Journal Title
XX(X):1–7
©The Author(s) 0000
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

# The UMA-VI Dataset: Visual-Inertial Odometry in Low-textured and Dynamic Illumination Environments

**David Zuñiga-Noël, Alberto Jaenal, Ruben Gomez-Ojeda, and Javier Gonzalez-Jimenez**

**Abstract**

This paper presents a visual-inertial dataset gathered in indoor and outdoor scenarios with a handheld custom sensor rig, for over 80 min in total. The dataset contains hardware-synchronized data from a commercial stereo camera (Bumblebee®2), a custom stereo rig and an inertial measurement unit. The most distinctive feature of this dataset is the strong presence of low-textured environments and scenes with dynamic illumination, which are recurrent corner cases of visual odometry and SLAM methods. The dataset comprises 32 sequences and is provided with ground truth poses at the beginning and the end of each of the sequences, thus allowing to measure the accumulated drift in each case. We provide a trial evaluation of five existing state-of-the-art visual and visual-inertial methods on a subset of the dataset. We also make available open source tools for evaluation purposes, as well as the intrinsic and extrinsic calibration parameters of all sensors in the rig. The dataset is available for download at http://mapir.uma.es/work/uma-visual-inertial-dataset

**Keywords**

Sensor Fusion, Localization, Mapping, Visual Tracking, Calibration and Identification

## 1 Introduction

*Visual Odometry* (VO) and *SLAM* techniques have greatly improved over the past years. As a consequence, state-of-the-art methods such as ORB-SLAM (Mur-Artal et al. 2015), DSO (Engel et al. 2018) or PL-SLAM (Gomez-Ojeda et al. 2019), for instance, achieve impressive performance results in real-time. However, there are still open problems that require further research before these techniques become *robust* enough for long-term application (Cadena et al. 2016).

This is the case of *low-textured* environments and scenes with *dynamic illumination* (Figure 1). The main issue with little textured scenes is the lack of enough salient features for reliable estimations, which can even lead to a complete system failure (Figure 1a). On the other hand, the changing light condition renders the visual tracking more challenging and thus affects the quality of the estimated trajectory (Figure 1b).

In this context, *Inertial Measurement Units* (IMUs) have proven to be of valuable help to gain robustness and precision over purely visual techniques (Leutenegger et al. 2015). *Visual-Inertial* (VI) fusion requires accurate synchronization between the sensors (Schubert et al. 2018), but such synchronization is hard to achieve in practice. Surely, this is one of the reasons that explains the lack of datasets for VI-based methods. Table 1 summarizes the most relevant datasets providing visual and inertial data. Please, note that only four of them use hardware synchronization for data acquisition. The PennCOSYVIO (Pfrommer et al. 2017) and OIVIO (Kasper et al. 2019) datasets are the most similar to ours since they also consider challenging textures and lighting. However, the former only provides 4 sequences (totaling 8.7 min) recorded on the same scene and under very similar lighting conditions, while the later focuses on onboard illumination for dark environments (mines, tunnels, etc). Therefore, we believe that the available datasets are not sufficient to test and validate VI odometry solutions under realistic settings.

In this paper we contribute a visual-inertial dataset in realistic environments with little texture and variable light conditions. It contains over 80 min of hardware synchronized IMU measurements and images from two stereo rigs, divided into 32 sequences. The trajectories for all sequences form large loops with the start, which allows to easily evaluate the accumulated drift of VI odometry methods, as proposed in Engel et al. (2016). Finally, we evaluated the performance of five start-of-the-art VI and VO solutions on a subset of the dataset: ORB_SLAM2 (Mur-Artal et al. 2017), PL-SLAM (Gomez-Ojeda et al. 2019), VINS-Mono (Qin et al. 2017), OKVIS (Leutenegger et al. 2015) and VINS-Fusion (Qin et al. 2019).

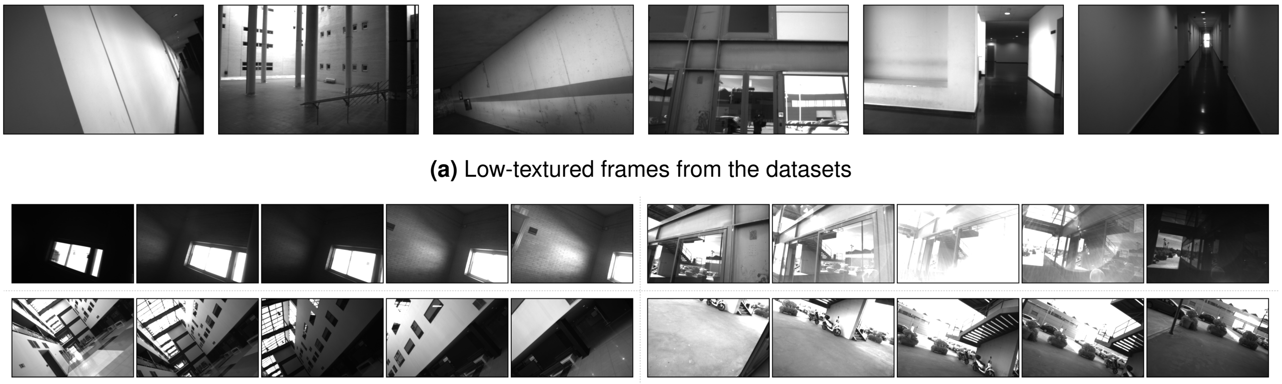
The rest of the paper is organised as follows. In Section 2, we describe the sensor setup used for data collection as well as the calibration process carried out to estimate both the intrinsic and extrinsic parameters of the unit. In Section 3, we provide an outline of the dataset itself, describe the loop-closure alignment procedure used to extract the ground truth information and explain the format in which the data is presented. We present a trial evaluation of state-of-the-art

Machine Perception and Intelligent Robotics (MAPIR) Group, University of Malaga, Spain.

**Corresponding author:**

David Zuñiga-Noël, Machine Perception and Intelligent Robotics (MAPIR) Group, System Engineering and Automation Department, University of Malaga, Campus de Teatinos, 29071 Malaga, Spain.

Email: dzuniga@uma.es



**(a)** Low-textured frames from the datasets

**(b)** Frame sequences with fast illumination changes

**Figure 1.** Our dataset is built from a number of visually challenging sequences, showing low-textured environments and scenes with dynamic illumination. The goal is to provide a benchmark for the evaluation of visual-inertial odometry algorithms in these real-world situations.

visual and visual-inertial solutions in Section 4. Finally, the contributions of this paper are summarized in Section 5.

## 2 Sensor Unit Description

### 2.1 Sensor Setup

We designed a custom VI sensor unit for data collection purposes. Our VI sensor consists of two stereo rigs and a three-axis inertial unit, as depicted in Figure 2. The individual characteristics of each sensor are described next and briefly summarized in Table 2.

- The Bumblebee®2 (BB2-08S2C-25) stereo camera provides stereo, Bayer encoded color images with 1024×768 px resolution with auto control of the exposure time and sensor gain. The cameras have a 1/3" Sony ICX204 CCD sensor with global shutter and a 2.5 mm lens with 96° Horizontal Field of View (HFoV) each. The stereo camera has a 12 cm baseline, and synchronous stereo images were recorded at 12.5 Hz.
- The custom stereo rig was built from two IDS uEye UI-1226LE-M cameras, each providing 752×480 px monochrome images with hardware auto-exposure. The cameras have a 1/3" Mobisense MT9V032STM CMOS sensor with global shutter and a 3.5 mm lens with 60° HFoV. The custom stereo rig has a 25.5 cm baseline, and synchronous stereo images were recorded at 25 Hz.
- The IMU is a XSens MTi-28A53G35 3D motion sensor, providing angular rates and specific force measurements in three perpendicular axes. The device is intrinsically calibrated from factory to output corrected measurements, which we logged at 250 Hz.

The data from the sensors were recorded with a consumer-grade Acer Travelmate P259 series laptop into a high speed Samsung 970 EVO solid-sate drive. The Bumblebee®2 cameras were connected to the laptop through a shared FireWire 400 bus, while the uEye cameras and the XSens IMU were connected through USB 3.0 and 2.0 connections, respectively.

For synchronization purposes, first we configured all sensors to run in external trigger mode for data acquisition. Then, we programmed an ATmega328P microcontroller to generate the trigger signals at the specified rates, for each sensor individually. Even though each sensor has its own trigger signal, we guarantee accurate time synchronization by using the same clock to generate all triggers. The remaining small, constant time delays specific to each sensor are further calibrated for additional accuracy.

### 2.2 Calibration

The dataset contains raw data, intrinsic and extrinsic calibration parameters as well as temporal offsets. In the following we describe how these parameters were estimated. We also provide the corresponding calibration sequence, allowing custom calibration methods to be used with our dataset.

*2.2.1 Stereo Camera Calibration* The intrinsic parameters (the projection parameters of each camera as well as the relative spatial transformation between the two cameras of each stereo setup) were calibreated for each stereo rig independently. For that purpose, we recorded the calibration pattern (an AprilTag grid Olson 2011) while slowly moving the VI sensor unit in front of it. The final calibration parameters were estimated using the calibration toolbox *Kalibr*[*], presented in Furgale et al. (2013).

*2.2.2 IMU Noise Calibration* For VI sensor fusion, the noise of the IMU measurements has to be characterized. Typically, it is assumed that IMU measurements are perturbed by white noise and a slowly varying bias (random walk). We estimated these parameters from the Allan deviation function of our VI sensor resting for a long period (more than 100 h), as described in Schubert et al. (2018).

*2.2.3 Camera-IMU Extrinsics and Time Delays* We calibrated the extrinsics as well as the time-sinchronization offsets for each camera with respect to the IMU. For that purpose, we again recorded the calibration pattern (AprilTag grid Olson 2011) while moving the VI sensor unit in front

[*] https://github.com/ethz-asl/kalibr

**Table 1.** Comparison of existing datasets for Visual-Inertial Odometry/SLAM.

| Dataset | Environment | Motion Type | Sensor Configuration | Ground-truth |
|---|---|---|---|---|
| KITTI (Geiger et al. 2012) | Outdoors | Car | Stereo/IMU/Laser [1] | INS/GNSS |
| Malaga Urban (Blanco-Claraco et al. 2014) | Outdoors | Car | Stereo/IMU [1] | GPS |
| UMich NCLT (Carlevaris-Bianco et al. 2016) | Indoors/Outdoors | Segway | Omni/IMU/Laser [1] | GPS/IMU/Laser |
| EuRoC MAV (Burri et al. 2016) | Indoors | MAV | Stereo/IMU [2] | MoCap |
| Zurich Urban (Majdik et al. 2017) | Outdoors | MAV | Monocular/IMU [1] | Visual pose |
| PennCOSYVIO (Pfrommer et al. 2017) | Indoors/Outdoors | Handheld | Stereo/IMU [1,2] | Visual pose (markers) |
| TUM VI (Schubert et al. 2018) | Indoors/Outdoors | Handheld | Stereo/IMU [2] | MoCap (partial) |
| ADVIO (Cortés et al. 2018) | Indoors/Outdoors | Handheld | Stereo/Depth/IMU[1] | IMU |
| KAIST Urban (Jeong et al. 2019) | Outdoors | Car | Stereo/IMU [1] | GPS/FOG/Encoder/LiDAR |
| OIVIO (Kasper et al. 2019) | Indoors/Outdoors | Handheld | Stereo/IMU[1,2] | Visual Pose (partial) |
| Ours | Indoors/Outdoors | Handheld | Stereo/IMU [2] | Visual pose (partial) |

[1] software synchronized [2] hardware synchronized

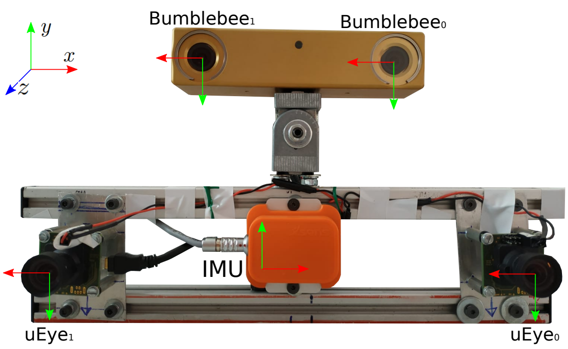


**Figure 2.** The visual-inertial sensor unit used for dataset collection. It contains a stereo camera (Bumblebee®2), a custom stereo rig with two cameras (uEye) and an XSens Inertial Measurement Unit (IMU). All sensors are hardware-synchronized by means of a microcontroller.

**Table 2.** Summary of the main features of each sensor.

| Sensor | Model | Rate | Features |
|---|---|---|---|
| Stereo Camera | PointGrey Bumblebee®2 | 12.5 Hz | Color, 1024×768, Auto-exposure, auto-gain |
| Stereo Rig | 2 x IDS uEye UI-1226LE-M | 25 Hz | Grayscale, 752×480, Auto-exposure |
| IMU | XSens MTi-28A53G35 | 250 Hz | 3D Accelerometer, 3D Gyroscope |

of it, trying to excite the three axes of the IMU with rotation and translation movements. The calibration was performed with good scene illumination and slow motions in order to minimize motion blur. We used again *Kalibr* (Furgale et al. 2013) to estimate the extrinsic parameters and time delays of the cameras.

*2.2.4 Photometric Calibration* In order to enable direct VI methods to be tested with our dataset, we also calibrated the sensor's response function and lens vignetting map for the uEye monochrome cameras. To calibrate the response function, we recorded a static scene with different exposure times (ranging from 0.07 to 20 ms with the smallest steps allowed by the sensor). For the vignette calibration, we recorded a known marker (ArUco tag Garrido-Jurado et al. 2014) on a white planar surface while moving the VI sensor in order to observe it from different angles. The photometric calibration was carried out using a modified version[†] of the code provided by the TUM MonoVO dataset (Engel et al. 2016).

## 3 Visual Inertial Dataset

### 3.1 Dataset description

We recorded 32 sequences for the evaluation of VI motion estimation methods, totalling ∼80 min of data. The dataset covers challenging conditions (mainly illumination changes and low textured environments) at different degrees and a wide range of scenarios (including corridors, parking, classrooms, halls, etc) from two different buildings at the University of Malaga. In general, we deliver at least two different sequences within the same scenario, with different illumination conditions or following different trajectories. All sequences were recorded with our VI rig handheld, including a few during which the person holding the rig was mounted on a moving car. An overview of the sequences included in the dataset is presented in Table 3.

We grouped the evaluation sequences into different categories, depending on the type of challenges that they address:

- **Low-texture:** indoor sequences showing environments lacking distinctive features or with repetitive textures
- **Indoor:** sequences containing dark and light areas in indoor scenarios
- **Outdoor:** sequences containing natural illumination changes present in outdoor environments
- **Indoor-Outdoor dynamic illumination:** sequences containing the typical illumination changes of indoors/outdoors transitions
- **Indoor with illumination changes:** indoor sequences with forced artificial lightning changes (spotlights, lights on and off)
- **Sun overexposure:** challenging sequences with a blinking effect in the uEye cameras caused by direct exposure to the sunlight, which saturates the imaging sensors and causes rapid fluctuations in the exposure time

[†] https://github.com/AlbertoJaenal/mono_dataset_code

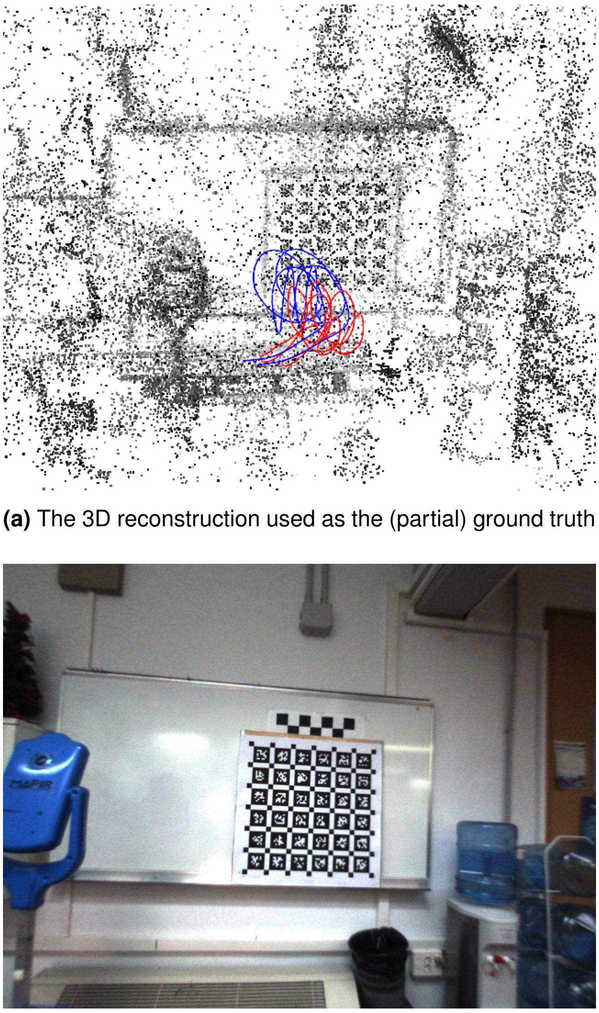
**Table 3.** Overview of the sequences included in the dataset.

| Sequence type | Number of sequences | Time |
|---|---|---|
| Calibration sequences | | |
| Cam calibration | 3 for uEye, 3 for Bumblebee | 29.3 min |
| IMU/Cam calibration | 3 for uEye, 3 for Bumblebee | 8.6 min |
| Photometric calibration (uEye) | 2 for response, 2 for vignette | 14.4 min |
| IMU static | 1 sequence | 141.3 h |
| Evaluation sequences | | |
| Low-texture | 5 sequences | 7.9 min |
| Indoor | 5 sequences | 8.4 min |
| Outdoor | 6 sequences | 20.4 min |
| Indoor-Outdoor dynamic illumination | 5 sequences | 15 min |
| Indoor with illumination changes | 5 sequences | 10 min |
| Sun overexposure | 6 sequences | 18.3 min |

## 3.2 Loop-Closure Alignment

Due to the specific characteristics of our dataset (long trajectories with indoor-outdoor changes within a single sequence), it is not feasible to track the position of our VI sensor unit with an independent external reference system (such as a Vicon mo-cap) to obtain the ground truth data. For this reason, we decided to compute partial[‡] ground truth data from visual input, following the approach proposed in Engel et al. (2016). The approach consists of designing looped trajectories with the same start and end point. The sequences begin and end observing the same, well-textured, easy-to-track scene with smooth motions for approximately 10 s. This way, the ground truth poses for the start and end segments of each trajectory can be computed through 3D reconstruction techniques and then used to evaluate the accumulated drift of an odometry solution, as in Schubert et al. (2018); Kasper et al. (2019) (see Figure 3). As noted in Engel et al. (2016), the loop-closure module of the algorithms to be evaluated should be disabled.

The 3D reconstruction was performed with the Structure-from-Motion (SfM) pipeline *COLMAP* (Schonberger and Frahm 2016). For simplicity, we used only one of the two stereo rigs to compute the reference poses of the start and end segments. We choose the Bumblebee®2 stereo camera over the custom stereo rig mainly due to its larger FoV. Since *COLMAP* does not explicitly support stereo cameras, the reconstructions are computed from two independent monocular cameras, and then the metric scale is recovered by imposing the calibrated baseline in a final optimization step.



**(a)** The 3D reconstruction used as the (partial) ground truth

**(b)** The well-textured, easy-to-track scene

**Figure 3.** Example of the loop-closure alignment between the first (start segment, in red) and the last 10 s (end segment, in blue) of a sequence performed with *COLMAP* (Figure 3a). To improve the accuracy of the reconstruction, we included the calibration pattern in the loop-closure (Figure 3b).

## 3.3 Evaluation Metrics

We measure the accumulated tracking inaccuracies of a trajectory using the *alignment* error proposed in Engel et al. (2016). This metric reflects the overall performance of the method in a single value, and it is equally affected by the accumulated drifts in scale, rotation and translation over the whole trajectory. Additionally, it allows to compare in a direct way methods with different observability modes (like monocular, stereo or visual-inertial).

To determine the *alignment* error, first we need to compute the transformations that best align the estimated trajectory to the reference start and end segments independently (see Figure 4):

$$T_s^{gt} \triangleq \underset{T \in \mathrm{Sim}(3)}{\arg\min} \sum_{i \in S} \left(\hat{p}_i - T \oplus p_i\right)^2 \tag{1}$$

$$T_e^{gt} \triangleq \underset{T \in \mathrm{Sim}(3)}{\arg\min} \sum_{i \in E} \left(\hat{p}_i - T \oplus p_i\right)^2 \tag{2}$$

where $S \subset [1, \ldots, n]$ and $E \subset [1, \ldots, n]$ represent the subsets of frame indices for the start and end segments of a sequence with $n$ stereo frames in total, and $\hat{p}_i, p_i \in \mathbb{R}^3$ the ground truth and estimated positions for the $i$-th frame, respectively. The $\oplus$ operator refers to the pose-point composition operator (Blanco-Claraco 2010).

Finally, the *alignment* error is defined to be the translational root-mean-square-error between the estimated

[‡] Since our dataset aims to capture different visual challenges, we are unable to provide ground truth data from visual input for the whole trajectory.

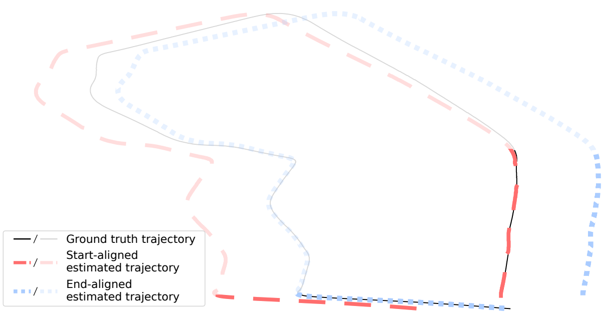


**Figure 4.** Example of the trajectory alignment approach used for evaluation. The reference, ground truth trajectory $\hat{p}_i$ is represented in black, the estimated trajectory aligned to the start segment $T_s^{gt} \oplus p_i$ in blue (dotted) and the estimated trajectory aligned to the end segment$T_e^{gt} \oplus p_i$ in red (dashed). Only the poses whose index $i \in S \cup E$ are highlighted.

trajectory itself when aligned to the start and end segments:

$$e_{align} \triangleq \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left\|T_s^{gt} p_i - T_e^{gt} p_i\right\|_2^2} \tag{3}$$

Additionally, the accumulated drift $T_{drift}$ can be computed from the start and end alignment transformations:

$$T_{drift} \triangleq T_e^{gt} \oplus (\ominus T_s^{gt}) \tag{4}$$

where $\ominus$ represents the inverse pose operator (Blanco-Claraco 2010). Then the rotation, translation and scale drift can be easily extracted from $T_{drift}$ as:

$$e_r = \arccos\left(\frac{\mathrm{trace}(R_{drift}) - 1}{2}\right) \tag{5}$$

$$e_t = \|t_{drift}\| \tag{6}$$

$$e_s = s_{drift} \tag{7}$$

where $(R_{drift}, t_{drift}, s_{drift}) = T_{drift} \in \mathrm{Sim}(3)$. Note that these drift values can also be used for a detailed performance evaluation.

## 3.4 Data Format

Each sequence is packed into a single zip file containing the data gathered with our VI unit. The layout of the zip files is depicted in Figure 5, and it is very similar to the layout used in the EuRoC dataset (Burri et al. 2016). The data of each sensor is divided into sub-folders. The specific contents for each type of sensor are described next.

- *Camera*: the observations from cameras are represented as images, stored in the data sub-folder in PNG lossless format. The Bumblebee®2 cameras are referred to as cam0 and cam1 (left and right, respectively), while the uEye cameras as cam2 and cam3 (left and right, respectively). In the former, they are stored as RGB color images[§], while as grayscale images in the latter. In both cases, the images provided are unrectified (as they were captured). Additionally, the data.csv file contains plain text image-timestamp[¶] associations as well as their respective acquisition exposure times (in ns), as Comma Separated Values (CSV).

```
<sequence_id>
├─cam0
│  ├─data
│  │  ├─1549036653683940058.png
│  │  ├─1549036653763940058.png
│  │  └─...
│  └─data.csv
├─cam1
│  ├─data
│  │  └─...
│  └─data.csv
├─cam2
│  └─...
├─cam3
│  └─...
├─imu0
│  ├─data.csv
│  └─bias_priors.csv
└─imu0_trajectory.csv
```

**Figure 5.** Example path layout for a single sequence.

- *IMU*: the inertial measurements and their acquisition timestamps are stored in the file data.csv as plain text (CSV). Each row contains: timestamp (in ns), gyroscope (in rad/s) and accelerometer (in m/s$^2$). The file bias_priors.csv contains gyroscope and accelerometer bias priors in a similar format as the IMU measurements, estimated for each trajectory with Leutenegger et al. (2015).
- *Ground truth*: the partial ground truth trajectory (for the start and end segments) is provided in the imu0_trajectory.csv as plain text (CSV). This is the reference trajectory, as described by the IMU and computed from the Bumblebee®2 camera as described in Section 3.2 (applying the calibrated extrinsic parameters and time delay). Note that the ground truth trajectory is provided at the camera rate (12.5 Hz). Each row contains: timestamp (in ns), translation (in m) and rotation (as a Hamiltonian unit quaternion).

## 3.5 Calibration Data

The calibration parameters of our VI unit (estimated as described in Section 2.2) are provided in a separate zip package. The parameters are stored in plain text YAML format. The package contains the following files:

- bumblebee_camchain.yaml: intrinsic and extrinsic parameters for the Bumblebee®2 stereo camera. The extrinsic parameters with respect to the IMU are also included.

[§] The raw color images from the Bumblebee®2 stereo camera are captured with a Bayer pattern. For practical reasons, the images provided in the dataset have already been converted to the three-channel RGB representation.

[¶] These are raw acquisition timestamps, without delay or exposure compensation.

- ueye_camchain.yaml: intrinsics and extrinsics for the uEye stereo rig, and the extrinsics with respect to the IMU.
- xsens_imu.yaml: noise density and random walk intrinsic parameters of the IMU.

Additionally, the per-pixel attenuation factors of vignetting calibration are provided for the uEye cameras in PNG format.

## 4 Evaluation of Existing Methods

We present a trial evaluation of state-of-the-art methods, in which we evaluated the accumulated drift in a representative subset of the dataset, covering all different categories. We selected five methods with different observability modes: ORB_SLAM2 (Mur-Artal et al. 2017) and PL-SLAM (Gomez-Ojeda et al. 2019) as pure visual stereo systems; VINS-Mono (Qin et al. 2017) as a monocular visual-inertial solution and OKVIS (Leutenegger et al. 2015) and VINS-Fusion (Qin et al. 2019) as stereo visual-inertial approaches.

The results are summarized in Tables 4 and 5, showing the mean over 5 executions of the alignment errors as well as the accumulated translational, rotational and scale drift for the uEye and Bumblebee®2 rigs, respectively. For the monocular case, we used only the left camera of the stereo rigs. The evaluation shows that ORB_SLAM2 is hardly able to complete the sequences and PL-SLAM presents more robustness to visual tracking failure since it relies on both points and line segments features. This suggests that visual systems may require extra information to cope with such challenges. On the other hand, the results indicate general performance improvements when using an IMU in addition to the visual system. The monocular VI case (VINS-Mono) exhibits large drifts in some cases, while the stereo VI methods are robust enough to complete these sequences.

## 5 Conclusions

We have presented a new indoor-outdoor Visual-Inertial dataset that aims to provide means for the evaluation of odometry and SLAM methods in real-world, visually challenging situations. The dataset focuses on changing light conditions and low-textured scenes in a wide variety of environments. Every sequence contains a large loop-closure at the end that allows to measure the accumulated drift. The sequences were recorded with a handheld custom VI sensor unit. Our sensor unit consists of four cameras, divided in two stereo rigs and an IMU, all hardware-synchronized by a microcontroller. We evaluated the performance of state-of-the-art methods in a subset of the dataset, showing the challenge that it represents for pure visual systems and even for VI approaches. Open source evaluation tools, as well as the calibration parameters of the whole VI unit are available for download at http://mapir.uma.es/work/uma-visual-inertial-dataset

### Acknowledgements

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal used for this research.

### Funding

This work has been funded by the Government of Spain under project FPU17/04512 and project WISER DPI2017-84827-R.

### References

Blanco-Claraco JL (2010) A tutorial on SE(3) transformation parameterizations and on-manifold optimization. Technical report, University of Malaga.

Blanco-Claraco JL, Moreno-Dueñas FÁ and González-Jiménez J (2014) The Málaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario. *The International Journal of Robotics Research* 33(2): 207–214.

Burri M, Nikolic J, Gohl P, Schneider T, Rehder J, Omari S, Achtelik MW and Siegwart R (2016) The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research* 35(10): 1157–1163.

Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, Reid I and Leonard JJ (2016) Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics* 32(6): 1309–1332.

Carlevaris-Bianco N, Ushani AK and Eustice RM (2016) University of Michigan North Campus long-term vision and lidar dataset. *The International Journal of Robotics Research* 35(9): 1023–1035.

Cortés S, Solin A, Rahtu E and Kannala J (2018) ADVIO: An authentic dataset for visual-inertial odometry In: *Proceedings of the European Conference on Computer Vision (ECCV)* pp. 425–440.

Engel J, Koltun V and Cremers D (2018) Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(3): 611–625.

Engel J, Usenko V and Cremers D (2016) A photometrically calibrated benchmark for monocular visual odometry. *arXiv preprint arXiv:1607.02555*.

Furgale P, Rehder J and Siegwart R (2013) Unified temporal and spatial calibration for multi-sensor systems. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 1280–1286.

Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ and Marín-Jiménez MJ (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47(6): 2280–2292.

Geiger A, Lenz P and Urtasun R (2012) Are we ready for autonomous driving? The KITTI vision benchmark suite. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3354–3361.

Gomez-Ojeda R, Moreno FA, Zuñiga-Noël D, Scaramuzza D and Gonzalez-Jimenez J (2019) PL-SLAM: a stereo SLAM system through the combination of points and line segments. *IEEE Transactions on Robotics* 35(3): 734–746.

Jeong J, Cho Y, Shin YS, Roh H and Kim A (2019) Complex urban dataset with multi-level sensors from highly diverse urban environments. *The International Journal of Robotics Research* 38(6): 642–657.

**Table 4.** Translational, rotational and scale drift results for uEye cameras. L refers to sequences where the tracking got lost and D refer to excessively drifted sequences. The most accurate results in each group are **highlighted**, i.e. closest to zero in general and closest the unity for $e_s$.

| Dataset category | ORB_SLAM2 | | | | PL-SLAM | | | | VINS-Mono | | | | VINS-Fusion | | | | OKVIS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ |
| Low-texture | L | | | | 3.72 | 5.38 | 60.88 | 0.52 | 11.55 | 11.21 | 35.7 | 0.92 | **1.50** | **0.50** | **7.30** | **1.02** | 2.31 | 0.61 | 21.75 | 0.98 |
| Indoor | L | | | | L | | | | D | | | | **0.77** | **0.29** | **5.53** | **0.99** | 7.94 | 0.78 | 13.74 | 0.97 |
| Outdoor | L | | | | D | | | | 2.55 | 1.35 | 6.57 | 0.99 | **0.33** | **0.29** | **0.57** | **1.00** | 2.48 | 0.42 | 10.14 | 1.00 |
| Indoor-Outdoor | L | | | | L | | | | D | | | | D | | | | **1.07** | **0.79** | **2.60** | **1.01** |
| Illumination changes | L | | | | L | | | | 1.24 | 0.38 | 5.96 | 0.84 | **0.32** | **0.25** | **3.07** | **1.01** | 0.75 | 0.72 | 1.88 | 0.96 |
| Sun overexposure | L | | | | L | | | | 22.64 | 7.57 | 60.22 | 0.32 | 2.97 | **1.05** | 9.69 | **1.02** | **2.51** | 2.38 | **3.22** | 1.05 |

**Table 5.** Translational, rotational and scale drift results for Bumblebee®2 cameras. Sequences marked with * lost tracking in more than 80% of executions. The most accurate results in each group are **highlighted**, i.e. closest to zero in general and closest the unity for $e_s$

| Dataset category | ORB_SLAM2 | | | | PL-SLAM | | | | VINS-Mono | | | | VINS-Fusion | | | | OKVIS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ | $e_{align}$ (m) | $e_t$ (m) | $e_r$ (°) | $e_s$ |
| Low-texture | 8.88* | 6.43* | 31.15* | 1.01* | 7.11 | 3.97 | 43.7 | 0.62 | L | | | | **0.74** | 0.71 | **4.92** | **1.01** | 2.49 | **0.42** | 25.58 | 0.97 |
| Indoor | L | | | | 8.56 | 13.58 | 63.3 | 0.45 | D | | | | **0.47** | **0.18** | **2.99** | **0.97** | 10.08 | 1.62 | 33.16 | 0.95 |
| Outdoor | 15.80* | 20.24* | 54.42* | 0.82* | L | | | | 2.21 | **0.92** | 8.16 | 1.01 | **1.10** | 1.35 | **5.55** | **0.99** | 3.44 | 1.58 | 17.77 | 0.95 |
| Indoor-Outdoor | L | | | | 19.91 | 23.55 | 67.4 | 0.74 | 3.32 | 1.05 | 7.39 | 0.94 | 0.97 | 1.38 | 2.17 | 1.01 | **0.65** | **0.57** | **1.95** | **0.99** |
| Illumination changes | L | | | | L | | | | 1.53 | 0.96 | 6.46 | 0.86 | **0.51** | **0.18** | 5.14 | **1.00** | 0.89 | 0.92 | **0.99** | 0.94 |
| Sun overexposure | L | | | | L | | | | 9.88 | 6.53 | 21.57 | 0.82 | **1.38** | **1.18** | **4.20** | **1.00** | 2.91 | 1.43 | 8.59 | 1.10 |

Kasper, M and McGuire, S and Heckman, C (2019) A Benchmark for Visual-Inertial Odometry Systems Employing Onboard Illumination.In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 5256–5263.

Leutenegger S, Lynen S, Bosse M, Siegwart R and Furgale P (2015) Keyframe-based visual–inertial odometry using nonlinear optimization. *The International Journal of Robotics Research* 34(3): 314–334.

Majdik AL, Till C and Scaramuzza D (2017) The Zurich urban micro aerial vehicle dataset. *The International Journal of Robotics Research* 36(3): 269–273.

Mur-Artal R, Montiel JMM and Tardos JD (2015) ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics* 31(5): 1147–1163.

Mur-Artal R and Tardos JD (2017) ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics* 33(5): 1255–1262.

Olson E (2011) AprilTag: A robust and flexible visual fiducial system. In: *2011 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 3400–3407.

Pfrommer B, Sanket N, Daniilidis K and Cleveland J (2017) PennCOSYVIO: A challenging visual inertial odometry benchmark. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 3847–3854.

Qin T, Li P and Shen S (2018) VINS-Mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics* 34(4): 1004–1020.

Qin T, Cao S, Pan J and Shen S (2019) A General Optimization-based Framework for Global Pose Estimation with Multiple Sensors. *arXiv preprint arXiv:1901.03642*

Schonberger JL and Frahm JM (2016) Structure-from-motion revisited. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4104–4113.

Schubert D, Goll T, Demmel N, Usenko V, Stückler J and Cremers D (2018) The TUM VI benchmark for evaluating visual–inertial odometry. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 1680–1687.