Vehicle Positioning in GNSS-Deprived Urban Areas by Stereo Visual-Inertial Odometry

Milad Ramezani, Student Member, IEEE and Kourosh Khoshelham

Abstract—Accurate and continuous positioning in global navigation satellite system (GNSS) deprived urban areas is crucial for autonomous vehicles and mobile mapping systems. To achieve this goal, we propose a stereo visual-inertial odometry approach using a multistate constraint Kalman filter (MSCKF). In contrast with the conventional MSCKF, in which an inertial navigation system (INS) propagates the vehicle motion, and then the propagation is corrected by measurements of salient features extracted from images of a single camera, we update the propagation with observations extracted from images of a stereo pair of cameras. This way, additional constraints across the stereo pairs of images are exploited to improve the pose estimation. Experimental results on several KITTI datasets show that the stereo MSCKF outperforms the mono one achieving an average positioning error of 0.9% of the trajectory length compared to 1.3% for the mono approach and 2.5% for INS-only integration. These results show that visual-inertial odometry has a promising potential for vehicle positioning in short periods of GNSS signal outage.

Index Terms—Positioning, visual-inertial odometry, camera pose estimation, INS, Kalman filter.

I. INTRODUCTION

Accurate and continuous positioning of vehicles is a main requirement of autonomous vehicles and mobile mapping systems. Global Navigation Satellite System (GNSS) is a key technology to meet this demand in outdoor environments. However, in densely built-up areas such as urban canyons and through tunnels, GNSS observations are not properly available. In such areas, accurate and reliable vehicle positioning becomes a challenge.

Inertial sensors, which are typically integrated with GNSS, can provide position estimates in the absence of GNSS observations. However, positioning by an Inertial Navigation System (INS) is characterized by drift, i.e., the positioning errors accumulate and grow largely over time. This makes the INS unsuited for positioning even in short periods of GNSS signal loss.

Recent advances in the manufacturing of optical sensors have made it possible to build small, light, and inexpensive digital cameras which can be installed in many manned or unmanned platforms. The low cost, weight, and power consumption of these optical sensors make them ideal for aiding inertial navigation, in areas where GNSS has weak coverage such as urban canyons.

Inertial and visual sensors have complementary characteristics. INS can provide an initial estimate of the camera pose at each image acquisition, and visual features can constrain the INS drift. Moreover, there is no guarantee that navigation systems purely based on vision can provide continuous positioning due to the fact that they are highly dependent on visual features and therefore perform poorly in featureless environments.

The general approach to navigation based on the integration of cameras with inertial sensors is known as Visual-Inertial Odometry (VIO). A large body of research has been done on VIO in robotics and computer vision [1]–[9]. The algorithms which have been used to integrate inertial measurements with image observations are mainly classified into two groups: optimization-based approaches and filtering methods. Optimization algorithms estimate unknown parameters through an iterative minimization of a least squares error function [5], [10], [11]. Despite the fact that these algorithms provide better estimation, they have to confine the available measurements in order to achieve real-time pose estimation. On the contrary to optimization methods, filtering algorithms estimate parameters in a recursive fashion as soon as image observations become available. In the past few decades, many filtering methods have been proposed [1], [2], [6], [7], [9], [12]–[15] to solve the real-time problem. These methods are either computationally expensive, and therefore unsuitable for real-time positioning, or do not fully exploit the constraints that a tracked feature provides [1]. An exception, however, is the MSCKF, which employs the maximum geometric constraints that a feature provides from multiple camera poses [16].

In the conventional MSCKF, an INS propagates the motion and then the propagation is updated by observations of salient features tracked along the entire sequence of images captured by a single camera. A challenge in this approach is the low reliability of features tracked in a single image sequence, which affects the accuracy of pose estimation. In order to solve this issue, in this paper, we propose a stereo visual-inertial odometry using MSCKF. While matching features along the sequential images constraints the estimated pose, their matching across the stereo pair of images provides an additional constraint resulting in an improved pose estimation accuracy. We evaluate the
stereo MSCKF and compare its performance with the mono one through experiments on several KITTI datasets [17].

The paper proceeds with a review of main visual and visual-inertial navigation approaches to vehicle positioning without GNSS in Section II. Section III presents our method for positioning based on stereo visual-inertial odometry using MSCKF. Experiments and results are discussed in Section IV. Finally, conclusions are drawn and possible directions for future research are discussed in Section V.

II. RELATED WORK

Visual Odometry (VO) first proposed by Nister et al. [18] is a popular approach to motion estimation using one or more cameras. In a typical VO approach, salient image features are detected and matched between pairs of sequential images. Relative poses between the images are estimated by using the matched feature points, and these are then refined by using 3D points reconstructed from the visual features. Because of this incremental pose estimation, VO suffers from trajectory drift, although the magnitude of the drift is smaller than inertial dead reckoning which is based on the INS-only integration. Besides pose estimation, the visual observations can also be used to simultaneously construct a map of the area where the system is operating. This approach is commonly referred to as Simultaneous Localization and Mapping (SLAM) [19]–[23]. Using SLAM, the accuracy of less than 1 meter is achievable [24]. However, VO and SLAM are unlikely to achieve continuous positioning in environments which lack sufficient texture and geometric features.

In order to provide continuous motion estimation, visual observations can be integrated with inertial measurements. This approach is known as Visual-Inertial Odometry (VO) [25]. Several works have been done to integrate cameras with an INS. Strelow and Singh [10] define two algorithms, one in a batch process and another in a recursive algorithm, to estimate the motion and scene structure simultaneously. The batch process requires that entire measurements from images be available before computation begins, therefore it is not suited for real-time operation. Although the recursive method is suitable for real-time motion estimation, it only uses the scene points visible in the most recent image to update the state, thereby resulting in loss of information.

To exploit visual observations from multiple images while ensuring real-time operation, Leutenegger et al. [11] present an approach to tightly couple image observations with readings from an INS in a SLAM framework. In a non-linear optimization, the reprojection error of visual features is integrated with an INS error component in a fully probabilistic manner. Considering some frames as keyframes, the old states are partially marginalized to maintain a bounded-sized optimization window. This approach is introduced as Sliding Window Filter (SWF) in [5]. The shortcoming of this method is that not all the measurements of a visible feature are exploited, because the old states are discarded.

A popular estimator for the incorporation of visual and inertial measurements is the Extended Kalman Filter (EKF) [26]. In an EKF-based method, the state vector including motion parameters as well as INS biases is propagated according to a linear system model. The propagation is later updated by image observations through a measurement model. When the 3D position of scene points are estimated along the motion parameters, the approach is known as EKF-SLAM [7], [13], [27]. EKF-SLAM benefits from the correlations between camera motion and 3D position of visible features to yield better estimates. However, exploiting these correlations in a correct form imposes computational complexity affecting real-time operation in areas with many features [16].

In comparison to EKF-SLAM, algorithms that exploit image observations without jointly estimating the position of features to drive constraints between pairs of images are computationally more efficient. One of the constraints is driven from epipolar geometry that connects the current frame with the previous frame. Many works use epipolar constraints for the integration of visual measurements with readings from an INS [25], [28], [29]. The main drawback of the pairwise approaches is that the constraints are just between pairs of sequential images, i.e. not all constraints are considered at once, therefore resulting in a loss of information for the features that are observed from multiple poses. In order to resolve this issue, Mourikis and Roumeliotis [1] developed a Multi-State Constraint Kalman Filter (MSCKF) which imposes constraints between multiple camera poses. In MSCKF, when a scene feature is observed from multiple sequential camera poses, the state including motion parameters such as velocity, position, and INS biases as well as the camera poses from which the feature is visible, is updated by a measurement model which expresses the geometric constraints. The linear complexity of this algorithm in the number of features makes it suitable for real-time and precise motion estimation in large scale real world environments [1], [16]. Despite the fact that MSCKF exploits the maximum potential of visual information, it is highly sensitive to the accuracy of feature measurements. Because the 3D positions of features are estimated over a single sequence of images, not all measurements are sufficiently precise to contribute to the update step, resulting in less accurate motion estimation.

To improve the estimation of 3D position, Usenko et al. [30] develop a visual-inertial odometry with stereo cameras. Using stereo cameras, depth information is obtained from both static and temporal stereo. Static stereo refers to the fixed-baseline images of the stereo camera, whereas temporal stereo relates to images from the same camera taken at different camera poses in time. Experimental evaluation of this system shows better performance in rapid motion and substantial illumination changes compared to state-of-the-art keypoint-based methods. Moreover, combining static with temporal stereo, the system exploits the reconstruction of very small and very large-scaled environments at the same time to provide better motion estimation. Nevertheless, because the optimization used in this system is non-linear, all state variables have to be marginalized out except the current image, its predecessor, and its reference keyframe to achieve real-time performance.

Overall, the review of the literature shows the lack of a visual-inertial odometry method that exploits the maximum potential of visual measurements taken from more than one camera. Motivated by the potential of static stereo in more
accurate reconstruction and using the maximum potential of geometric constraints extracting from observing a feature, we present a stereo visual-inertial odometry using MSCKF.

III. STEREO MULTI-STATE CONSTRAINT KALMAN FILTER

Though the MSCKF takes full advantage of the constraints that a scene point provides, it has been used only for the integration of readings from an INS and visual observations taken from a single camera. To improve the quality of feature tracking and effectiveness of measurements in the update step, we exploit the fixed baseline between the pair of cameras in the stereo configuration for imposing additional constraints on the motion estimation.

Fig. 1 represents the principle steps of stereo MSCKF. Features are extracted and matched between the left and right camera images. The state is augmented with the pose parameters of the left camera. Given the calibration parameters between two cameras, the pose of the right camera is calculated from the left one. Finally, the state is updated by the measurements obtained from both cameras at each timestep.

A. Formulation of MSCKF State Vector

The state vector of INS, \( \mathbf{X}_{\text{INS}} \in \mathbb{R}^{16} \), is described as:

\[
\mathbf{X}_{\text{INS}} = \left[ \mathbf{q}_E^T \varepsilon \mathbf{v}_B^T \mathbf{p}_B^T \mathbf{b}_u^T \mathbf{b}_g^T \right]^T
\]

where, \( \mathbf{q}_E^B \) is the \( 4 \times 1 \) unit quaternion representing the rotation of the INS resolved in the body frame \( \{B\} \), vectors \( \varepsilon \mathbf{v}_B \), and \( \varepsilon \mathbf{p}_B \) are the linear velocity and the position of the INS with respect to the Earth Centered Earth Fixed (ECEF) frame \( \{E\} \), respectively. The \( 3 \times 1 \) vectors \( \mathbf{b}_u \) and \( \mathbf{b}_g \) express the gyroscope and accelerometer biases, respectively. Following (1), the INS error-state, \( \delta \mathbf{X}_{\text{INS}} \in \mathbb{R}^{15} \) is given by:

\[
\delta \mathbf{X}_{\text{INS}} = \left[ \delta \mathbf{\Phi}_B^T \varepsilon \delta \mathbf{v}_B^T \varepsilon \delta \mathbf{p}_B^T \delta \mathbf{b}_u^T \delta \mathbf{b}_g^T \right]^T
\]

where, the \( 3 \times 1 \) vector \( \delta \mathbf{\Phi}_B \) defines the perturbation of the INS attitude in the body frame. This vector can be linearly approximated by the quaternion error evolution [31]:

\[
\delta \mathbf{q} \approx \begin{bmatrix} 1 - \frac{1}{2} \varepsilon \delta \Phi^T \end{bmatrix}
\]

The full EKF state vector includes the state vector of INS, and \( M \) camera poses from which visual features are visible. This \((16 + 7M) \times 1 \) EKF state vector is defined as:

\[
\mathbf{X} = [\mathbf{X}_{\text{INS}}^T \mathbf{q}_E^T \varepsilon \mathbf{p}_C^T \ldots \mathbf{q}_E^T \varepsilon \mathbf{p}_C^T \ldots \mathbf{q}_E^T \varepsilon \mathbf{p}_C^T]^T
\]

Errors involved in the INS error-state propagation include attitude error, velocity error, position error, and biases.

1) Attitude error propagation [31]:

We describe the attitude-error propagation resolved in the body frame.

\[
\delta \mathbf{\Phi}_B \approx - (\mathbf{\Phi}_B \times \delta \mathbf{\Phi}_B - \mathbf{d}_g)
\]

where \( \delta \mathbf{\Phi}_B \) is the time derivative of the attitude error, the vector \( \mathbf{\Phi}_B \) is the estimated angular velocity resolved in the body frame. \( \mathbf{\Phi}_B \) is obtained from the sensed angular velocity, \( \omega_m \), and the estimated gyroscopic bias, \( \mathbf{d}_g \):

\[
\mathbf{\Phi}_B \approx \mathbf{\omega}_m - \mathbf{d}_g
\]

2) Velocity error propagation [32]:

The velocity error, where it is referenced to and resolved in the ECEF frame, is described as:

\[
\varepsilon \delta \mathbf{v}_B^T \approx - (\mathbf{C}_B^E \hat{\mathbf{f}}) \times \mathbf{C}_B^E \delta \mathbf{\Phi}_B - 2 \Omega_e \varepsilon \delta \mathbf{v}_B^T
\]

\[
+ \frac{2g_0(\hat{L}_b)}{\varepsilon r_S(\hat{L}_b)} \varepsilon \mathbf{p}_B \varepsilon \mathbf{p}_B^T \delta \mathbf{p}_B - \mathbf{C}_B^E \mathbf{b}_u
\]

where, \( \mathbf{C}_B^E \) is the rotation matrix from frame \( \{B\} \) to frame \( \{E\} \), \( \hat{\mathbf{f}} \) is the estimated specific force in the body frame related to the measured specific force, \( \mathbf{f}_m \), and the estimated accelerometer bias, \( \mathbf{b}_u \), by:

\[
\hat{\mathbf{f}} = \mathbf{f}_m - \mathbf{b}_u
\]

The matrix \( \Omega_e \) is the skew-symmetric of the Earth-rotation vector, \( \| \mathbf{x} \| \) is the Euclidean norm of the vector \( \mathbf{x} \). The scalar parameter, \( g_0(\hat{L}_b) \) is the surface gravity calculated from Somigliana model [33], and \( \varepsilon r_S(\hat{L}_b) \) is the geocentric radius at the surface given by [32]:

\[
\varepsilon r_S(\hat{L}_b) = R_E (\hat{L}_b) \sqrt{\cos^2 \hat{L}_b + (1 - \varepsilon^2)^2 \sin^2 \hat{L}_b}
\]

In the previous expression, \( \hat{L}_b \) is the geodetic latitude of the body, and \( R_E \) is the transverse radius of curvature.
3) Position error propagation [32]:

The time derivative of the position error in frame \( \{ E \} \) is simply the velocity error described as:

\[
\dot{\varepsilon} \delta \mathbf{p}_B^T = \varepsilon \dot{\mathbf{v}}_B^T
\]

(12)

4) Biases propagation:

The accelerometer and gyroscope biases are assumed not to vary in time. Thus:

\[
\dot{b}_a = 0, \quad \dot{b}_g = 0
\]

(13)

C. Linearised System Model

The linearised continuous system model defined for the INS error-state, \( \mathbf{X}_{INS}(t) \) is:

\[
\delta \mathbf{X}_{INS}(t) = \mathbf{F}(t) \delta \mathbf{X}_{INS}(t) + \mathbf{G}(t) \mathbf{W}_{INS}(t)
\]

(14)

where, \( \delta \mathbf{X}_{INS} \) is the time derivative of the INS error-state, \( \mathbf{F} \) is the system matrix, with respect to the ECEF frame given by:

\[
\mathbf{F} = \begin{bmatrix}
-I_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3} \\
0_{3,3} & 0_{3,3} & \mathbf{C}_{E}^T & 0_{3,3} \\
0_{3,3} & 0_{3,3} & 0_{3,3} & \mathbf{I}_{3,3} \\
0_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3}
\end{bmatrix}
\]

(15)

where,

\[\mathbf{F}_{21} = [-\mathbf{C}_{B}^T \varepsilon \mathbf{p}_B], \quad \text{and} \quad \mathbf{F}_{23} = \frac{2\varepsilon}{t_s(t_s + t_\sigma)} \mathbf{p}_B^T \varepsilon \mathbf{p}_B^T \]

(32), the \( \mathbf{G} \) matrix and \( \mathbf{W}_{INS} \) reflect system noise and its propagation in time. Assuming the system noise is zero-mean, white Gaussian, modelling the INS measurement noises, i.e. \( \mathbf{W}_{INS} = \mathbf{w}_g \mathbf{w}_r \mathbf{w}_{bdg} \mathbf{w}_{bda} \mathbf{w}_a \), where \( w_g, w_r, w_{bdg}, \) and \( w_{bda} \) are gyro and accelerometer noises and their corresponding bias variations, respectively, the matrix \( \mathbf{G} \) is given by:

\[
\mathbf{G} = \begin{bmatrix}
-I_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3} \\
0_{3,3} & 0_{3,3} & -\mathbf{C}_E^T & 0_{3,3} \\
0_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3} \\
0_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3} \\
0_{3,3} & 0_{3,3} & \mathbf{I}_{3,3} & 0_{3,3} \\
0_{3,3} & 0_{3,3} & 0_{3,3} & 0_{3,3}
\end{bmatrix}
\]

(16)

D. Discrete-Time Implementation

Consider that \( t_s \) is the INS sampling time, the continuous time equation is transformed into a discrete-time equation as [32]:

\[
\mathbf{X}_k = \mathbf{\Pi}_{k,k-1} \mathbf{X}_{k-1}
\]

(17)

where, \( \mathbf{\Pi}_{k,k-1} \) is the \( 15 \times 15 \) transition matrix from step \( k - 1 \) to \( k \) linked to the system matrix, \( \mathbf{F} \) by:

\[
\mathbf{\Pi}_{k,k-1} \approx \mathbf{I} + \mathbf{F}(t_k, t_{k-1}) t_s
\]

(18)

Furthermore, the EKF covariance matrix, which has to be propagated is introduced as:

\[
\mathbf{P}_{k|k} = \begin{bmatrix}
\mathbf{P}_{I_{k,k}} & \mathbf{P}_{I_{C,k}} \\
\mathbf{P}_{C_{I,k}} & \mathbf{P}_{C_{C,k}}
\end{bmatrix}
\]

(19)

where \( \mathbf{P}_{I_{k,k}} \) is the \( 15 \times 15 \) covariance matrix of the evolving INS state, \( \mathbf{P}_{C_{C,k}} \) is the \( 6M \times 6M \) covariance matrix of the camera pose parameters, \( \mathbf{\xi}_C = [\mathbf{p}_C^T, \mathbf{p}_C^T]_c, i = 1 \cdots M \), and \( \mathbf{P}_{I_{C,k}} \) is the correlation between the errors in the INS state and the camera pose estimates. Using this notation, the propagated covariance matrix is given by [16]:

\[
\mathbf{P}_{k+1|k} = \begin{bmatrix}
\mathbf{P}_{I_{k+1,k}} & \mathbf{P}_{I_{C,k}}(k+1,k) \\
\mathbf{P}_{C_{I,k}}(k+1,k)^T & \mathbf{P}_{C_{C,k}}
\end{bmatrix}
\]

(20)

E. State Augmentation

As soon as an image is recorded, camera pose parameters are added to the state to augment the state. These parameters are related to the INS pose estimate By:

\[
\mathbf{q}_C^E = \mathbf{q}_S^E \otimes \mathbf{q}_C^B
\]

(21)

where \( \mathbf{q}_C^E \) is the unit quaternion representing rotation from frame \( \{ E \} \) to the camera frame, \( \{ C \} \), and \( \mathbf{q}_S^E \) defines the rotation from the body frame to the camera frame computed off-line from the calibration between camera and INS. This calibration is possible by ceaseless-time batch estimation and continuously incorporating time offset between outputs of the INS and the camera within a maximum likelihood estimator [34]. The symbol, \( \otimes \), denotes the quaternion multiplication.

The position of camera with respect to the ECEF frame is also defined as [16]:

\[
\varepsilon \mathbf{p}_C^E = \varepsilon \mathbf{p}_S^B + \mathbf{C}_E^B \mathbf{p}_C^E
\]

(22)

where \( \varepsilon \mathbf{p}_C^E \) is the position of camera with respect to the ECEF frame and \( \mathbf{C}_E^B \mathbf{p}_C^E \) is the displacement vector between the camera frame and the body frame obtained through calibration process.

Eventually, the covariance matrix of the EKF is augmented, accordingly [16]:

\[
\mathbf{P}_{k|k} \equiv \begin{bmatrix}
\mathbf{I}_{6N+15} & \mathbf{J} \\
\mathbf{J}^T & \mathbf{I}_{6N+15}
\end{bmatrix}
\]

(23)

where \( \mathbf{J} \) is the Jacobian matrix of (21) and (22) with respect to the state vector at timestep \( k \) given by [16]:

\[
\mathbf{J} = \begin{bmatrix}
\mathbf{C}_E^B & 0_{3,3} & 0_{3,3} & 0_{3,6} & 0_{1,6N}
\end{bmatrix}
\]

(24)

F. Measurement Model

In a standard EKF, the measurement vector, \( \mathbf{y} \) is described as a linear model of the state and the corresponding white noise sources:

\[
\mathbf{y} = \mathbf{H}\delta\mathbf{X} + \mathbf{n}
\]

(25)

where, \( \mathbf{n} \) is white noise, and \( \mathbf{H} \) is the measurement Jacobian matrix determined from the known properties of the system.

Considering a single feature observed from a set of \( N_f \) camera poses, \( i \in \mathcal{S} \), the position of feature \( j \) in the frame \( \{ C_i \} \) of a
perspective camera is defined as:

$$c_i \mathbf{p}_f^T_j = \begin{bmatrix} c_i, X_j \\ c_i, Y_j \\ c_i, Z_j \end{bmatrix} = C_{E}^{c_i} \left( \hat{\mathbf{p}}_f^T_j - \hat{\mathbf{p}}_E^T \right)$$ (26)

where, $c_i \mathbf{p}_f^T_j$ is the position of feature $f_j$ with respect to the $i$th camera frame, $C_{E}^{c_i}$ is the rotation matrix from the Earth frame to the $i$th camera frame, $\hat{\mathbf{p}}_f^T$ is the position of $i$th camera in global frame, and $\hat{\mathbf{p}}_E^T$ is the position of feature $f_j$ in frame $\{E\}$, which is unknown and must be estimated (cf. Appendix).

Suppose that the camera is pre-calibrated, the visual measurement model is described as:

$$z_{i}^{(j)} = \frac{1}{\sqrt{Z_i}} \begin{bmatrix} c_i, X_j \\ c_i, Y_j \end{bmatrix} + \eta_{i}^{(j)}$$ (27)

where, $z_{i}^{(j)}$ is the measurement of feature $f_j$ in the perspective camera frame $C_i$. $\eta$ is the $2 \times 1$ noise vector on image, with corresponding covariance matrix $R = \sigma_{\eta}^2 \mathbf{I}_{2 \times 2}$.

Having estimated the global position of the feature, the measurement residual for the left camera, which is augmented in the state, is described as:

$$r_{l_i}^{(j)} = z_{l_i}^{(j)} - z_{i}^{(j)}$$ (28)

where, $z_{l_i}^{(j)}$ is the measurement of feature $f_j$ in the left camera frame $C_{l_i}$ calculated from (27). Linearising the residuals with respect to camera pose parameters and the 3D position of features, the previous equation is expressed as:

$$r_{l_i}^{(j)} \simeq H_{l_i}^{(j)} \delta X + H_{l_i}^{(j)} \varepsilon \hat{\mathbf{p}}_f + n_{l_i}^{(j)}$$ (29)

where, $\varepsilon \hat{\mathbf{p}}_f$ is the error in the position estimate of $f_j$, and $H_{l_i}^{(j)}$ are the Jacobian matrices of measurement $z_{l_i}^{(j)}$ with respect to the state and the feature position, respectively, given by:

$$H_{l_i}^{(j)} = \begin{bmatrix} 0 & J_{l_i}^{(j)}(c_i, \hat{\mathbf{p}}_f, \wedge) & -J_{l_i}^{(j)} C_{E}^{c_i} & 0 \end{bmatrix}$$ (30)

$$H_{l_i}^{(j)} = J_{l_i}^{(j)} C_{E}^{c_i}$$ (31)

In (30), the dimension of the left and the right zeros are $2 \times (15 + 6(i - 1))$ and $2 \times 6(N - i)$ respectively. The Jacobian matrix $J_{l_i}^{(j)}$ is expressed by:

$$J_{l_i}^{(j)} = \frac{1}{(c_i, Z_i)^2} \begin{bmatrix} c_i, Z_j & 0 & -c_i, \hat{X}_j \\ 0 & c_i, Z_j & -c_i, \hat{Y}_j \end{bmatrix}$$ (32)

Now we define the pose of the right camera, $\xi c_i = [q_{C_i}^{T}, \varepsilon \mathbf{p}_C^{T}]^T$, with respect to the left camera pose and the relative calibration parameters between them:

$$\varepsilon \mathbf{p}_C^{T} = \varepsilon \mathbf{p}_C^{T} - \dot{C}_{E}^{c_i} C_{C_i}^{T} C_{E}^{c_i} \mathbf{p}_{C_i}$$ (33)

where, $C_{C_i}^{c_i}$ and $C_{C_i}$ are the rotation matrix and displacement vector between the left and the right camera, respectively. Matrix $\dot{C}_{C_i}$ represents the rotation between the left camera and the Earth frame.

$$q_{E}^{c_i} = q_{C_i}^{c_i} \otimes q_{E}^{c_i}$$ (34)

where, the unit quaternion $q_{E}^{c_i}$ represents the rotation from the left camera coordinate system to the right one. We also obtain the position of feature $f_j$ with respect to the right camera frame at $i$th pair of camera:

$$c_i \mathbf{p}_f = C_{E}^{c_i} (\varepsilon \mathbf{p}_f - \varepsilon \mathbf{p}_C) = C_{C_i}^{c_i} c_i \mathbf{p}_f + c_i \mathbf{p}_C$$ (35)

Now, the linearisation of measurement residuals in (28) around the estimates for the left camera and the feature position is defined as:

$$r_{l_i}^{(j)} \simeq H_{l_i}^{(j)} \delta X + H_{l_i}^{(j)} \varepsilon \hat{\mathbf{p}}_f + n_{l_i}^{(j)}$$ (36)

where, $H_{l_i}^{(j)}$ and $H_{l_i}^{(j)}$ are the Jacobians of measurement $f_j$ from the right camera pose with respect to the error state and the global position of the feature, respectively, given by:

$$H_{l_i}^{(j)} = \begin{bmatrix} 0 & J_{l_i}^{(j)}(c_i, \hat{\mathbf{p}}_f, \wedge) & -J_{l_i}^{(j)} C_{E}^{c_i} C_{E}^{c_i} & 0 \end{bmatrix}$$ (37)

$$H_{l_i}^{(j)} = J_{l_i}^{(j)} C_{E}^{c_i} C_{E}^{c_i}$$ (38)

By stacking the residuals $r_{l_i}^{(j)}$ and $r_{l_i}^{(j)}$ separately, we obtain a $4N_j \times 1$ residual vector, $r^{(j)} = [r_{l_i}^{T} r_{l_i}^{T}]^T$, where $N_j$ is the number of stereo camera poses in which feature $f_j$ is visible. By combining (29) and (36), we form the block of all residuals as:

$$r^{(j)} \simeq H_{X}^{(j)} \delta X + H_{f}^{(j)} \varepsilon \hat{\mathbf{p}}_f + n^{(j)}$$ (39)

where,

$$H_{X}^{(j)} = \begin{bmatrix} H_{X}^{(j)} \mathbf{H}_{f}^{(j)} \mathbf{n}^{(j)} \end{bmatrix}, \quad \mathbf{H}_{f}^{(j)} = \begin{bmatrix} H_{f}^{(j)} \mathbf{n}^{(j)} \end{bmatrix}, \quad \mathbf{n}^{(j)} = \begin{bmatrix} n_{l_i}^{(j)} \end{bmatrix}$$ (40)

Since the error state, $\delta X$, correlates with the error of feature position, $\varepsilon \hat{\mathbf{p}}_f$, the residuals cannot be directly applied in an EKF estimation. To resolve this issue, the residual vector, $r^{(j)}$ is projected onto the left null-space of the $4N_j \times 3$ matrix $H_{f}^{(j)}$ [16]:

$$A^T r^{(j)} \simeq A^T H_{X}^{(j)} \delta X + A^T n^{(j)}$$ (41)

where, $A$ is a $4N_j \times (4N_j - 3)$ unitary matrix with the columns from the basis of the left null-space of $H_f$. With a slight change in notation, we define the standard measurement model used is an EKF as:

$$r^0_0 = H_{0}^{(j)} \delta X + n_0^{(j)}$$ (42)

where,

$$r^0_0 = A^T r^{(j)}, \quad H_{0}^{(j)} = A^T H_{X}^{(j)} \mathbf{n}^{(j)} = A^T n^{(j)}$$ (43)

projecting $r^{(j)}$ onto the left null-space of the matrix $H_{f}^{(j)}$, the term related to the error in the position estimate of $f_j$ is cancelled out, and $r_0$ can now be used in the filter update step.

G. EKF Update

Having matched salient features from pairs of images in static stereo and tracked them over a sequence of camera poses, the observations are entirely used as geometric constraints to update
Lastly, the EKF state is updated by the Kalman gain matrix, \( \mathbf{K} \), given by:

\[
\mathbf{K} = \mathbf{P}_{k|k} \mathbf{H}^T \left( \mathbf{R}_{k|k} \mathbf{P}_{k|k}^T + \mathbf{R}_{\mathbf{n}} \right)^{-1}
\]

where, \( \mathbf{P} \) is the state covariance matrix, and \( \mathbf{R}_{\mathbf{n}} \) is the covariance matrix of the noise component, \( \mathbf{Q}_{\mathbf{n}}^T \mathbf{r}_0 \), computed from propagation of uncertainty, i.e. \( \mathbf{R}_{\mathbf{n}} = \mathbf{Q}_{\mathbf{n}}^T \mathbf{r}_0 \mathbf{Q}_{\mathbf{n}} \).

The state vector is corrected with the reduced measurement using:

\[
\mathbf{X}_{k+1|k+1} = \mathbf{X}_{k+1|k} + \mathbf{K} \mathbf{Q}_{\mathbf{n}}^T \mathbf{r}_0
\]

Simply, the posterior error covariance is updated with:

\[
\mathbf{P}_{k+1|k+1} = (\mathbf{I} - \mathbf{KR}_{k|k}) \mathbf{P}_{k+1|k} (\mathbf{I} - \mathbf{KR}_{k|k})^T + \mathbf{KR}_{k|k} \mathbf{Q}_{\mathbf{n}} \mathbf{K}^T
\]

where, \( \mathbf{I} \) is a \((6M + 15) \times (6M + 15)\) identity matrix.

IV. EXPERIMENTS AND RESULTS

To evaluate the performance of the VIO approach in the mono and stereo mode, an experiment was conducted using KITTI datasets 0001, 0005, and 0020 [17]. Datasets 0001 and 0005 contains 108 and 154 pairs of images, respectively, both captured on 26 September 2011. Dataset 0020 consists of 330 pairs of images recorded on 30 September 2011. Sample frames from the datasets are shown in Fig. 2. The images in all datasets have the same size of 1392 \times 512, and were recorded at approximately 0.1 second intervals. The baseline between the two cameras in the static stereo is 54 cm which is sufficient for initial 3D point estimation. Moreover, the datasets include inertial measurements attained by an OXTS RT3003 inertial and GPS navigation system at a measurement rate of 100 Hz. The RTK GNSS measurements serve as the ground truth in these datasets. It is worth noting that the INS used by KITTI is a high quality inertial sensor and its inertial data has been purified regarding gravity, biases, and integration error. Therefore, it is expected that the INS only solution provides a good motion estimate competing with the visual-inertial integration.

The entire algorithm was implemented in MATLAB R2015a on an ordinary PC with a 3.4 GHz Intel(R) Core(TM)i7 processor and 8 GB RAM.

A. Feature Detection, Matching and Tracking

To extract and match features, we used Speeded Up Robust Features (SURF) [36]. In stereo vision, features were matched by applying the approximate matching method described in [37]. Having matched the features across the static stereo, we tracked the corresponding pairs of features along
Fig. 3. Features tracked in mono vision (Top) and stereo vision (Bottom) from KITTI dataset 0001.

Fig. 4. Statistics on the number of tracked features in each frame over the three datasets 0001, 0005, and 0020 for both mono and stereo modes.

The number of features tracked varies between the mono and stereo mode. Fig. 4 shows a boxplot of the number of tracked features in each frame over the three datasets 0001, 0005, and 0020 for both mono and stereo modes. The lower number of features in stereo mode is due to the fact that not all SURF features can be matched across the stereo pair. However, as it can be seen, in the stereo mode the variation in the number of tracked features is significantly less than that in the mono mode, which indicates that the tracking performs more consistently along the image sequence.

Table I summarizes the average number of all tracking features along the sequence of images in the mono and the stereo modes on the used datasets as well as their processing time averaged over 10 runs of the program. It is clear that cross matching rejects approximately half of the features tracked along the image sequence. However, the overall time of feature tracking in the stereo mode is almost two times more than that in the mono one.

### B. Accuracy of Stereo Visual-Inertial MSCKF

To assess the accuracy of the estimated trajectories, we used translational RMSE, defined as the Root Mean Square of the three translation errors at each timestep. The translational error at each timestep was calculated from the estimated position and the ground truth. The total translational error was measured as the Euclidean norm of all RMSEs, whereas ARMSE, noted in Table II, was calculated as the average of RMSEs. Since GNSS does not provide the attitude of the vehicle, there is no ground truth available for rotational RMSE. For this reason, we do not include the rotational accuracy of the stereo MSCKF in the evaluation.

Tracking salient features, the rectified image observations, i.e. features corrected with respect to the intrinsic parameters of the camera, were taken as input to estimate the 3D position using Gauss-Newton least-squares minimization [40]. In contrast to the mono MSCKF in which scene features are reconstructed based on observations over a sequence of images, we employed the observations from both the static stereo and multiple images...
over time. This way, more 3D point estimates meet the sufficient precision contributing to the EKF update step. In addition, to have the best initialization, we calculated the initial 3D position with the triangulation between the first and the last frame in which the feature is visible. Fig. 5 shows the trajectories obtained by the mono and the stereo VIO compared with the INS-only integration as well as the ground truth. While VIO in both modes performs much better than inertial positioning, the stereo VIO provides a more accurate trajectory than mono VIO in the used datasets.

Fig. 6 shows the cumulative distribution of translational errors for the proposed stereo method compared with the mono and INS only integration. The performances brought here are for the optimum camera pose window. Stereo MSCKF performs better than the other methods and reaches an average accuracy of 0.9% of the trajectory length compared to 1.3% and 2.5% for the mono and INS only integration, respectively. Note that the drift of INS-only integration is quite low because KITTI inertial data was obtained by a precise INS(OCTS RT3003), and has been filtered for biases and integration errors. Using a low-grade INS, it is anticipated that INS-only integration would result in a larger drift, therefore the benefit of VIO would be more apparent.

Table II summarizes the performance of stereo MSCKF compared with mono and INS on the basis of Translational Average Root Mean Squared Error (ARMSE) and RMSE values over the entire sequence of frames for the tested KITTI datasets.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduced a stereo visual inertial odometry method based on Multi State Constrained Kalman Filter for vehicle positioning in GNSS-deprived environments. The principle advantage of this method is that, in addition to using image observations along a single sequence of images, we exploit the epipolar constraint across the static stereo pairs of images captured by a stereo camera system. Furthermore, for the 3D reconstruction, we exploit observations from both static stereo and multiple images in time. This way, more robust features contribute to the update step resulting in a more accurate positioning. Experimental evaluation using three KITTI datasets shows that the stereo VIO outperforms the mono MSCKF despite the fact that the number of measurements in mono MSCKF is generally larger than that in the stereo VIO. This indicates that features tracked from both static and temporal stereo are more robust, therefore contributing to more accurate estimation of parameters.

Despite the fact that the presented VIO method reduces the trajectory drift, it cannot eliminate the drift completely. In practice, this system is able to provide positioning with sub-meter accuracy only in short periods of GNSS signal loss.

In future work, we explore global position estimation by matching georeferenced geometries from a 3D City Information Model (CIM) with features extracted from images captured by an omnidirectional camera. In addition, we investigate the
use of other features such as Oriented FAST and rotated BRIEF (ORB) [41] and Harris [42] and analyse the outcomes.

APPENDIX

In order to estimate the position of the tracked feature, \( f_j \), MSCKF uses an inverse-depth least square Gauss-Newton optimization. The input of the optimization is the rectified image coordinates of feature \( f_j \) and the corresponding camera poses.

If \( C_j \) is the first frame in which the feature \( f_j \) is visible, the position of the feature with respect to the camera frame \( \{C_i\} \) is given by:

\[
C_i \cdot p_{f_j} = C_{C_j}^{C_i} C_j \cdot p_{f_j} + C_{C_j}^{C_i} p_{C_j} = \left( C_{C_j}^{C_i} \begin{bmatrix} c_j \cdot X_j \\ c_j \cdot Y_j \\ c_j \cdot Z_j \end{bmatrix} \right) + C_{C_j}^{C_i} p_{C_j} \quad (51)
\]

By considering \( \alpha_j = \frac{c_j \cdot X_j}{c_j \cdot Z_j} \), \( \beta_j = \frac{c_j \cdot Y_j}{c_j \cdot Z_j} \), and \( \gamma_j = \frac{1}{C_j \cdot Z_j} \), (52) can be reformulated as:

\[
C_{C_j}^{C_i} p_{f_j} = \frac{1}{\gamma_j} \begin{bmatrix} h_{11} (\alpha_j, \beta_j, \gamma_j) \\ h_{12} (\alpha_j, \beta_j, \gamma_j) \\ h_{13} (\alpha_j, \beta_j, \gamma_j) \end{bmatrix} \quad (52)
\]

where, \( h_{11}, h_{12}, \) and \( h_{13} \) are scalar functions of \( \alpha_j, \beta_j, \) and \( \gamma_j \), respectively. Substituting (53) into (27), the model for perspective camera is expressed as:

\[
z_{i}^{(j)} = \frac{1}{h_{13}} \begin{bmatrix} h_{11} \\ h_{12} \end{bmatrix} + \eta_{i}^{(j)} \quad (53)
\]

The vector of camera measurement error used in the least-square system then is defined as:

\[
r_{i}^{(j)} (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) = z_{i}^{(j)} - \frac{1}{h_{13}} (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \begin{bmatrix} h_{11} (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \\ h_{12} (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \end{bmatrix} \quad (54)
\]

Computing the entire residuals of feature \( f_j \), the estimated value of \( \alpha, \beta, \) and \( \gamma \) can be obtained and finally, the 3D position of feature \( f_j \) can be estimated by Gauss-Newton Least Square minimization as:

\[
\varepsilon^{(j)} p_{f_j} = \frac{1}{\beta_j} C_{C_j}^{C_i} \begin{bmatrix} \hat{\alpha}_j \\ \hat{\beta}_j \end{bmatrix} + \varepsilon p_{C_j} \quad (55)
\]

REFERENCES


Kourosh Khoshelham received the Ph.D. degree in photogrammetry and laser scanning from the Hong Kong Polytechnic University, Hong Kong, in 2004. He is a Senior Lecturer with the Department of Infrastructure Engineering, The University of Melbourne, Melbourne, VIC, Australia. He has worked as an Assistant Professor with the University of Twente, Delft University of Technology, and University of Tehran. He is the Co-Chair of the ISPRS WG IV/5 on indoor/outdoor seamless modelling, LBS and mobility, Co-Chair of Indoor 3D 2017 in Wuhan, and Co-Chair of Research@Locate 2018, which will be held in Adelaide, Australia. His research interests include mobile mapping, positioning, and information extraction from images and point clouds.

Milad Ramezani received the B.S. degree in geomatics engineering from The University of Isfahan, Isfahan, Iran, and the M.S. degree in photogrammetry from K.N. Toosi University of Technology, Tehran, Iran, in 2009 and 2011, respectively. He is currently working toward the Ph.D. degree in visual-inertial odometry at the University of Melbourne, Melbourne, VIC, Australia. He has established a system titled “A vertical-flight surveyor” patented by the Iranian Emblem the Judiciary State Organization for registration of Deeds and Properties. He has also participated in many research projects including aerial and satellite image processing as well as surveying. His research areas of interest include SLAM, visual-based navigation, Kalman filter estimation, nonlinear optimization, image processing, and omnidirectional cameras.