A Learning-based Approach to Cover Short-term Camera Failure in a Monocular Visual Inertial Odometry System

Y. Tian* and M. Compere**

*Embry-Riddle Aeronautical University, Department of Mechanical Engineering
Daytona Beach, FL, USA tiany@my.erau.edu
** Embry-Riddle Aeronautical University, Department of Mechanical Engineering
Daytona Beach, FL, USA comperem@erau.edu

ABSTRACT

Localization and navigation have multiple established methods in sensing and fusion algorithms. Visual Inertial Odometry (VIO) has drawn attention recently, however, one disadvantage is camera susceptibility to disturbances such as fast motions, and moving objects. Existing researchers usually test their algorithms assuming good camera performance. In this research, we propose a learning-based method to estimate pose during brief periods of camera failure or occlusion. A Long Short-Term Memory (LSTM) network is trained during periods of good camera operation and, once trained, the LSTM provides an alternative pose estimate, available as soon as camera failure is detected. We tested our algorithm by removing the visual inputs and comparing Kalman Filter, IMU-only, and pre-trained LSTMs results. The results indicate the implemented LSTM increased the positioning accuracy by 76.2% and orientation accuracy by 26.5%.

Keywords: lstm, sensor fusion, localization, visual-inertial odometry

1 INTRODUCTION

Localization, the key to navigation, is a popular topic in the engineering research field. The commonly-used sensors are GPS (Global Positioning Sensor), Inertial Measurement Unit (IMU), Lidar and Camera. However, if only one sensor is utilized, its failure can easily lead to the entire navigation system's failure. In order to address this problem, a growing number of studies fuse multiple sensors to achieve optimal performance, such as Global Navigation Satellite System (GNSS) and Global Inertial Navigation System (GINS) using GPS and an inertial measurement unit (IMU), and Visual-Lidar odometry system and Visual-Inertial odometry (VIO) system that use camera with Lidar and IMU respectively. As augmented reality and virtual reality have become popular, visual-inertial odometry has drawn more attention. However, current loosely-coupled VIO studies focus on combining well-known visual odometry algorithms with inertial measurements, assuming the camera is functional during the entire navigation procedure. It is obvious that the camera's performance significantly degrades from disturbances like fast motion, lighting condition change and moving objects. Aside from degraded performance, a complete camera failure or brief occlusion can easily happen in the real world. There is very little research focused on this failure. In this study, we propose a learning-based approach to cover the short-term camera failure using LSTM in a monocular visual-inertial odometry system.

2 BACKGROUND

This section summarizes previous work in the literature on monocular visual-inertial odometry. There are two types of VIO methods: Tightly-coupled and Loosely-coupled. Tightly-coupled methods merge inertial measurements into the visual odometry calculation process. However, they usually need complex formulations. Loosely-coupled methods calculate visual and inertial odometry estimates individually, and then combine these two values together. In this study, our goal is to have simpler formulas that are more easily understood. We focus on Loosely-coupled methods and compare the proposed algorithm with a Kalman Filter.

2.1 Kalman Filter

For the loosely-coupled method, Kalman Filter is a commonly-used sensor fusion algorithm. Weiss and Siegwart [1] proposed a loosely-coupled approach with an Extended Kalman Filter (EKF). In their work, the metric scale estimation remains independent of the visual algorithm. Kelly and Sukhatme [2] used an Unscented Kalman Filter (UKF) that is better designed to handle non-linearities compared to EKF. Their solution self-calibrates the transformation between the camera and IMU, while simultaneously localizing the body and mapping the environment. Lynen et al. [3] designed a multi-sensor fusion in an EKF framework with Micro Aerial Vehicle (MAV) data that is robust to long term missions. All these filter-based implementations consider the visual odometry algorithm as an individual system and only use the pose and covariance result provided by it.

2.2 Deep Learning

Recently, artifically intelligence technology has been

developed rapidly, deep-learning based approaches have been applied to solve visual-inertial odometry problems and shown their advantages in robustness and accuracy. There are two main methods. The first one has a similar concept with the loosely-coupled methods. It uses the learning-based method only for visual or inertial part, respsectively. The other feeds both visual and inertial inputs into the neural network for training. Both methods have similar architecture, starting with a Convolutional Neural Network (CNN) for the feature extraction and feeding features to an Long-Short Term Memory (LSTM) neural network [4]-[7]. S. Wang et al [4] designs a CNN-LSTM architecture for visual odometry. This DeepVO algorithm tremendously enhances the visual odometry performance by 65.9% compared to another monocular VO method (VISO2_M). J. R. Rambach et al [6] trains an LSTM network with inertial measurements as inputs and ground truth as the target. After the LSTM model is well-trained, it can produce poses without the complex modeling of IMU biases, noises, and sensor calibration. Meanwhile, visual odometry is calculated using a feature-based method. A Kalman Filter is used to produce the visual-inertial odometry. VINet developed by R. Clark et al [7] is the first end-to-end trainable method. It uses images and IMU data as inputs to a CNN-RNN network for pose estimation and shows a significant improvement compared to the state-of-art VIO algorithms such as the LIBVISOvisual odometry algorithm and the EKF sensor fusion algorithm.

From the above summary, we can draw a conclusion that those algorithms work well during functional sensor operations. However, it is obvious that the final estimation procedure requires the involvement of visual inputs, which means that upon camera failure such as an occluded view from obstacles, there will be a high probability that the performance cannot meet the expectation. To cover these kinds of short-term camera failures and achieve higher stability in practical, we proposed a novel method using a Long Short-Term Memory (LSTM) neural network. LSTM neural network is a typical type of recurrent neural network (RNN) that can accommodate the sequential information. The benefit of LSTM is that it can address the vanishing gradient compared to RNN. Todd Barker and Martin Akerblad[8] trained an LSTM network using inertial measurements and ground truth, and the results show a significant improvement in translation. Figure 1 shows an LSTM network architecture. Compared with an ordinary RNN that has only one layer repeating throughout the tme, an LSTM neural network has three different gates: *(1)* forget gate, *(2)* input gate, and *(3)* output gate [8].

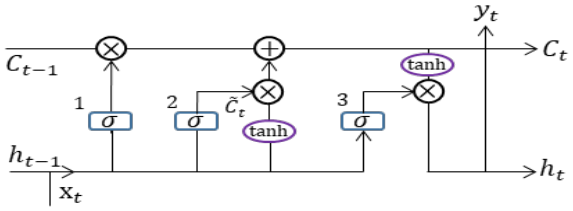

Figure 1. The Architecture of LSTM Network

2.3 Coordinate Frame

It is important to clarify the coordinate frames during localization and navigation. There are two main coordinates used in this study, body-fixed frame(b) and global frame(g). Body-fixed frame(b) is attached to the IMU. Global frame(g) can also be recognized as the navigation frame.

3 METHODOLOGY

This section describes the proposed LSTM-based odometry approach using inertial measurements. The designed architecture is shown in Figure 2.

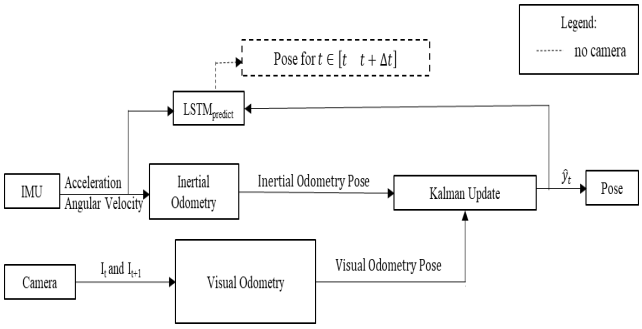

Figure 2. The LSTM VIO Architecture

In this study, visual odometry is calculated using a feature-based method. It extracts a set of features from each image and matches them across multiple frames. Camera pose can be estimated by minimizing the reprojection errors between feature pairs. Inertial odometry is calculated by integrating the acceleration and angular velocity vectors. Accelerometer senses acceleration in 3 axes, however, these measurements are always affected by sensor bias and sensor noise. Bias is not constant throughout time, but bias rate can be modeled as Gaussian White Noise. Commonly, sensor noise is modeled as Gaussian White Noise as well. Eq (1) presents the acceleration equation with these uncertainties:

$$a = R_{bg}(a_m - b_a - n_a) - g \tag{1}$$

where a is the linear acceleration, R_{bg} is the rotation matrix from the body-fixed frame to the global frame, a_m is the measured acceleration, g is the gravity vector, b_a is the sensor bias and n_a is the sensor noise.

The gyroscope can sense the 3 degree-of-freedom angular velocity. Similar to the accelerometer, the gyroscope measurements can be easily influenced by gyro bias and noise. The true gyro value can be modeled as Eq (2) with two similar uncertainty terms:

$$\omega = \omega_m - b_g - n_g \tag{2}$$

where ω is the angular velocity, ω_m is gyro measurements, b_g is gyro bias, and n_g is gyro noise.

The kinematic model of the IMU can be model using Eq (3) to (6).

$$P_g = \int v_g, v_g = \int a \tag{3}$$

$$\Omega = \int \omega \tag{4}$$

$$\dot{R} = S(\Omega)R \tag{5}$$

$$R = \begin{bmatrix} \cos\theta\cos\psi & -\cos\phi\sin\psi + \sin\theta\cos\psi\sin\phi & \sin\phi\sin\psi + \sin\theta\cos\psi\cos\phi \\ \cos\theta\sin\psi & \cos\phi\cos\psi + \sin\theta\sin\psi\sin\phi & -\sin\phi\cos\psi + \sin\theta\sin\psi\cos\phi \\ -\sin\theta & \cos\theta\sin\phi & \cos\theta\cos\phi \end{bmatrix} \tag{6}$$

Where P_g and v_g are the position and the velocity of the IMU in the global frame, R is the rotation matrix that relates the body frame to the global frame, $S(\Omega)R = \Omega \times R$. While the camera is functionally working, a Kalman Filter is used to fuse visual and inertial odometry. Kalman Filter predictions are used as the ground truth for the following training process.

In this research, our model is a 2-layer LSTM network with 6 inputs, 3 degree-of-freedom accelerations, and 3 degree-of-freedom angular velocities. We perform forward feed and back propagation operations for training. Once the LSTM is well-trained, it can be activated when a camera failure occures. The definition of a camera failure is the matched features between two sequential images for 5 visual-data update cycles. If the average number of the matched features for these 5 cycles was less than 30, we will assume the camera is experiencing a complete failure, and start to activate the LSTM model. The network takes inertial measurements as inputs and starts to provide alternative pose estimations.

4 EXPERIMENTS AND RESULTS

This section presents the detailed training process and the relevant analysis of results. The proposed LSTM network was trained on EuRoC MAV, and was tested on EuRoC MAV [9] and KITTI VO benchmark datasets [10].

4.1 Training

The training dataset used in this study is EuRoC MAV dataset [9] that contains stereo images and synchronized inertial measurements, and ground truth. As we focus on monocular visual-inertial odometry, only images collected by cam 0 are used here. This database contains 11 various sequences. In this study, 5 sequences (Vicon Room 1 01-03 and Vicon Room 2 01-02) for training and Vicon Room 2 03 for testing. The visual sensor is Aptina MT9V034 global shutter and WVGA image sensors. The visual sampling rate is 20 fps. The inertial sensor is ADIS16448, whose detailed specifications are listed in Table 1. The IMU sampling rate is 200 HZ.

Table 1. Specifications of ADIS 16448 Sensor.

Accelerometer	Range	± 250˚/s ± 500˚/s ± 1000˚/s
	Bias Stability （1σ）	14.5˚/hr
	Sampling Rate	200 HZ
Gyroscope	Range	± 18 g
	Bias Stability（1σ）	0.25 mg
	Sampling Rate	200 HZ

The visual and inertial odometry are calculated respectively, and fused with a Kalman Filter. If the camera provides high-quality images, this monocular visual-inertial odometry system outcome will be Kalman Filter pose estimation. These poses are also used as the training ground truth.

Table 2. Specifications of the Proposed LSTM

Learning Rate	0.25
Training Epoch	100
Batch Size	128

The network was trained in Python using TensorFlow. The computing capacity in this study is an Asus PC with an i5 Intel CPU and 32 GB memory. Table 2 presents the important factors for training. The proposed LSTM network was trained with 0.25 as the learning rate, 100 as the training epoch and 128 as the batch size. The loss function is defined as the Mean Square Error(MSE) of positions P_g and orientations Ω in Eq (7)

$$\theta^* = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{t} || P_j - \hat{P}_j ||_2^2 + || \Omega_j - \hat{\Omega}_j ||_2^2 \tag{7}$$

where N is the number of samples, P is the global positions, W are orientations, ||.|| is 2-norm, theta* is the optimal parameter of the LSTM network and the function argmin locates the theta that minimizes the loss function.

4.2 Results

To evaluate and validate this algorithm, we removed the visual inputs and compared Kalman Filter, IMU-only, and pre-trained LSTMs results. The performances of the proposed LSTM model are shown in Figure 3 and Table 3. For testing, Sequence Vicon Room 2 03 of EuRoC MAV [9] was chosen, which is characterized as a difficult indoor environment. Figure 3 presents the model loss. It is important to note that both training and testing losses converge. The final value of training loss, 1.8, is slightly higher than that of testing loss, 1.5. Table 3 illustrates the mean error in translation and orientation, which is computed by comparing the model results with the database ground truth. It is obvious that LSTM network outperforms both IMU-Only and Kalman Filter. The accuracy in

translation improves by 76.2%. The mean error drops from 1.38 m to 0.43 m. With respect to orientation, the mean error reduces from 1.38 rad to 1.02 rad by 26.5%.

Figure 3. Proposed LSTM Model Loss

Table 3. Comparisons between IMU-Only, Kalman-Filter, and LSTM in Translation (m) and Orientation (rad).

Method	Translations Mean Error (m)	Orientation Mean Error (rad)
IMU-Only	0.88	1.18
Kalman Filter	1.38	1.38
LSTM	0.43	1.02

In order to explore the capability of this pre-trained LSTM model, we did another test with Sequence 9 from KITTI benchmark [10], that uses a different inertial sensor, OXTS RT 3003. As is shown in Table 4, our LSTM model fails to achieve similar accuracy and precision as the test with EuRoC dataset. Moreover, IMU-Only and Kalman Filter work better than this pre-trained model in both translation and orientation. This reveals the limitation of the proposed method. The performance of the LSTM is good while using the same or similar IMU, however, it can not handle another visual-inertial odometry system using an IMU with different specifications without retraining.

Table 4. Comparisons between IMU-Only, Kalman-Filter, and LSTM in Translation (m) and Orientation (rad) using KITTI dataset.

Method	Translations Mean Error (m)	Orientation Mean Error (rad)
IMU-Only	8.33	1.21
Kalman Filter	18.87	1.58
LSTM	33.17	3.03

5 CONCLUSION

This study proposes a novel learning-based method to cover brief camera failure or occlusion for monocular visual-inertial odometry system. In this research, we trained a 2-layer LSTM network with inertial measurements as inputs and visual-inertial Kalman-Filter pose estimations as the target. For evaluating, visual inputs were removed. The results of Kalman Filter, IMU-only, and pre-trained LSTMs were compared. The LSTM model outperforms both IMU-Only and Kalman Filter. The results demonstrate that the network improved the accuracy in translation by 76.2% and 26.5% in orientation. It was demonstrated that the proposed learning-based method maintains a high-quality and stable odometry performance during camera failure. However, the additional test with KITTI dataset [10] demonstrates the limitation of this approach. The pre-trained LSTM model only works with the IMU with same or similar specifications. Re-training is essential while using a different sensor.

REFERENCE

[1] Achtelik, M. W., Lynen, S., Weiss, S., Kneip, L., Chli, M., & Siegwart, R. “Visual-inertial SLAM for a small helicopter in large outdoor environments.” In 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems(pp. 2651-2652). IEEE. 2012.

[2] Kelly, J., & Sukhatme, G. S. “Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration.” The International Journal of Robotics Research, 30(1), 56-79. 2011.

[3] Leutenegger, Stefan, et al. "Keyframe-based visual–inertial odometry using nonlinear optimization." The International Journal of Robotics Research 34.3 : 314-334. 2015.

[4] S. Wang, R. Clark, H. Wen, and N. Trigoni. “Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks.” In Robotics and Automation (ICRA), 2017 IEEE International Conference on, pp. 2043–2050. IEEE, 2017.

[5] R. Li, S. Wang, Z. Long, and D. Gu, “Undeepvo: Monocular visual odometry through unsupervised deep learning,” arXiv preprint arXiv:1709.06841, 2017.

[6] J. R. Rambach, A. Tewari, A. Pagani, D. Stricker. “Learning to Fuse: A Deep Learning Approach to Visual-Inertial Camera Pose Estimation.” ISMAR 2016: 71-76.

[7] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni. “Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem.” In AAAI, pp. 3995–4001, 2017b.

[8] Barker, Todd, and Martin Åkerblad. "A Study on Long Short-Term Memory Networks Applied to Local Positioning." 2018.

[9] Burri, Michael, et al. "The EuRoC micro aerial vehicle datasets." The International Journal of Robotics Research 35.10: 1157-1163. 2016.

[10] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the KITTI vision benchmark suite,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.