Pose Estimation with a Lightweight Visual-Inertial Neural Network for Agile UAV Flights

Kevin Guzman-Duran*[1],Alejandro Gutierrez-Giles[1] , and Jose Martinez-Carranza*[1]
[1]Instituto Nacional de Astrofisica, Optica y Electronica, Puebla, Mexico

ABSTRACT

This paper presents a pose estimation system for UAVs designed for agile flights, utilizing the UZH-FPV Drone Racing dataset, which features high-speed, aggressive 6DoF trajectories for state estimation and drone racing. Our approach operates at 30 Hz, combining visual and inertial information through a sensor fusion technique incorporating temporal data into the input images. We use the lightweight convolutional neural network DeepPilot4Pose, which receives monocular images concatenated with inertial information. We process these data to calculate the UAV's position in real-time. Despite using a relatively small number of images for training compared to the trajectory length, our results show significant improvements in pose estimation accuracy and robustness in dynamic environments.

1 INTRODUCTION

In recent years, autonomous drones have grown exponentially in various fields, such as exploration, surveillance, supply delivery and cinematography. The ability of these Unmanned Aerial Vehicles (UAVs) to perform agile and precise manoeuvres in complex environments is crucial to their success in these applications. However, getting a UAV to navigate autonomously and efficiently in dynamic, high-speed scenarios presents a significant challenge due to the need for highly accurate and fast pose (position and orientation) estimation.

The central problem we are addressing is that UAVs, when flying at high speed and performing complex manoeuvres, demand pose estimation systems with high accuracy and low latency. However, traditional methods, such as GPS [1, 2] or additional hardware-based systems [3, 4, 5, 6, 7], are often not feasible due to their cost, weight, and power consumption, especially for small drones. Moreover, these methods could be more reliable in indoor or dense urban environments where GPS signals can be blocked or reflected [8, 2]. In such scenarios, there is an urgent need for a new, lightweight, efficient solution that can be easily integrated into UAVs and operate in real-time. Several methods on pose estimation for UAVs have been proposed in recent literature. For example, DeepPilot4Pose [9, 10], a compact convolutional neural network for visual pose estimation that runs onboard, flies in an indoor environment where no GPS or external sensors are available.

Our proposal presents a contribution to the field of pose estimation for UAVs. Our primary innovation lies in incorporating temporal and inertial data into the DeepPilot4Pose neural network, significantly improving pose estimation accuracy and robustness in dynamic environments. The DeepPilot4Pose network used in our approach has been optimized to fuse visual and inertial information at high speeds, providing real-time pose estimates at 30 Hz. This capability is crucial to enable the UAV to ”see” and ”feel” its environment through its cameras and sensors, using this information to calculate its position in real time.

Significantly, our system determines the absolute position of the UAV based on the vision system. The system can identify specific environmental features by leveraging visual information, ensuring precise localisation within a global reference frame.

We also evaluate our proposal using the UZH-FPV Drone Racing dataset [11], which provides images of aggressive flights and visual-inertial odometry data. Despite using a relatively small number of images for training compared to the trajectory length, our results show significant improvements in pose estimation accuracy and robustness in dynamic environments.

This paper is organised as follows: Section 2 provides a literature review on pose estimation; Section 3 describes the dataset used; Section 4 provides a description of the methodology for data collection, preprocessing and fusion; Section 5 describes the analysis of the results; and conclusions and future work are given in Section 6.

2 RELATED WORK

Visual-inertial odometry (VIO) and inertial measurement unit (IMU) based odometry have seen significant advances in recent years, driven by the need for accurate and robust pose estimation in various applications, such as robotics, autonomous vehicles, and augmented reality.

Visual odometry (VO) involves estimating the motion of a camera by analyzing captured image sequences. Early contributions in this field laid the foundation for modern VIO systems. [12] introduced a feature-based method for VO that uses feature matching and geometric triangulation to estimate

*Department of Computer Science at INAOE. Email addresses: {kevin.guzman, carranza}@inaoep.mx camera motion. Their work established a framework for subsequent research in visual odometry.

Inertial odometry (IO) uses IMU data to estimate pose. Traditional methods, such as the Strap Inertial Navigation System (SINS), rely on double integration of acceleration data. However, these methods suffer from significant drift due to noise and sensor biases, making them unsuitable for long-term navigation tasks. The challenges posed by drift and noise in IMU data have motivated the development of more sophisticated algorithms that can provide accurate pose estimates over extended periods.

Geometric methods for VIO combine visual and inertial data using mathematical models and estimation filters. Among them is VINS-MONO by Qin [13], which employs a sliding window optimisation approach to fuse camera and IMU data. This method operates at approximately 10 Hz and is known for its robustness and accuracy, although it requires significant computational resources. OKVIS [14] integrates an extended Kalman filter (EKF) with sliding window optimisation, achieving high accuracy at a processing frequency of around 30 Hz at the expense of increased computational complexity. These methods have been instrumental in advancing the state of the art in VIO.

The advent of deep learning has led to the development of end-to-end odometry solutions that can learn complex representations directly from raw sensor data. VINet [15] is a pioneering work that frames VIO as a sequence-to-sequence learning problem using long short-term memory (LSTM) networks. This approach takes advantage of temporal dependencies in the data to improve the pose estimation accuracy, operating at a frequency of approximately 10 Hz. Similarly, IONet [16] employs LSTM networks to predict displacements from IMU data, focusing on 2D trajectory estimation with an operational frequency of about 10 Hz. AbolDeepIO [17] combines convolutional neural networks (CNNs) with LSTM networks to estimate 3D translational and rotational shifts from IMU data. This approach demonstrates significant improvements over traditional methods, with a processing frequency of approximately 15 Hz, showing the potential of deep learning in inertial odometry. These deep learning-based methods represent a shift toward leveraging large data sets and complex models to achieve superior performance under challenging conditions.

Hybrid methods combine geometric principles with deep learning techniques to leverage the strengths of both approaches. DeepAVO [18] introduces a four-branch network with spatial-channel attention to improve visual pose estimation accuracy, operating at approximately 20 Hz. This method integrates geometric constraints with deep learning models to improve robustness and accuracy. TLIO [19] represents another significant advance, combining a neural network-based displacement estimator with an EKF to achieve accurate inertial odometry. TLIO operates at around 50 Hz, highlighting the efficiency of tight integration between learned models and traditional filtering techniques.

Visual-inertial and inertial odometry landscape has evolved considerably, from early geometric methods to modern deep learning and hybrid approaches. Pioneering work laid the foundation for robust pose estimation, while recent developments have leveraged advanced learning techniques to push the limits of accuracy and robustness. Integrating deep learning with traditional geometric methods offers a promising direction for future research, aiming to achieve reliable and accurate odometry in a wide range of challenging environments.

By providing a comprehensive review of these advances, this section highlights key contributions and ongoing efforts in odometry, emphasizing the importance of combining geometric insights with data-driven models to improve the performance and applicability of odometry systems.

3 DATASET FOR EXPERIMENTAL STUDIES

This section delves into the dataset that forms the basis of our method's evaluation, shedding light on the algorithm's robustness. For agile UAV flight pose estimation, we leveraged the UZH-FPV Drone Racing dataset [11]. This dataset is distinguished by its high velocity and is purpose-built to capture daring and highly agile manoeuvres, making it a crucial resource for dynamic and demanding flight scenarios.

Inspired by the Autonomous Drone Racing challenge [20], the UZH-FPV Drone Racing Dataset, a widely used resource in visual-inertial odometry (VIO) applications, was developed by ETH Zurich and is publicly accessible. This dataset comprises grayscale images captured from a stereo camera mounted on a Micro Aircraft (MAV). It also includes simultaneous data from the accelerometer and gyroscope of the onboard IMU sensor, providing a comprehensive set of data for various analyses. Additionally, the dataset provides information on the calibration and noise values of the camera and IMU sensor, further enhancing its usability.

Using a high-speed data set such as the UZH-FPV Drone Racing is critical for developing and evaluating pose estimation algorithms capable of supporting agile UAV flight. The UZH-FPV Drone Racing Dataset recordings were conducted in indoor and outdoor environments, spanning different environments and camera orientations. The environments contain various objects to provide texture. The dataset is divided into easy, medium, and hard difficulty categories based on illumination, MAV speed, and image blur.

In our study, we used the following datasets: Indoor forward facing: 3-01, 6-02, 9-01, 10-01 and Indoor 45-degree downward facing: 2-01, 4-01, 9-01, 12-01, 13-02, 14-03. The first digit represents the trajectory, while 01, 02 and 03 indicate easy, medium, and hard difficulty categories.

To generate the UZH-FPV Drone Racing dataset, the researchers employed a first-person view (FPV) racing quadcopter equipped with sensors and piloted aggressively by an expert pilot. The MAV hardware included sensors for visual and inertial measurements. IMU measurements were recorded at 500 Hz, while visual data were captured at 30 Hz. The IMU and cameras were synchronized so that half of the exposure coincided with the IMU measurements. A Leica Nova MS60 laser tracker, mounted on top of the MAV, was used to obtain reference data at a frequency of 500 Hz.

4 METHODOLOGY

In this study, visual and inertial information is evaluated in three different steps to perform pose estimation. These are the visual step, which uses the raw images in the data sets; the Inertial Image step, which combines the IMU data; and finally, the fusion step, which combines both information. Figure 1 shows the architecture of the proposed work.

4.1 Visual Data

Our study begins with the UZH-FPV Drone Racing dataset, a robust and reliable source of raw stereo images. We vertically concatenate the left mono and right mono images to form a single image. Then, we introduce our unique technique that incorporates the temporal dimension into the analysis through a three-dimensional concatenation of frames. This technique involves concatenating the current frames (n seconds) with frames captured at moments $n-5$ and $n-10$ seconds. The result is an RGB image where each channel represents a different time point, providing a rich representation of spatial and temporal information.

This concatenation allows us to capture relative motion between the camera and objects in the environment more accurately. The resulting images are used in the visual feature extraction steps, allowing the model to learn the dynamics of motion in the environment more effectively, as seen in Figure 1. In this way, the raw data from the camera is supplemented with motion information.

4.2 Inertial Data

The IMU sensors contain self-motion information, such as linear acceleration a and angular velocities ω. In some cases, the IMU also provides orientation information, e.g. roll-pitch-yaw angles (ϕ,θ,ϕ). These sensors, which are very sensitive to movement, operate at high frequency. In this study, the IMU frequencies for UZH-FPV Drone Racing are 500 Hz. In the data set, the IMU frequency is approximately 16 times the camera frequency. Therefore, there are 16 IMU data between any two frames.

In the inertial step, the goal is to use the raw IMU values between frames for pose estimation. In previous studies, time-dependent IMU raw data is generally directly fed to the LSTM layers, resulting in temporal IMU features. This study extracts features from inertial data differently from other studies to take advantage of the lightweight convolutional network architecture. Instead of using RNN-based LSTM as in traditional methods. We seek to use IMU data differently by relying on the robust estimation ability of CNN models.

$$\widehat{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

The accelerometer, gyroscope, and orientation values are first normalized to convert numerical IMU data into images due to their different lower and upper limits. Normalization adjusts these values to a standard range between 0 and 1. These normalized values are then multiplied by 255 to obtain a single-channel grayscale image with values ranging from 0 to 255. The resulting image comprises 144 pixels arranged in a 9×16 grid. Each pixel in this image represents the intensity of the corresponding accelerometer, gyroscope, or orientation value. In this way, the IMU data is transformed into a two-dimensional grayscale image, facilitating visualization and analysis and enabling better integration with the CNN models used in our study.

The same process as visual information is carried out to take advantage of temporality, i.e., incorporating the information corresponding to $n-5$ frames and $n-10$ into the remaining RGB channels.

4.3 Visual-Inertial Data

Finally, for data fusion we propose a process that involves the concatenation of visual information with inertial information. This three-dimensional concatenation of frames provides a representation that is rich in spatial and temporal information. The fusion of the frame and IMU information used in the fusion step and the temporal analysis representation are shown in Figure 2. Our preliminary results show a significant improvement in pose estimation when using both visual and inertial information, compared to using either one alone. The real power lies in the combination of both data sources, which allows us to capture the relative motion between the camera and objects in the environment more effectively, leading to accurate and robust pose estimation. The final image resulting from this fusion has a size of 224x224x3. These advanced data processing techniques ensure that our system can operate in real-time and quickly adapt to changes in the environment, significantly improving the accuracy and robustness of pose estimation in autonomous UAV navigation applications.

4.4 Pose Estimation

For 3D pose estimation, the DeepPilot4Pose [10] neural network is used. This network is capable of processing the fused images generated in the previous steps.

The DeepPilot4Pose network architecture comprises several convolutional layers that extract high-level features from the input images. These features are combined into fully connected layers to estimate the 3D position in the coordinates x, y and z. Additionally, a Savitzky-Golay filter [21] is applied to the output to smooth the pose estimates, thereby reducing noise and improving the estimation stability.

The DeepPilot4Pose network is a real-time powerhouse, processing the images at a frequency of 30 Hz, perfectly

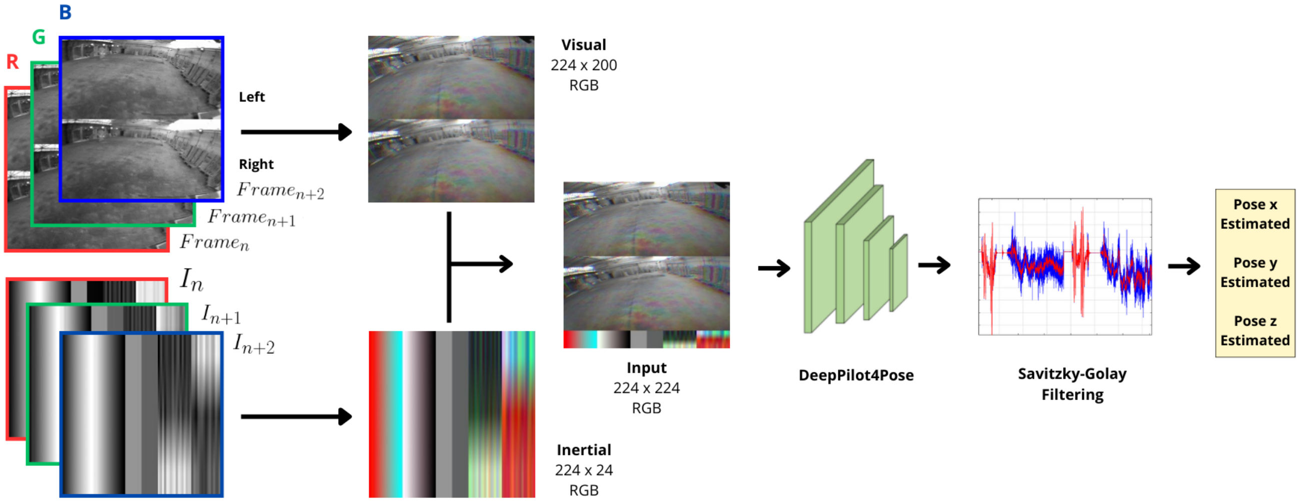


Figure 1: The architecture of the proposed approach.

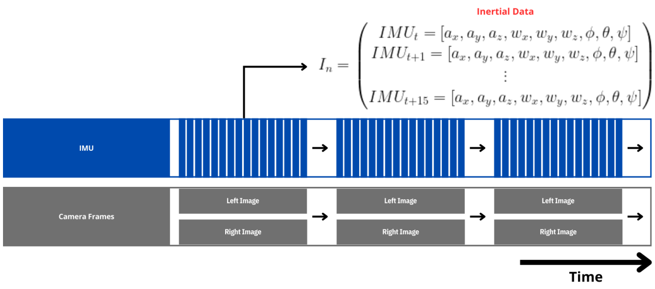

Inertial Data
$I_n = \begin{pmatrix} IMU_t = [a_x, a_y, a_z, w_x, w_y, w_z, \phi, \theta, \psi] \\ IMU_{t+1} = [a_x, a_y, a_z, w_x, w_y, w_z, \phi, \theta, \psi] \\ \vdots \\ IMU_{t+15} = [a_x, a_y, a_z, w_x, w_y, w_z, \phi, \theta, \psi] \end{pmatrix}$
IMU
Camera Frames
Left Image
Right Image
Left Image
Right Image
Left Image
Right Image
Time

Figure 2: Temporal relationship between the IMU and the frames.

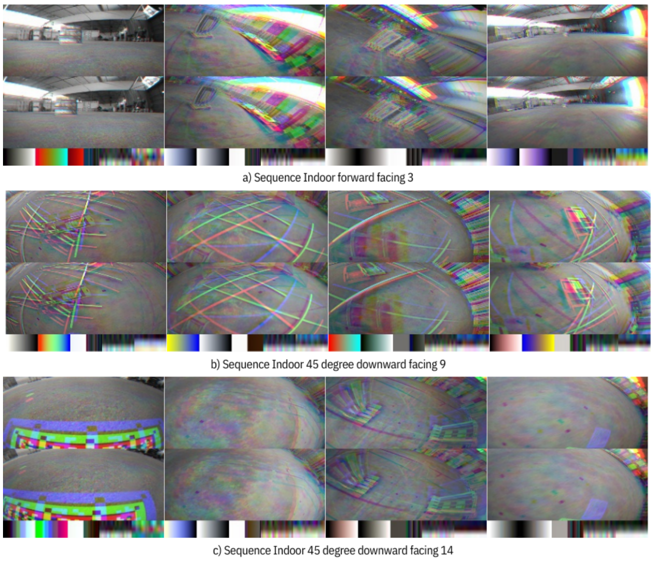
matching the camera image capture frequency. This capability ensures the system can provide accurate and timely pose estimates, making it an ideal tool for autonomous UAV navigation.

5 EXPERIMENTS AND RESULTS

This section analyzes the results of the proposed method applied to the UZH-FPV Drone Racing dataset. The training-test ratio for the data sets is estimated to be 80%-20%, taking at least one complete lap for training and another for testing. A desktop computer with NVIDIA GeForce RTX 3060 GPU was used during training and testing. The proposed algorithm calculates the average pose estimation time for the test data to be around 5 ms.

The root mean square error (RMSE) values are used at each of the three steps obtained from the test data Inside forward 03, 06, 09, 10 and Inside 45 degrees down 02, 04, 09,12, 13, 14. in this study, are given in Table 1.

In Table 1, the columns represent different aspects of the dataset and the results of the experiments. The columns include the maximum speed (V_{max}) of the UAV during the trajectory, the length of the trajectory in meters, the number of images used for training, the number of images used for testing, and the RMSE values for each method: Visual only, Inertial only, and the fusion of both.


a) Sequence Indoor forward facing 3

b) Sequence Indoor 45 degree downward facing 9

c) Sequence Indoor 45 degree downward facing 14

Figure 3: Illustration of a sequence of images representing the characteristics of a trajectory. The figure shows the complexity of the images, which can vary in texture from low to high, influencing the accuracy of pose estimation.

It is evident that prediction performance generally decreases as the scene and flight conditions in the UZH-FPV Drone Racing dataset become more challenging. Factors such as image layout and UAV speed, which can negatively affect the texture information of the environment, are expected to hinder prediction performance under such conditions. The

Sequences (Indoor)	$V_{\max}$ (m/s)	Length (m)	Training	Testing	Visual	Inertial	Fusion
Forward 3	9.50	22	1219	238	1.160	2.183	**0.864**
Forward 6	12.5	26	670	217	1.672	2.234	**1.623**
Forward 9	11.4	26	565	285	2.128	5.061	**0.967**
Forward 10	9.49	27	580	300	2.767	3.265	**2.136**
45 degree 2	6.97	27	1090	365	3.507	5.492	**3.387**
45 degree 4	6.55	24	840	200	1.750	5.514	**1.534**
45 degree 9	11.2	24	430	200	1.964	**1.535**	1.571
45 degree 12	4.33	28	590	597	**3.164**	5.384	3.294
45 degree 13	7.92	25	730	304	**1.214**	3.299	2.081
45 degree 14	9.54	22	815	170	1.414	**0.462**	0.867

Table 1: RMSE (m) values of the estimated UAV in the UZH-FPV Drone Racing dataset.

crux of the matter lies in using visual-inertial features together, ensuring minimal error due to fusion and providing more robust pose estimation in the face of these challenging conditions. The fusion results in Table 1 demonstrate that much more stable results are obtained than in the other two steps.

To better understand the performance of the UAV's predicted trajectory, the x, y, and z trajectories of the training data values and the position values for the test are visualised in Figure 4.

In this Figure 4, it is observed that when the test trajectory is similar to the one used in training and the greater the number of laps that represent the training, the performance in position estimation improves significantly. This occurs because the model has learned the specific trajectory and movement pattern characteristics during training, allowing it to make more accurate predictions in similar situations. However, due to the length and complexity of the trajectories, any slight difference can result in a decrease in performance since the model can only partially generalise to trajectories that deviate significantly from those observed during training.

Comparing the RMSE values for each experiment, it is observed that the best pose estimation is achieved using fused visual and inertial information. Visual information alone provides the second-best results, while inertial information alone is the least accurate. There are only a few cases where visual or inertial information obtains better results than information with the fusion of both sensors. This highlights the advantage of combining visual and inertial data for accurate posture estimation.

Additionally, it is essential to note that trajectories 45 degrees 9, 12, 13 and 14 show RMSE values where the highest values are not obtained by the fusion information, particularly in trajectories 45 degrees 9 and 14. These trajectories involve images with less texture and more repetition since the camera is tilted at 45 degrees and mainly captures the grey floor. This lack of distinguishing features makes it more challenging for the model to estimate the position accurately. As illustrated in the sequence images of Figure 3, these conditions can significantly affect the estimate's accuracy. In the case of sequence 14, the inertial information is reduced to the fusion threshold, highlighting that the maximum speed is 9.54 m/s, which leads us to think that due to the blurred information that captures the visual information, it hinders estimating the pose accurately.

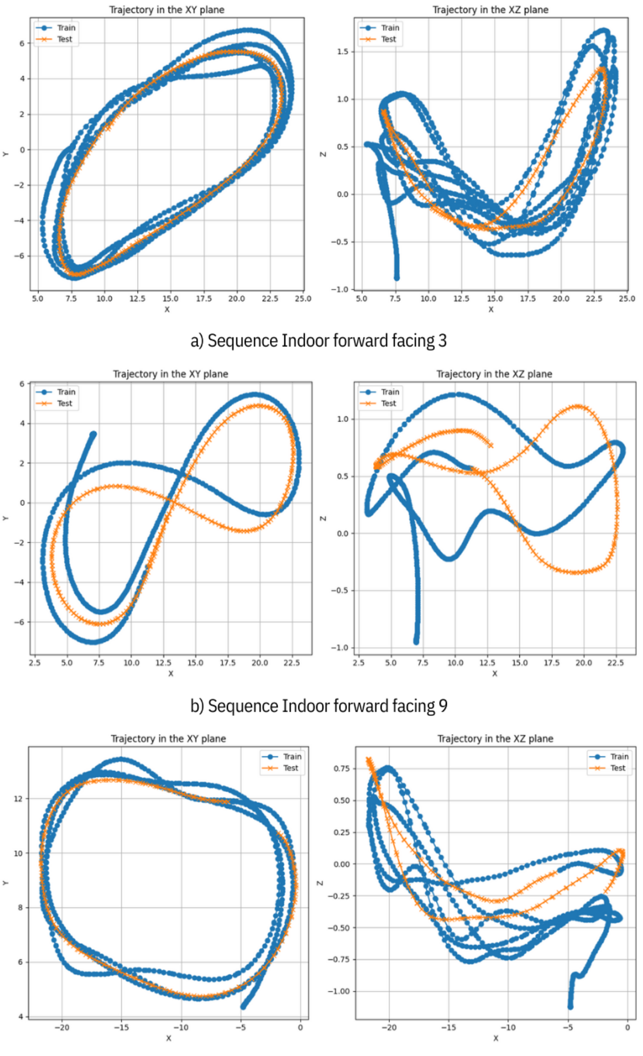
Figures 5, 6, and 7 illustrate the obtained results, showing graphs that compare the trajectories of the actual position and the estimated position in the x, y, and z axes. Additionally, the results are compared for estimations using only inertial and visual information and the fusion of both. These comparisons allow us to evaluate the accuracy of each approach and highlight the significant improvement achieved by fusing visual and inertial data.

Finally, it is crucial to highlight that the DeepPilot4Pose neural network operates at 30 Hz. This high operating frequency is essential for real-time autonomous navigation applications, as it allows for rapid and accurate pose updates. Additionally, this frequency could be increased if images are captured at a higher rate, further improving the system's accuracy and responsiveness in dynamic environments.

6 Conclusion

In this work, we have addressed the problem of pose estimation in agile UAV flights, a critical task for applications in robotics and automation, especially in dynamic and challenging environments. Accurate position estimation is essential for autonomous navigation and executing complex manoeuvres in real time.

Our method, is based on the fusion of visual visual and inertial data through out the use of a compact Convolutional Neural Network. We carried out experimental tests using the well-known UZH-FPV Drone Racing dataset. The network employs a three-dimensional frame concatenation technique. By incorporating the temporal dimension into the analysis, this technique creates an RGB image where each channel represents a different time point. This unique representation, rich in spatial and temporal information, allows us to capture motion dynamics more effectively, thereby improving the accu-



a) Sequence Indoor forward facing 3

b) Sequence Indoor forward facing 9

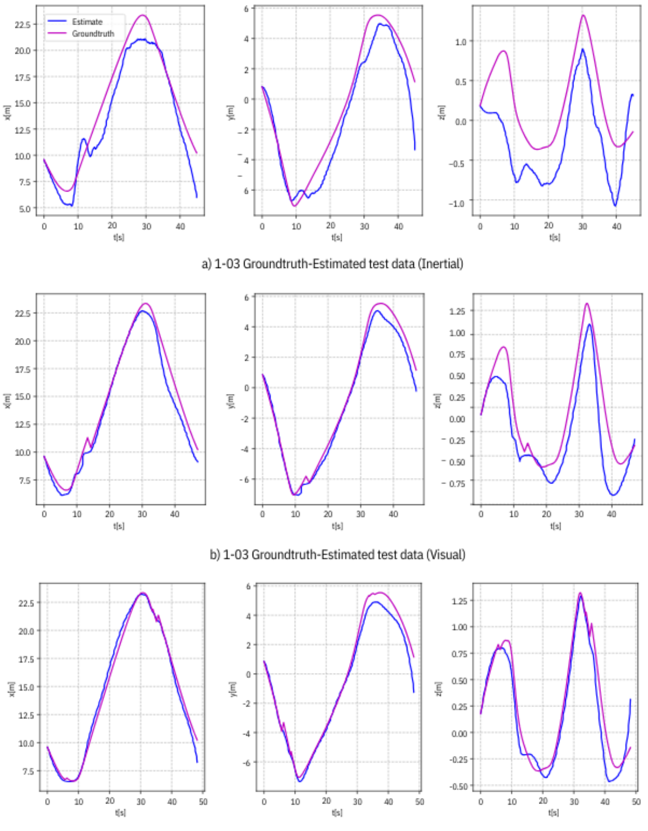
c) Sequence Indoor 45 degree downward facing 14

Figure 4: Comparison of two trajectories: a training trajectory and a test trajectory for pose estimation.

racy of pose estimation.

Our research has yielded tangible benefits, notably enhancing the robustness and accuracy of the pose estimation algorithm in the challenging scenarios captured in the UZH-FPV dataset. In terms of average error, our approach obtained an enhanced performance through out the fusion of the visual and intertial data in terms of accuracy. Another advantage is the lower memory consumption, while maintaining a processing time of 30 Hz, suitable for UAV applications with limited resources.

In future studies, we plan to evaluate our approach against other prominent methods mentioned in the literature review using the same UZH-FPV Drone Racing dataset. This will provide a broader perspective on our method's performance relative to other approaches, enhancing the robustness and reliability of our findings.



a) 1-03 Groundtruth-Estimated test data (Inertial)

b) 1-03 Groundtruth-Estimated test data (Visual)

c) 1-03 Groundtruth-Estimated test data (Fusion)

Figure 5: Groundtruth-estimated data from the Indoor forward facing 3 sequence.

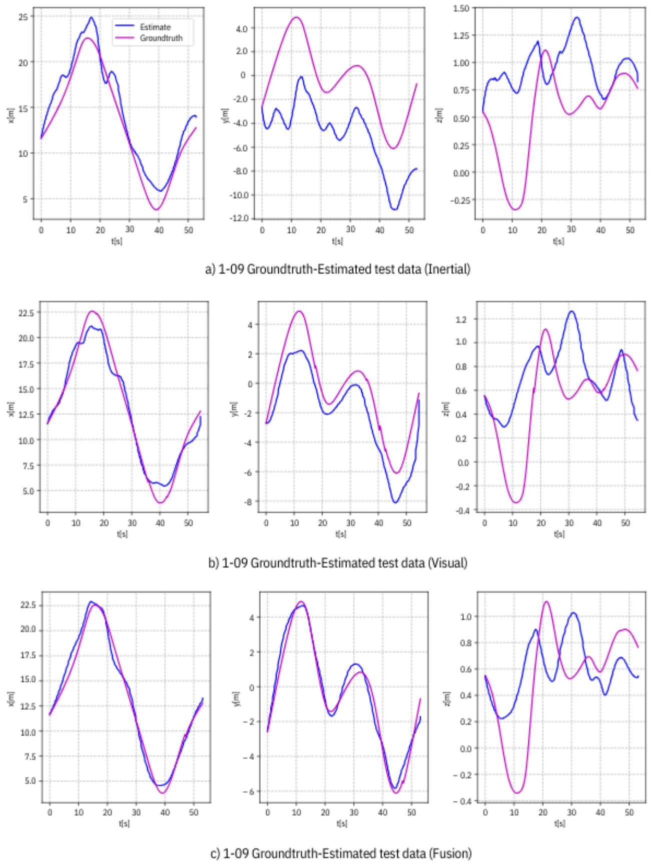
Additionally, we consider several promising directions to expand and improve our research: to investigate techniques to improve the algorithm's robustness in adverse environmental conditions, such as variable lighting, changing weather conditions, and environments with dynamic obstacles.

References

[1] Shah Zahid Khan, Mujahid Mohsin, and Waseem Iqbal. On gps spoofing of aerial platforms: a review of threats, challenges, methodologies, and future research directions. *PeerJ Computer Science*, 7:e507, 2021.

[2] Ganesan Balamurugan, J Valarmathi, and VPS Naidu. Survey on uav navigation in gps denied environments. In *2016 International conference on signal processing, communication, power and embedded system (SCOPES)*, pages 198–204. IEEE, 2016.

[3] John McConnell, Fanfei Chen, and Brendan Englot. Overhead image factors for underwater sonar-based slam. *IEEE Robotics and Automation Letters*, 7(2):4901–4908, 2022.

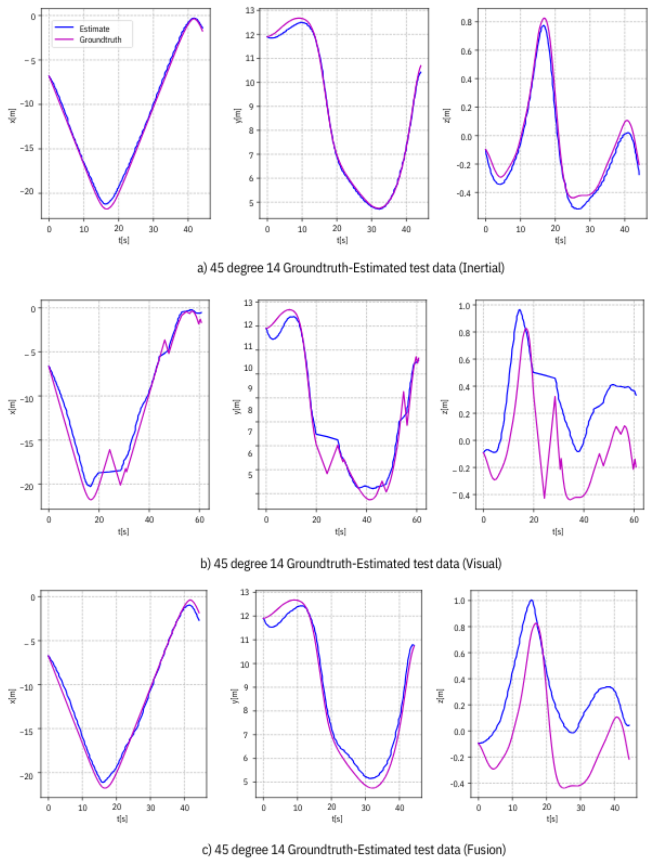


a) 1-09 Groundtruth-Estimated test data (Inertial)

b) 1-09 Groundtruth-Estimated test data (Visual)

c) 1-09 Groundtruth-Estimated test data (Fusion)

Figure 6: Groundtruth-estimated data from the Indoor forward facing 9 sequence.



a) 45 degree 14 Groundtruth-Estimated test data (Inertial)

b) 45 degree 14 Groundtruth-Estimated test data (Visual)

c) 45 degree 14 Groundtruth-Estimated test data (Fusion)

Figure 7: Groundtruth-estimated data from the Indoor 45 degree downward facing 14 sequence.

[4] Jie Qian, Kaiqi Chen, Qinying Chen, Yanhong Yang, Jianhua Zhang, and Shengyong Chen. Robust visual-lidar simultaneous localization and mapping system for uav. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

[5] Jung-Cheng Yang, Chun-Jung Lin, Bing-Yuan You, Yin-Long Yan, and Teng-Hu Cheng. Rtlio: Real-time lidar-inertial odometry and mapping for uavs. *Sensors*, 21(12):3955, 2021.

[6] Ahmed Elamin, Nader Abdelaziz, and Ahmed El-Rabbany. A gnss/ins/lidar integration scheme for uav-based navigation in gnss-challenging environments. *Sensors*, 22(24):9908, 2022.

[7] Matěj Petrlík, Tomáš Krajník, and Martin Saska. Lidar-based stabilization, navigation and localization for uavs operating in dark indoor environments. In *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 243–251. IEEE, 2021.

[8] Yueyan Zhi, Zhangjie Fu, Xingming Sun, and Jingnan Yu. Security and privacy issues of uav: A survey. *Mobile Networks and Applications*, 25(1):95–101, 2020.

[9] Leticia Oyuki Rojas-Perez and Jose Martinez-Carranza. Deeppilot: A cnn for autonomous drone racing. *Sensors*, 20(16):4524, 2020.

[10] Leticia Oyuki Rojas Pérez and Jose Martinez-Carranza. Deeppilot4pose: a fast pose localisation for mav indoor flight using the oak-d camera. *Journal of Real-Time Image Processing*, 20, 02 2023.

[11] Jeffrey Delmerico, Titus Cieslewski, Henri Rebecq, Matthias Faessler, and Davide Scaramuzza. Are we ready for autonomous drone racing? the UZH-FPV drone racing dataset. In *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2019.

[12] David Nistér, Oleg Naroditsky, and James Bergen. Visual odometry. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. Ieee, 2004.

[13] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[14] Stefan Leutenegger, Simon Lynen, Michael Bosse, Roland Siegwart, and Paul Furgale. Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research*, 34(3):314–334, 2015.

[15] Ronald Clark, Sen Wang, Hongkai Wen, Andrew Markham, and Niki Trigoni. Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

[16] Changhao Chen, Xiaoxuan Lu, Andrew Markham, and Niki Trigoni. Ionet: Learning to cure the curse of drift in inertial odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[17] Mahdi Abolfazli Esfahani, Han Wang, Keyu Wu, and Shenghai Yuan. Aboldeepio: A novel deep inertial odometry network for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):1941–1950, 2019.

[18] Ran Zhu, Mingkun Yang, Wang Liu, Rujun Song, Bo Yan, and Zhuoling Xiao. Deepavo: Efficient pose refining with feature distilling for deep visual odometry. *Neurocomputing*, 467:22–35, 2022.

[19] Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. Tlio: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020.

[20] Hyungpil Moon, Jose Martinez-Carranza, Titus Cieslewski, Matthias Faessler, Davide Falanga, Alessandro Simovic, Davide Scaramuzza, Shuo Li, Michael Ozo, Christophe De Wagter, et al. Challenges and implemented technologies used in autonomous drone racing. *Intelligent Service Robotics*, 12:137–148, 2019.

[21] Malek Karaim, Aboelmagd Noureldin, and Tashfeen B Karamat. Low-cost imu data denoising using savitzky-golay filters. In *2019 International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–5. IEEE, 2019.