisely aligned to 3D geo-coordinates. It renders imagery from a 3D viewpoint, set by the requested predicted pose from our inference engine. This rendered image is then matched to current video frame, by comparing detected features. If the matching succeeds after refinement, it results in a set of 2D-3D tie-points with associated uncertainties, which represent matches from 2D feature points on the video frame to 3D positions for features on the reference image. For details of our geo-registration process, we refer to [10], [14]–[16].

Our 2D-3D geo-registration process currently requires one second to register one camera video frame. For cameras with frequency higher than 1Hz, it skips frames to maintain real-time performance.

2) Feature Association : The feature association process (Figure 3) tracks features across consecutive video frames. For each tracked feature at the current frame, if this feature is not set with 3D geo-referenced value from previous tracked frames, it keeps requesting 3D information from 2D-3D geo-registration process. If the geo-registration process succeeds, 3D values for these features can be retrieved by ray-tracing techniques using geo-registration results.

Since our inference engine collects measurements and updates estimation at 1Hz (Section III-C), it naturally handles the delayed information from geo-registration process and does not cause additional latency. All tracked features with or without 3D geo-registered values will be fed into the extended sliding-window factor graph framework. These derived measurements propagate geo-referenced information to future frames, which are associated with same feature tracks but without successful geo-registration.

B. 2D-3D Tie-Points

In our system, we define the navigation state for a given time as $x = \{\Pi, v, b\}$. Each state x covers three kinds of nodes: pose node Π includes 3d translation t (body in global) and 3d rotation R (global to body), velocity node v represents 3d velocity, and b denotes sensor-specific bias block such as IMU bias. To simplify the notation, we assume all sensors have the same center, which is the origin of the body coordinate system.

Note there may be too many 2D-3D tie-points returned from successful geo-registration process. To reduce computation cost, we select only high-quality 2D-3D tie-points based on their uncertainties computed from geo-registration process. We then formulate one unary factor for each selected 2D-3D tie-point, between a 2D feature on the current video frame and the matched 3D geo-referenced point from the reference image.

Since the uncertainty of these selected matches is small, we treat the 3D position as a fixed 3D point in the geo-referenced world coordinate system. We then transform this fixed 3D point Y to the body coordinate system as $Z = [\ Z_1 \quad Z_2 \quad Z_3\ ]^T$, based on rotation R_i and translation t_i in state x_i. Since this factor only involves variables $\Pi = (R, t) \in x$, the measurement model of the observation in normalized image coordinates with noise η is as follows.

$$z_i = [\ Z_1/Z_3 \quad Z_2/Z_3\ ] + \eta,\ Z = R_i(Y - t_i) \qquad (4)$$

We compute the measurement residual and linearize the estimates as:

$$r_i = z_i - \hat{z}_i = z_i - h(\hat{\Pi}_i) \simeq H_{R_i}\delta R_i + H_{t_i}\delta t_i + \eta \qquad (5)$$

where H_{R_i} and H_{t_i} are the Jacobians of the measurement z_i with respect to R_i and t_i as follows. Note we use error quaternion to represent attitude error for computation.

$$H_{R_i} = J\left[\hat{R}_i(\hat{Y} - \hat{t}_i)\right], H_{t_i} = J(-\hat{R}_i) \quad (6)$$

$$J = \begin{bmatrix} 1/\hat{Z}_3 & 0 & -\hat{Z}_1/\hat{Z}_3^2 \\ 0 & 1/\hat{Z}_3 & -\hat{Z}_2/\hat{Z}_3^2 \end{bmatrix} \quad (7)$$

This factor formulation is applied to all selected 2D-3D tie-points passed from the successful geo-registration process. These unary factors provide immediate absolute information to update estimation.

C. Geo-Registered Feature Tracks

For each tracked feature from the feature association process, we use the 3-stage method [9] to model this feature across multiple navigation states if there is no 3D geo-referenced value available. This method is based on the maturity of the estimation of the underlying 3D location of the landmark for the tracked feature, and models the feature by two factor classes (Figure 2): binary and extrinsic.

The first stage avoids unstable initialization of the 3D location of the landmark points while still incorporating the landmark image observation into binary factor formulation for optimization. The second stage utilizes the extrinsic factor to estimate both navigation states and the 3D location of the associated landmark. Once the uncertainty of the 3D landmark state becomes small, the third stage switches back to binary factor formulation but treats the computed 3D location of the landmark as a fixed quantity when estimating the navigation state, saving computation time.

If the 3D geo-referenced information for the tracked feature is available, the tracked feature can be directly formulated as a binary factor between two navigation states, which are correspondent to consecutive video frames. This way avoids the construction and estimation of underlying 3D landmark state in the 3-stage method. Consider a single feature s tracked from state x_{i-1} to state x_i, this factor only involves variables $\Pi = (R, t) \in x$. The nonlinear measurement model for observations of s on $\Pi_{i-1}$and Π_i is

$$z_k = h(P_s, \Pi_k) + n_k = h(P_s^k) + n_k, k = i-1, i \quad (8)$$

where $P_s = [Z_1\, Z_2\, Z_3]^T$ is the unknown 3D position of this feature in world coordinate system, $P_s^k = R_k(P_s - t_k)$ is the 3D feature position transformed from geo-referenced world coordinate system to body coordinate system on state x_k, $h(P_s^k) = \begin{bmatrix} Z_1/Z_3 & Z_2/Z_3 \end{bmatrix}^T$is the projection on the normalized image plane, and n_k is the 2-dimension image noise vector with covariance matrix $C_k = \sigma_{im}^2 I_2$.

Since the geo-registration process is successful, P_s can be directly set from geo-registration results. This way ensures the formulation directly uses the absolute 3D geo-referenced value to form the binary constraint. After this 3D estimation is obtained, we compute the measurement residual and linearize the estimates of Π_k and P_s as:

$$r_k = z_k - \hat{z_k} = z_k - h(\hat{P_s}, \hat{\Pi_k}) \simeq H_{\Pi_k}\delta\Pi_k + H_{s_k}\delta P_s + n_k \quad (9)$$

where H_{Π_k} and H_{s_k} are the Jacobians of the measurement z_k with respect to Π_k and P_s respectively. We then stack z_{i-1} and z_i together as:

$$r \simeq H_\Pi \delta\Pi + H_s \delta P_s + n \quad (10)$$

where $r = [r_{i-1}; r_i]$, $H_\Pi = \left[H_{\Pi_{i-1}}; H_{\Pi_i}\right]$, $H_s = \left[H_{s_{i-1}}; H_{s_i}\right]$, and $n = [n_{i-1}; n_i]$ with covariance matrix $C = \sigma_{im}^2 I_4$. To marginalize out state P_s in the formulation, we project r on the left nullspace of H_s to get r_o using a unitary matrix U whose columns form the basis of the left nullspace of H_s:

$$r_o = U^T(z - \hat{z}) \simeq U^T H_\Pi \delta\Pi + U^T n = H_o \delta\Pi + n_o \quad (11)$$

where r_o is a 1-dimension vector after projection. Then we split H_o into $H_{o_{i-1}}$ and H_{o_i} for state $\Pi_{i-1}$and Π_i respectively. This results in the following linearized constraint between two states for our factor formulation, and has been shown [11] to yield better results than epipolar constraints.

$$r_o = H_{o_{i-1}}\delta\Pi_{i-1} + H_{o_i}\delta\Pi_i + n_o \quad (12)$$

There are two major differences between our factor formulation and the feature track model used in [8], [11]. First, unlike [8], the 3D geo-referenced information is used in our feature track formulation. Since geo-registration is operating at 1-Hz and may not always be successful, our model extends the influence from past geo-registration results to new navigation states which are connected by same feature track. Second, our formulation generates a binary factor immediately for a feature tracked across two consecutive frames. This way is different than constructing only a single measurement with all involved frames when a tracked feature breaks [11], but still maintains consistent 3D estimation through all frames connected by the same feature.

V. Navigation System

In this section, we show how we formulate inertial measurements into factors. We also describe how the estimation consistency is maintained in our system.

A. IMU Motion Model

Note our extended sliding-window factor graph framework (Figure 2) is able to integrate many types of sensors, and a single factor typically encodes only one sensor measurement. However, IMU sensors produce measurements at a much higher rate than other sensor types. To fully utilize high-frequency IMU data while saving the time to create factors, we design a single factor to summarize multiple consecutive IMU measurements. A navigation state is only created at the time when a non-IMU measurement (such as measurements from a video camera frame) comes or no non-IMU measurement arrives after a certain interval (such as one second), and the IMU factor is built to connect two sequential navigation states by integrating IMU measurements between them.

We formulate this factor using an error-state IMU propagation mechanism [11], and implement it [9] as a binary factor

between two consecutive states x_{i-1} and x_i. It generates 6 degrees of freedom relative pose and corresponding velocity change as the motion model. It also tracks the IMU-specific bias as part of the state variables for estimating motion.

We use this factor instead of traditional process models in our system. The linearization point to integrate non-IMU measurements at x_i is computed from this factor, which is based on the linearization point for x_{i-1} and IMU readings between x_{i-1} and x_i. If there is no IMU available, we use constant velocity assumption as the process model.

In contrast to traditional filtering techniques, the IMU motion factor is part of the full non-linear optimization process in our extended sliding-window factor graph framework. The value of IMU integration changes during re-linearization for iterative optimization.

B. Estimation Consistency

Our system maintains probabilistic consistency in estimating navigation states and their uncertainty. The extended sliding-window factor graph framework is a hybrid system consisting of nonlinear factors that are wholly within the sliding window, and linear factors that were adjacent to both variables inside the sliding window and variables that have been marginalized. All factors, states, as well as their associated full covariance matrices, are stored within the sliding time window. They are re-computed during re-linearization. In addition, the linearization point of any variable adjacent to a linear factor is kept constant to specifically assert the estimation consistency, as in [9].

VI. Experimental Results

This section demonstrates that our vision-aided approach provides precise aerial navigation solutions on two large-scale real scenarios both assuming GPS is not available for the navigation state estimate. These two scenarios are used to verify different aspects of our approach. The initial global position and orientation of the aerial vehicle is assumed known. Ground Truth is obtained by using the RTK differential GPS technique [12].

A. Scenario 1

This data set is collected using an aircraft that flies over a large area, including forest and urban cites. The airplane travels 38.9 kilometers in 10 minutes. The sensor set includes one 100Hz IMU and two downward-looking 5Hz monocular cameras. Note both cameras are used to track ground features across consecutive video frames. However, we only use one camera to perform geo-registration for obtaining absolute 3D information, due to the limitation of computational resource.

1) Geo-Registration: Figure 4 shows a few successful geo-registration results on Scenario 1. Note the reliability of geo-registration process heavily depends on the distinctness of ground features observed from the camera. For example, if the aerial image scene is mostly covered by trees when flying over the forest, geo-registration may easily fail. The geo-registration process is more robust if there are discriminative geometry features available, such as buildings and roads.

Fig. 4: Successful geo-registration results on Scenario 1. The left column shows two video frames, and the right column shows the matched reference images after geo-registration refinement.

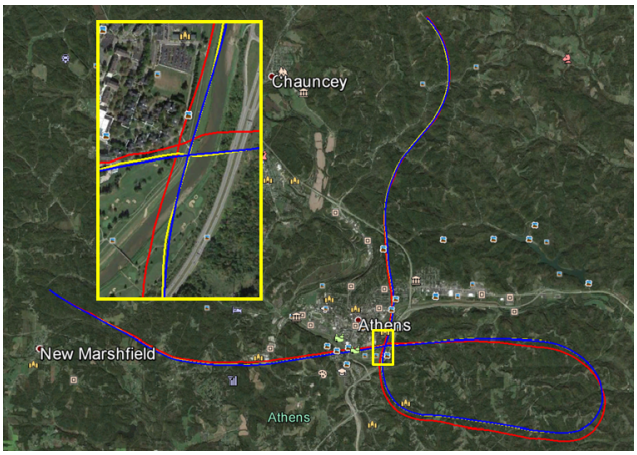

Fig. 5: The estimated trajectories for Scenario 1 (38.9 km, 10 min): Ground Truth (blue), result without geo-registration (red, 3D RMS error: 172.11 meters), and result with 2D-3D tie-points from geo-registration (yellow, 3D RMS error: 13.98 meters). The intersection region during navigation is also highlighted and enlarged for visualization.

However, the focus of this paper is to incorporate geo-registered information into visual measurements for navigation estimation, rather than improving the geo-registration process itself.

Figure 5 shows the 3D RMS error of the estimated trajectory for this scenario can be reduced from 172.11 meters to 13.98 meters, by adding 2D-3D tie-point measurements (Section IV-B) from successful geo-registration. Note without 2D-3D tie-points, the original estimation only relies on inertial data and relative feature track measurements using the 3-stage method [9]. These 2D-3D tie-points provide absolute 3D information to correct accumulated drift in estimation, and improves the navigation solution dramatically, especially when flying over urban cites (such as the highlighted region).

2) The Use of Geo-Referenced Information : The navigation solution can be further improved by sharing 3D geo-referenced information between our two kinds of visual measurements: 2D-3D tie-points and feature tracks (Section

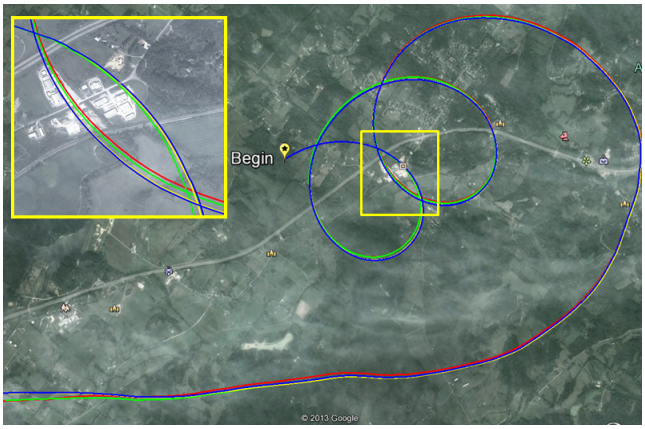

Fig. 6: The estimated trajectories for Scenario 2 (26.5 km, 7.5 min): Ground Truth (blue), result without geo-registration (red, 3D RMS error: 106.88 meters), result with geo-registered feature tracks (green, 3D RMS error: 38.92 meters), and result with 2D-3D tie-points from geo-registration (yellow, 3D RMS error: 10.52 meters). The highlighted region is enlarged for visualization.

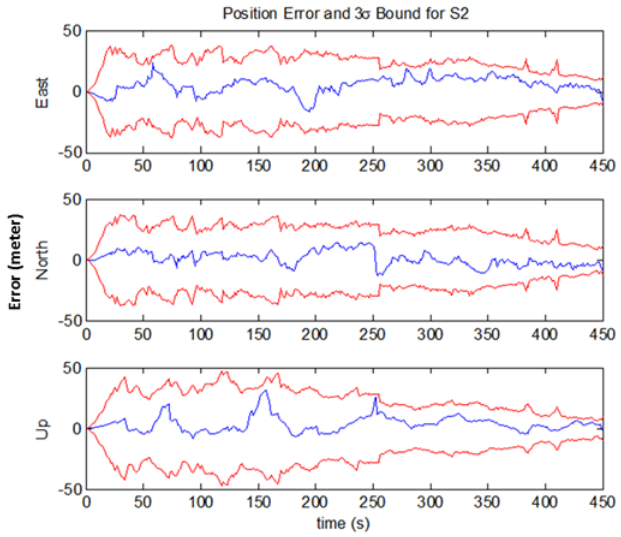

Fig. 7: Position error and 3 sigma bound in ENU frame for Scenario 2. Our smoother-based estimator is consistent for this scenario.

IV-C). Compared to the geo-registration process, our data association process (Figure 3) is more robust. The quality of feature tracking, including both the number of tracked features and the tracked length of features, is typically high in our system.

Incorporating 3D geo-referenced values into feature track measurements therefore extends the influence of absolute information to many navigation states which are connected by same feature tracks. It propagates geo-referenced information to future frames without direct geo-registration correction. For this scenario, 3D RMS error of the estimated trajectory can be further reduced from 13.98 meters to 9.83 meters, by utilizing 3D geo-referenced information in feature track formulation.

B. Scenario 2

This data set is collected using an aircraft that flies over mostly urban regions. The airplane travels total 26.5 kilometers in 7.5 minutes. The sensor set, sensor configuration, and system setting are all the same as in Section VI-A. However, the geo-registration process is more robust because there are discriminative features observed in more camera video frames during navigation.

1) The Improvement from Each Measurement Model: Since the successful rate of geo-registration process is higher for this scenario, we set our system first focus on demonstrating only the improvement from using geo-registered values in feature track formulation. The 2D-3D tie-point measurements are not used. Figure 6 shows the 3D RMS error of the estimated trajectory is reduced from 106.88 meters to 38.92 meters, by incorporating 3D geo-referenced values in feature track factors (Section IV-C). Instead of underlying 3D local landmark estimation, this new formulation provides more accurate relative constraints across navigation states.

We then set our system to show only the improvement due to direct absolute corrections from geo-registration. Figure 6 demonstrates the 3D RMS error can be reduced from 106.88 meters to 10.52 meters, by adding 2D-3D tie-points from geo-registration. If we incorporate geo-referenced information in both visual measurement models (2D-3D tie-points and geo-registered feature tracks), the 3D RMS error can be further reduced to 9.35 meters, our best result on this scenario.

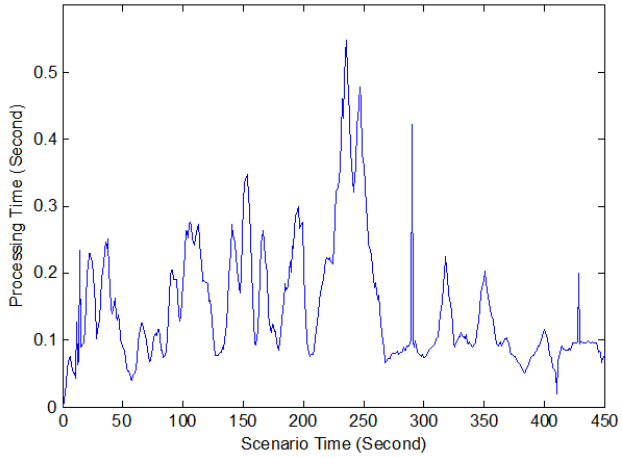

Fig. 8: The inference time along Scenario 2. Note the inference time is influenced by the dynamic number of measurement factors in the sliding window buffer, so it varies along the scenario.

2) Estimation Consistency : Figure 7 shows our smoother-based estimator (extended sliding-window factor graphs) is consistent for this scenario using inertial data and two kinds of geo-registered visual measurements, since all position errors are within 3 sigma uncertainty bounds. Note the successful rate of geo-registration process is high for this scenario, so absolute corrections are available very often during navigation to avoid the drift.

3) Computation Cost: Note we use a multi-threading architecture to support real-time navigation, and pre-processing

procedures such as geo-registration process are handled by different threads. Here we only show the processing time for our inference engine, without the pre-processing procedures. We use an incremental smoothing algorithm extended from iSAM2 [1] to perform periodic inference on our extended sliding-window factor graph framework with frequent relinearization. We set the update rate (inference frequency) as 1-Hz to generate navigation solutions and to accommodate delayed geo-registered measurements (see Section III-C) for all experiments. As shown in Figure 8, our incremental optimization takes less than 550 milliseconds on each update for this scenario to achieve real-time performance with satisfactory accuracy. All timing results were conducted on quad core Intel i7 CPU running at 2.70 GHz.

VII. Conclusions

In this paper, we present a new vision-aided aerial navigation approach which is capable of estimating precise absolute 3D pose using only inertial data and monocular cameras. Our approach treats each camera observation of a point feature as a single measurement, and tightly incorporates absolute geo-registered information into these visual measurements. We use a smoother-based inference framework called sliding-window factor graphs to fuse measurements from all sensors. This framework estimates states over a constant-length sliding window with fixed computational cost, and iteratively relinearizes highly nonlinear measurements for better estimation. We also extend this framework to accommodate delayed measurements from geo-registration process in an optimal way, by associating new factors with the corresponding navigation states when the measurement arrives within the smoothing window.

We introduce two kinds of visual measurements to incorporate geo-referenced information: 2D-3D tie-points and geo-registered feature tracks, based on our extended sliding-window factor graph framework. We model each 2D-3D tie-point, which is formed between a 2D feature on a camera video frame and the matched 3D geo-referenced point from the reference image, as an unary factor to provide 3D absolute information to update the navigation estimation. The 3D geo-referenced values from successful geo-registration process can also be utilized in our binary factor formulation for feature track measurements across consecutive video frames. This new formulation extends the influence of absolute information to many navigation states which are connected by same feature tracks. Experiments verify the importance of each visual measurement model, and demonstrate our approach provides accurate real-time solutions on large-scale aerial scenarios in GPS-denied setting.

Future work is to further enhance the quality of geo-registered visual measurements in our framework. Currently we only select high-quality (low-uncertainty) matches from geo-registration, and treat the 3D position of the matched point as a fixed quantity. However, same ground features may be detected and registered in geo-registration process more than once. We plan to store all geo-registered features in an online map. The 3D absolute positions for low-quality matches can be optimized in the map through multiple observations at different times. The uncertainty of the visual measurement for the same feature can then be decreased using multiple geo-registrations.

Acknowledgments

This material is based upon work supported by the DARPA All Source Positioning and Navigation (ASPN) Program under USAF/ AFMC AFRL Contract FA8650-13-C-7322. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the US government/ Department of Defense.

References

[1] M.Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," Intl. J. of Robotics Research, vol. 31, pp.217-236, Feb 2012.

[2] F. Lindsten, J. Callmer, H. Ohlsson, D. Tornqqvist, T. Schon, and F. Gustafsson, "Geo-referencing for UAV navigation using environment classification," in Proc. IEEE Intl Conf. on Robotics and Automation (ICRA), 2010.

[3] J. Farrell, "Aided navigation: GPS with high rate sensors," McGraw-Hill, 2008.

[4] T. Patterson, S. McClean, P. Morrow, and G. Parr, "Utilizing geographic information system data for unmanned aerial vehicle position estimation," in Proc. of Canadian Conf. on Computer and Robot Vision, 2011.

[5] D. Sim, R. Park, R. Kim, S. Lee, and I. Kim, "Integrated position estimation using aerial image sequences," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, 2002.

[6] G. Conte and P. Doherty, "Vision-based unmanned aerial vehicle navigation using geo-referenced information," in EURASIP Journal of Advances in Signal Processing, 2009.

[7] F. Kschischang, B. Fey, and H. Loeliger, "Factor graphs and the sum-product algorithm," IEEE Trans. Inform. Theory, vol. 47, no. 2, Feb 2001.

[8] A. Mourikis, N. Trawny, S. Roumeliotis, A. Johnson, A. Ansar, and L. Matthies, "Vision-aided inertial navigation for spacecraft entry, descent, and landing," IEEE Trans. on Robotics, vol. 25, no. 2, Apr 2009.

[9] H. Chiu, S. Williams, F. Dellaert, S. Samarasekera, and R. Kumar, "Robust vision-aided navigation using sliding-window factor graphs," in Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA), 2013.

[10] R. Kumar, H. Sawhney, J. Asmuth, A. Pope, and S. Hue, "Registration of video to geo-referenced imagery," in Proc. IEEE Intl. Conf. on Pattern Recognition (ICPR), 1998.

[11] A. Mourikis and S. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA), 2007.

[12] J. Sinko, "RTK performance in highway and racetrack experiments," Navigation, Vol. 50, No. 4, pp 265–275, 2003.

[13] K. Han, C. Aeschliman, J. Park, A. Kak, H. Kwon, and D. Pack, "UAV vision: feature based accurate ground target localization through propagated initializations and interframe homographies," in Proc. IEEE Intl Conf. on Robotics and Automation (ICRA), 2012.

[14] R. Kumar, S. Samarasekera, S. Hsu, and K. Hanna, "Registration of highly-oblique and zoomed in aerial video to reference imagery." in Proc. IEEE Intl Conf. on Pattern Recognition (ICPR), 2000.

[15] R. Kumar, H. Sawhney, S. Samarasekera, S. Hsu, H. Tao, Y. Guo, and K. Hanna, "Aerial video surveillance and exploitation." in Proc. of the IEEE, vol. 89, no. 10, Oct. 2001.

[16] R. Wiles, D. Hirvonen, S. Hsu, R. Kumar, W. Lehman, B. Matei, and W. Zhao, "Video geo-registration: Algorithms and quantitative evaluation", in Proc. IEEE Intl Conf. on Computer Vision (ICCV), 2001