2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Madrid, Spain, October 1-5, 2018

The TUM VI Benchmark for Evaluating Visual-Inertial Odometry

David Schubert*, Thore Goll*, Nikolaus Demmel*, Vladyslav Usenko*, Jörg Stückler and Daniel Cremers

***Abstract*— Visual odometry and SLAM methods have a large variety of applications in domains such as augmented reality or robotics. Complementing vision sensors with inertial measurements tremendously improves tracking accuracy and robustness, and thus has spawned large interest in the development of visual-inertial (VI) odometry approaches. In this paper, we propose the TUM VI benchmark, a novel dataset with a diverse set of sequences in different scenes for evaluating VI odometry. It provides camera images with 1024x1024 resolution at 20 Hz, high dynamic range and photometric calibration. An IMU measures accelerations and angular velocities on 3 axes at 200 Hz, while the cameras and IMU sensors are time-synchronized in hardware. For trajectory evaluation, we also provide accurate pose ground truth from a motion capture system at high frequency (120 Hz) at the start and end of the sequences which we accurately aligned with the camera and IMU measurements. The full dataset with raw and calibrated data is publicly available. We also evaluate state-of-the-art VI odometry approaches on our dataset.**

I. INTRODUCTION

Visual odometry and SLAM is a very active field of research with an abundance of applications in fields such as augmented reality or robotics. Variants include monocular ([1], [2]), stereo ([3], [4]) and visual-inertial ([5], [6], [7]) methods. Compared to one camera, adding a second one in a stereo setup provides better robustness and scale-observability. Adding an inertial measurement unit (IMU) helps dealing with untextured environments and rapid motions and makes roll and pitch directly observable. On the other hand, the camera complements the IMU with external referencing to the environment in 6 degrees of freedom.

To compare competing methods, it is necessary to have publicly available data with ground truth. Given the relevance of the topic of visual-inertial odometry, the availability of high-quality datasets is surprisingly small. Compared to single-camera, purely visual datasets, the challenge with a stereo visual-intertial dataset lies in the accurate synchronization of three sensors. A commonly used option for evaluating visual-inertial odometry is the EuRoC MAV dataset [8], but its image resolution and bit depth is not quite state-of-the-art anymore, and the number and variability of scenes is very limited.

For direct methods, which do not align pixel coordinates but image intensities, the assumption that the same 3D point has the same intensity in two different images should be satisfied. It has been shown that providing a photometric calibration that allows to compensate for exposure times, camera response function and lense vignetting is beneficial in this case [2], however it is not a common feature of existing datasets.

Fig. 1. The **TUM VI benchmark** includes synchronized measurements from an IMU and a stereo camera in several challenging indoor and outdoor sequences. The cameras are equipped with large field-of-view lenses (195°) and provide high dynamic range images (16 bit) at high resolution (1 MP) with linear response function. The figure shows example frames from the dataset.

In this paper, we propose the **TUM VI benchmark**, a novel dataset with a diverse set of sequences in different scenes, with 1024x1024 image resolution at 20 Hz, 16-bit color depth, known exposure times, linear response function and vignette calibration. An IMU provides 3-axis accelerometer and gyro measurements at 200 Hz, which we correct for axis scaling and misalignment, while the cameras and IMU sensors are time-synchronized in hardware. We recorded accurate pose ground truth with a motion capture system at high frequency (120 Hz) which is available at the start and end of the sequences. For accurate alignment of sensor measurements with the ground truth, we calibrated time offsets and relative transforms.

We evaluate state-of-the-art visual-inertial algorithms on our dataset. The full dataset with raw and calibrated data, together with preview videos, is available on:

https://vision.in.tum.de/data/datasets/visual-inertial-dataset

II. RELATED WORK

Datasets have in the past greatly fostered the research of visual odometry and SLAM algorithms. In table I we give an

* These authors contributed equally. The authors are with Technical University of Munich, 85748 Garching bei München, Germany {schubdav, gollt, demmeln, usenko, stueckle, cremers}@in.tum.de

TABLE I

COMPARISON OF DATASETS WITH VISION AND IMU DATA.

dataset	year	environ.	carrier	cameras	IMUs	time sync	ground truth	stats/props
Kitti Odometry [9]	2013	outdoors	car	1 **stereo** RGB 2x1392x512 @10Hz, 1 stereo gray 2x1392x512 @10Hz	OXTS RT 3003 3-axis acc/gyro @10Hz	sw	OXTS RT 3003 pose @10Hz, acc. <10cm	22 seqs, 39.2 km
Malaga Urban [10]	2014	outdoors	car	1 **stereo** RGB 2x1024x768 @20Hz	3-axis acc/gyro @100Hz	sw	GPS pos @1Hz, low acc	15 subseqs, 36.8 km
UMich NCLT [11]	2015	**in-/outdoors**	Segway	6 RGB (omni) 1600x1200 @5Hz	3-axis acc/gyro @100Hz	sw	fused GPS/IMU/laser pose @150Hz, acc≈10cm	27 seqs, 147.3 km
EuRoC MAV [8]	2016	indoors	MAV	1 **stereo** gray 2x752x480 @20Hz	ADIS16488 3-axis acc/gyro @200Hz	**hw**	laser tracker pos @20Hz, **motion capture pose @100Hz**, acc≈1mm	11 seqs, 0.9 km
PennCOSYVIO [12]	2017	**in-/outdoors**	handheld	4 RGB 1920x1080 @30Hz (rolling shutter), 1 **stereo** gray 2x752x480 @20Hz, 1 fisheye gray 640x480 @30Hz	ADIS16488 3-axis acc/gyro @200Hz, Tango 3-axis acc @128Hz / 3-axis gyro @100Hz	**hw** (stereo gray/ ADIS), sw	fiducial markers pose@30Hz, acc≈15cm	4 seqs, 0.6 km
Zurich Urban MAV [13]	2017	outdoors	MAV	1 RGB 1920x1080 @30Hz (rolling shutter)	3-axis acc/gyro @10Hz	sw	Pix4D visual pose, acc unknown	1 seq, 2 km
Ours (TUM VI)	2018	**in-/outdoors**	handheld	1 **stereo** gray 2x1024x1024 @20Hz	BMI160 3-axis acc/gyro @200Hz	**hw**	**partial motion capture pose @120Hz**, marker pos acc≈1mm (static case)	28 seqs, 20 km, **photometric calibration**

overview over the most relevant datasets that include vision and IMU data.

Visual odometry and SLAM datasets: The TUM RGB-D dataset [14] is focused on the evaluation of RGB-D odometry and SLAM algorithms and has been extensively used by the research community. It provides 47 RGB-D sequences with ground-truth pose trajectories recorded with a motion capture system. It also comes with evaluation tools for measuring drift and SLAM trajectory alignment. For evaluating monocular odometry, recently the TUM MonoVO dataset [15] has been proposed. The dataset contains 50 sequences in indoor and outdoor environments and has been photometrically calibrated for exposure times, lens vignetting and camera response function. Drift can be assessed by comparing the start and end position of the trajectory which coincide for the recordings. We also provide photometric calibration for our dataset, but additionally recorded motion capture ground truth in parts of the trajectories for better pose accuracy assessment. Furthermore, the above datasets do not include time-synchronized IMU measurements with the camera images like our benchmark.

For research on autonomous driving, visual odometry and SLAM datasets have been proposed such as Kitti [9], Malaga Urban dataset [10], or the Robot Oxford car dataset [16]. The Kitti and Malaga Urban datasets also include low-frequency IMU information which is, however, not time-synchronized with the camera images. While Kitti provides a GPS/INS-based ground truth with accuracy below 10 cm, the Malaga Urban dataset only includes a coarse position for reference from a low-cost GPS sensor. Our dataset contains 20 Hz camera images and hardware time-synchronized 3-axis accelerometer and gyro measurements at 200 Hz. Ground-truth poses are recorded at 120 Hz and are accurately time-aligned with the sensor measurements as well.

Visual-inertial odometry and SLAM datasets: Similar to our benchmark, some recent datasets also provide time-synchronized IMU measurements with visual data and have been designed for the evaluation of visual-inertial (VI) odometry and SLAM approaches. The EuRoC MAV dataset [8] includes 11 indoor sequences recorded with a Skybotix stereo VI sensor from a MAV. Accurate ground truth (approx. 1mm) is recorded using a laser tracker or a motion capture system. Compared to our benchmark, the sequences in EuRoC MAV are shorter and have less variety as they only contain recordings in one machine hall and one lab room. Furthermore, EuRoC MAV does not include a photometric calibration which is important to benchmark direct methods. Further datasets for visual-inertial SLAM are the PennCOSYVIO dataset [12] and the Zurich Urban MAV dataset [13]. However, they do not contain photometric calibration and as accurate ground truth or time-synchronization of IMU and camera images like our benchmark (cf. table I).

TABLE II
OVERVIEW OF SENSORS IN OUR SETUP.

Sensor	Type	Rate	Characteristics
Cameras	2 × IDS uEye UI-3241LE-M-GL	20 Hz	global shutter 1024x1024 16-bit gray
IMU	Bosch BMI160	200 Hz	3D accelerometer 3D gyroscope temperature
MoCap	OptiTrack Flex13	120 Hz	6D Pose infrared cameras
Light sensor	TAOS TSL2561	200 Hz	scalar luminance

III. SENSOR SETUP

Our sensor setup consists of two monochrome cameras in a stereo setup and an IMU, see fig. 2. The left figure shows a schematic view of all involved coordinate systems. We use the convention that a pose $\mathbf{T}_{BA} \in \mathrm{SE}(3)$ transforms point coordinates $\mathbf{p}_A \in \mathbb{R}^3$ in system A to coordinates in B through $\mathbf{p}_B = \mathbf{T}_{BA}\mathbf{p}_A$. For the coordinate systems, we use the following abbreviations,

I	IMU
C_0	camera 0
C_1	camera 1
M	IR-reflective markers
G	grid of AprilTags
W	world frame (reference frame of MoCap system)

The IMU is rigidly connected to the two cameras and several IR-reflective markers which allow for pose tracking of the sensor setup by the motion capture (MoCap) system. For calibrating the camera intrinsics and the extrinsics of the sensor setup, we use a grid of AprilTags [17] which has a fixed pose in the MoCap reference (world) system. In the following, we briefly describe the hardware components. An overview is also given in table II.

A. Camera

We use two uEye UI-3241LE-M-GL cameras by IDS. Each has a global shutter CMOS sensor which delivers 1024x1024 monochrome images. The whole intensity range of the sensor can be represented using 16-bit images, so applying a non-linear response function (usually used to increase the precision at a certain intensity range) is not required. The cameras operate at 20 Hz and are triggered synchronously by a Genuino 101 microcontroller.

The cameras are equipped with Lensagon BF2M2020S23 lenses by Lensation. These fisheye lenses have a field of view of $195°$ (diagonal), though our cameras record a slightly reduced field of view in horizontal and vertical directions due to the sensor size.

B. Light Sensor

We design our sensor setup to ensure the same exposure time of corresponding images for the two cameras. This way, both camera images have the same brightness for corresponding image points (which otherwise needs to be calibrated or estimated with the visual odometry). Furthermore, this also ensures the same center of the exposure time (which is used as the image timestamp) for two corresponding images and allows us to record accurate per-frame exposure times.

We use a TSL2561 light sensor by TAOS to estimate the required exposure time. The sensor delivers an approximate measurement of the illuminance of the environment. The relation of these measurements and the exposure times which are selected by the camera's auto exposure is approximately inversely proportional, as can be seen in fig. 3. We find its parameters using a least-squares fit and use it to set the exposure times of both cameras based on the latest illuminance measurement. This assumes that the change in scene brightness between the light measurement and the start of the exposure is negligible. Note that it is not necessary to reproduce the cameras' auto exposure control exactly as long as too dark or too bright images can be avoided. In most cases, the results of our exposure control approach are visually satisfying, but short video segments may be challenging.

C. IMU

Our sensor setup includes a Bosch BMI160 IMU, which contains 16-bit 3-axis MEMS accelerometer and gyroscope. IMU temperature is recorded, facilitating temperature-dependent noise models. We set its output rate to 200 Hz. The IMU is integrated in the Genuino 101 microcontroller board which triggers the cameras and reads the IMU values. This way, the timestamps of cameras and IMU are well aligned. We estimate the remaining small constant time offset (owing to the readout delay of IMU measurements) during the camera-imu extrinsics calibration which yields a value of $5.3\,\mathrm{ms}$ for our setup. We estimated this value once and corrected for it in both raw and calibrated datasets.

D. Motion Capture System

For recording accurate ground-truth poses at a high framerate of $120\,\mathrm{Hz}$, we use an OptiTrack motion capture system. It consists of 16 infrared Flex13 cameras which track the IR-reflective markers on the sensor setup. The MoCap system only covers a single room, so we cannot record ground truth for parts of the longer trajectories outside the room. Instead, all sequences start and end in the MoCap room such that our sequences provide ground truth at the beginning and the end.

IV. CALIBRATION

We include two types of sensor data in our dataset: raw data and calibrated data. The raw data is measured directly by the sensors as described so far, but cannot be used without proper calibration. In the following, we describe which calibrations we apply to the raw data in order to make it usable.

A. Camera Calibration

Firstly, we calibrate the camera intrinsics and the extrinsics of the stereo setup. We use one of the calib-cam sequences, where we took care to slowly move the cameras in front of the calibration grid to keep motion blur as small as possible.

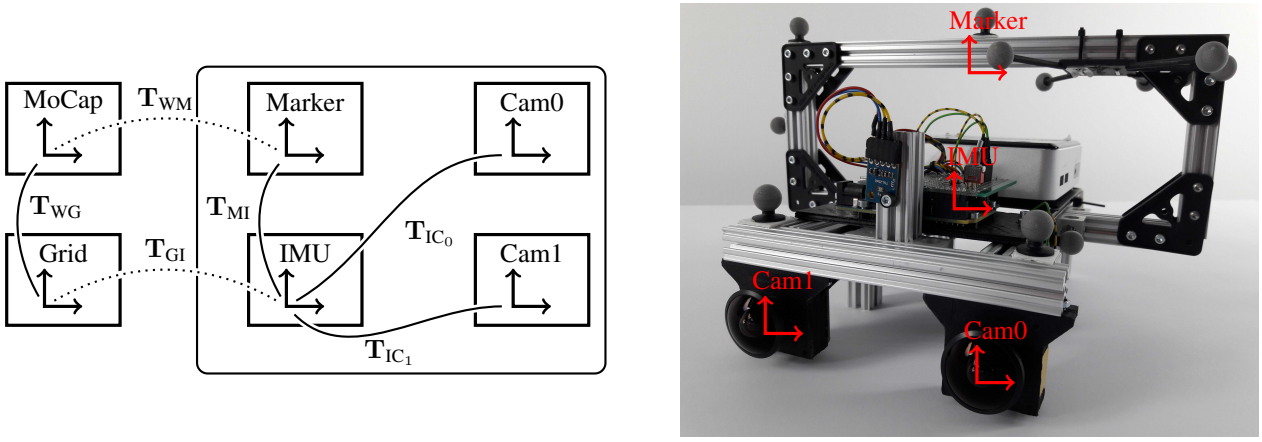

Fig. 2. Sensor setup. Left: Schematic view of the different coordinate systems. The rounded rectangle contains all components which are rigidly connected with the IMU coordinate system. A dotted line indicates a temporally changing relative pose when moving the sensor. Right: Photo of the sensor setup. It contains two cameras in a stereo setup, a microcontroller board with integrated IMU, a luminance sensor between the cameras and IR reflective markers.

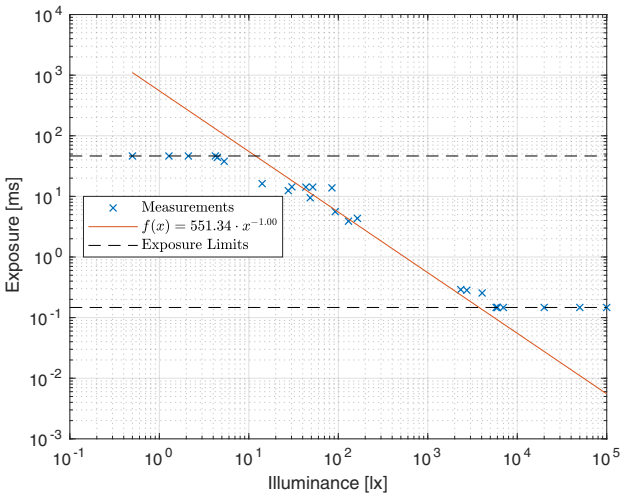

Fig. 3. Relation of illuminance measurements by our light sensor and corresponding exposure time settings by the camera's auto exposure mode. The dashed lines show the minimum and maximum exposure times possible. The red line shows the least-squares fit (without saturated values) which we use for estimating the next exposure time.

B. IMU and Hand-Eye Calibration

We then calibrate the extrinsics between IMU and cameras as well as between IMU and MoCap frame. Concurrently, we estimate the time-synchronization of IMU with MoCap measurements and IMU parameters such as axis alignment, scale differences and biases.

Specifically, we keep the camera intrinsics from the previous calibration fixed and optimize for

- the relative pose between cameras and IMU,
- the time shift between MoCap and IMU time,
- the time shift between camera and IMU time,
- the relative pose between the cameras,
- the relative poses $\mathbf{T}_{\text{MI}}$ and $\mathbf{T}_{\text{WG}}$,
- coarse initial accelerometer and gyroscope biases $\mathbf{b}_{\text{a}}$ and $\mathbf{b}_{\text{g}}$,
- axis scaling and misalignment matrices as in [18] $\mathbf{M}_{\text{a}}, \mathbf{M}_{\text{g}} \in \mathbb{R}^{3\times3}$.

The relative poses $\mathbf{T}_{\text{MI}}$ and $\mathbf{T}_{\text{WG}}$ are found through hand-eye calibration using a non-linear least squares fitting procedure. Using the relative poses, we convert raw MoCap poses $\mathbf{T}_{\text{WM}}$ to calibrated ground-truth poses $\mathbf{T}_{\text{WI}}$ for the IMU.

Additionally, we compensate for the time shift between MoCap and IMU time in the calibrated data. The time offset between MoCap and IMU has to be estimated for each sequence individually. To find the time offset, angular velocities are calculated from the MoCap poses and aligned with the gyroscope measurements. This is done — after a coarse alignment based on measurement arrival time — using a grid search with a stepsize of $100\,\mu$s. Then a parabola is fitted around the minimum and the minimum of the parabola is the resulting time offset. The results of this procedure can be seen in fig. 4. The ground-truth poses in the calibrated data are always given in IMU time.

We also compensate for axis/scale misalignment and initial biases of the raw accelerations $\mathbf{a}_{\text{raw}}$ and angular velocities $\boldsymbol{\omega}_{\text{raw}}$ using

$$\mathbf{a}_{\text{calibrated}} = \mathbf{M}_{\text{a}} \cdot \mathbf{a}_{\text{raw}} - \mathbf{b}_{\text{a}}\,, \tag{1}$$

$$\boldsymbol{\omega}_{\text{calibrated}} = \mathbf{M}_{\text{g}} \cdot \boldsymbol{\omega}_{\text{raw}} - \mathbf{b}_{\text{g}}\,. \tag{2}$$

The matrices $\mathbf{M}_{\text{a}}, \mathbf{M}_{\text{g}}$ account for rotational misalignments of gyroscope and accelerometer, axes not being orthogonal or axes not having the same scale. For $\mathbf{M}_{\text{g}}$, all 9 entries are optimized, whereas $\mathbf{M}_{\text{a}}$ is chosen to be lower triangular with 6 parameters. The remaining three parameters (rotation) are redundant and have to be fixed in order to obtain a well-constrained system.

In principle, it is not necessary to deduct $\mathbf{b}_{\text{a}}$ and $\mathbf{b}_{\text{g}}$, as inertial state estimation algorithms usually estimate a time-varying bias. However, we found that in our hardware setup there is a large IMU bias that is coarsely reproducible between sensor restarts and therefore approximate precalibration is reasonable. Note that estimating the biases

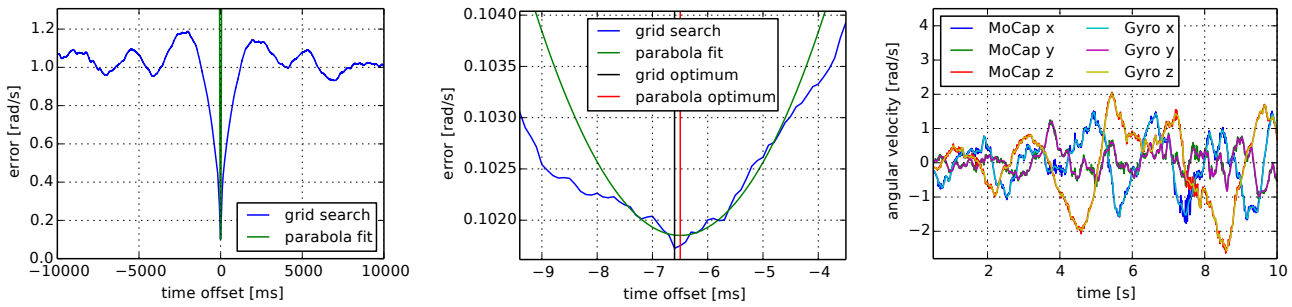

Fig. 4. Left and middle: Time alignment is performed using grid search. After a coarse initialization it is followed by parabola fitting to find the sub-discretization minimum. Right: Rotational velocities from gyroscope and MoCap after time alignment on the test sequence. MoCap angular velocities are computed using central differences on the orientation.

accurately from the sequences is still required for inertial state estimation.

For the calibration step, we use one of the calib-imu sequences which are recorded in front of the calibration grid with motions in all 6 degrees of freedom.

C. IMU Noise Parameters

For proper probabilistic modeling of IMU measurements in state estimation algorithms and accurate geometric calibration, the intrinsic noise parameters of the IMU are needed. We assume that our IMU measurements (accelerations or angular velocities) are perturbed by white noise with standard deviation σ_w and a bias that is slowly changing according to a random walk, which is an integration of white noise with standard deviation σ_b. To estimate these quantities, we analyse their Allan deviation $\sigma_\mathrm{Allan}(\tau)$ as a function of integration time τ. For a resting IMU with only white noise present, the Allan deviation relates to the white noise standard deviation as

$$\sigma_\mathrm{Allan}(\tau) = \frac{\sigma_\mathrm{w}}{\sqrt{\tau}}\,, \tag{3}$$

so the numerical value of the parameter σ_w can be found at $\tau = 1\,\mathrm{s}$. If the measurement is only perturbed by the bias, the relation is

$$\sigma_\mathrm{Allan}(\tau) = \sigma_\mathrm{b}\sqrt{\frac{\tau}{3}}\,, \tag{4}$$

which means the parameter can be found at $\tau = 3\,\mathrm{s}$. The relations between Allan deviation and integration time in Eqs. 3 and 4 can be found in [19]. White noise and bias dominate the Allan variance in different ranges of τ. Thus, in the log-log plot of $\sigma_\mathrm{Allan}(\tau)$ in fig. 5, a straight line with slope $-\frac{1}{2}$ has been fitted to an appropriate range of the data to determine σ_w, and a straight line with slope $\frac{1}{2}$ has been fitted to another range to determine σ_b.

D. Photometric Calibration

To enable good intensity matching for direct methods, we also provide vignette calibration. For this, we use the calibration code provided by the TUM MonoVO dataset[1] [15].

[1] https://github.com/tum-vision/mono_dataset_code

TABLE III
RMSE RPE OF THE EVALUATED METHODS ON 1 SECOND SEGMENTS

Sequence	OKVIS	ROVIO	VINS
room1	**0.013m / 0.43°**	0.029m / 0.53°	0.015m / 0.44°
room2	**0.015m / 0.62°**	0.030m / 0.67°	0.017m / 0.63°
room3	**0.012m / 0.63°**	0.027m / 0.66°	0.023m / 0.63°
room4	**0.012m** / 0.57°	0.022m / 0.61°	0.015m / **0.41°**
room5	**0.012m / 0.47°**	0.031m / 0.60°	0.026m / 0.47°
room6	**0.012m** / 0.49°	0.019m / 0.50°	0.014m / **0.44°**

The image formation model is given by

$$I(\mathbf{x}) = G\left(tV(\mathbf{x})B(\mathbf{x})\right)\,. \tag{5}$$

This means for an image point $\mathbf{x}$, light with intensity $B(\mathbf{x})$ is attenuated by a vignetting factor $V(\mathbf{x}) \in [0,1]$, then is integrated during the exposure time t, and finally is converted by a response function G into the irradiance value $I(\mathbf{x})$. In our case, we assume G linear, so the model simplifies to $I(\mathbf{x}) \propto tV(\mathbf{x})B(\mathbf{x})$. The given code requires images of a plane with a small calibration tag, taken from different viewpoints. It then alternatingly optimizes the texture of the wall (up to a constant factor) and a non-parametric vignette function. The result is a PNG image representing vignette values between 0 and 1 for each pixel.

V. DATASET

A. Sequences

Besides evaluation sequences, we also make our calibration data accessible such that users can perform their own calibration, even though we provide calibrated data and our calibration results. The sequences can be divided into the following categories.

- **calib-cam:** for calibration of camera intrinsics and stereo extrinsics. A grid of AprilTags has been recorded at low frame rate with changing viewpoints and small camera motion.
- **calib-imu:** for cam-imu calibration to find the relative pose between cameras and IMU. Includes rapid motions in front of the April grid exciting all 6 degrees of

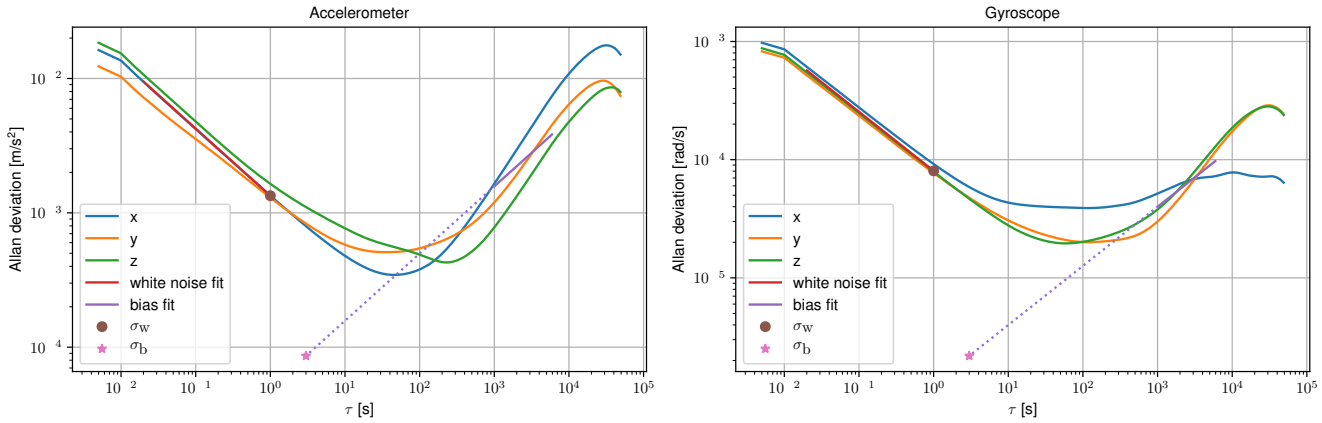

Fig. 5. Allan deviation of both accelerometer (left) and gyroscope (right). For the fit with slope $-1/2$ we averaged over all three dimensions and took the range $0.02 \leq \tau \leq 1$ into account. For the fit with slope $1/2$, the same averaging was done for the accelerometer, but for the gyroscope we only averaged the y-coordinate and the z-coordinate. The fit region is $1000 \leq \tau \leq 6000$. The assumed slope of 1/2 does not fit perfectly, which might be due to unmodeled effects such as temperature dependence. The numerical values of noise densities σ_w can be found at an integration time of $\tau = 1s$ on the straight line with slope $-1/2$, while bias parameters σ_b are identified as the value on the straight line with slope $1/2$ at an integration time of $\tau = 3s$. This results in $\sigma_w = 1.4 \times 10^{-3}\,\mathrm{m/s^2/\sqrt{Hz}}$, $\sigma_b = 8.6 \times 10^{-5}\,\mathrm{m/s^3/\sqrt{Hz}}$ for the accelerometer and $\sigma_w = 8.0 \times 10^{-5}\,\mathrm{rad/s/\sqrt{Hz}}$, $\sigma_b = 2.2 \times 10^{-6}\,\mathrm{rad/s^2/\sqrt{Hz}}$ for the gyroscope. The white noise parameters are similar to typical values provided by the manufacturer, $\sigma_w = 1.8 \times 10^{-3}\,\mathrm{m/s^2/\sqrt{Hz}}$ (accelerometer) and $\sigma_w = 1.2 \times 10^{-4}\,\mathrm{rad/s/\sqrt{Hz}}$ (gyroscope).

freedom. A small exposure has been chosen to avoid motion blur.

- **calib-vignette:** for vignette calibration. Features motion in front of a white wall with a calibration tag in the middle.
- **imu-static:** only IMU data to estimate noise and random walk parameters (111 hours standing still).
- **room:** sequences completely inside the MoCap room such that the full trajectory is covered by the ground truth.
- **corridor:** sequences with camera motion along a corridor and to and from offices
- **magistrale:** sequences featuring a walk around the central hall in a university building
- **outdoors:** sequences of a larger walk outside on a university campus
- **slides:** sequences of a walk in the central hall of a university building including a small part sliding in a closed tube with no visual features.

B. Format

1) ROS Bag Files: For each sequence, we provide three different ROS bag files, one raw bag and two calibrated ones. Raw bags contain the data as it has been recorded, i.e. before hand-eye, time shift or IMU calibration. They include the following topics.

`/cam0/image_raw`
`/cam1/image_raw`
`/imu0`
`/vrpn_client/raw_transform`

The first two contain the images of the cameras. Most fields in the messages are self-explanatory and follow standard conventions, but note that `frame_id` provides the exposure time in nanoseconds. In the IMU topic, we do not give the orientation, but we use the second entry of `orientation_covariance` to provide the temperature of the IMU in degree Celsius. The last topic contains the raw MoCap poses $\mathbf{T}_{WM}$. For each pose there is a timestamp in MoCap time, a translation vector and a rotation quaternion.

Calibrated bags contain the same topics as raw bags but with calibrated data. The differences are:

- MoCap poses have been aligned with the IMU frame (through hand-eye calibration, $\mathbf{T}_{WI}$),
- outlier MoCap poses have been removed with a median filter on positions,
- timestamps of the MoCap poses have been synchronized with the IMU time using the time shift calibration,
- IMU data has been processed according to eqs. (1) and (2).

We provide two kinds of calibrated bags: one with full resolution and one with quarter resolution (half resolution for each dimension). The downsampled version facilitates usage for users with storage or bandwidth limitations.

2) Calibration Files: We also provide geometric calibration files which have been obtained from the processed calibration bags using the Kalibr toolbox[2] [20]. They include intrinsic camera parameters for different models and the relative poses between cameras and IMU. Additionally, the vignette calibration result is given for each camera in PNG format as described in section IV-D.

VI. EVALUATION

A. Evaluation Metric

To evaluate the performance of tracking algorithms on the dataset, we use different evaluation metrics. The *absolute trajectory error* is used, which is the root mean squared

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

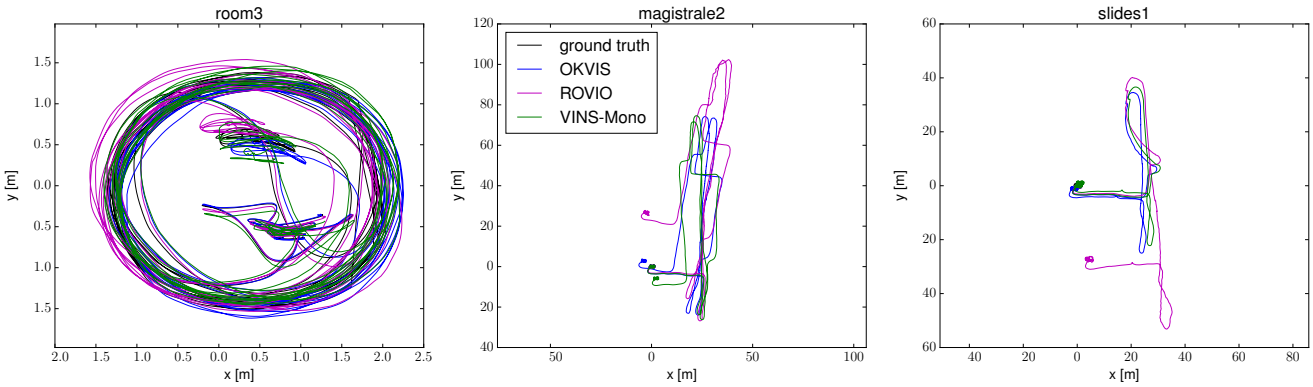

Fig. 6. Results of evaluated methods for room3, magistrale2 and slides1 sequences from our dataset. The ground truth is shown in black for the segments of the trajectory where it is available. The presented results are obtained with synchronous processing, without enforcing real-time and otherwise default parameters (except VINS-Mono for which non-realtime version is not implemented). Noise parameters are set to inflated values from the Allan plots in fig. 5 to account for unmodeled noise and vibrations.

difference of ground-truth 3D positions $\hat{\mathbf{p}}_i$ and the corresponding tracked positions $\mathbf{p}_i$, aligned with an optimal $\mathrm{SE}(3)$ pose $\mathbf{T}$,

$$r_{\text{ate}} = \min_{\mathbf{T}\in\mathrm{SE}(3)} \sqrt{\frac{1}{|I_{\text{gt}}|}\sum_{i\in I_{\text{gt}}} \left\|\mathbf{T}\mathbf{p}_i - \hat{\mathbf{p}}_i\right\|^2}\,. \tag{6}$$

All tracked poses where ground truth is available are used, which corresponds to indices I_{gt}. For most sequences, this is the case at the start and at the end, but for some sequences, there is ground truth throughout.

For visual odometry without global optimization, another reasonable quantity is the *relative pose error*. Following [14], it is defined as

$$r_{\text{rpe}} = \sqrt{\frac{1}{|I_{\text{gt},\Delta}|}\sum_{i\in I_{\text{gt},\Delta}} \left\|\mathrm{trans}(\mathbf{E}_i)\right\|^2}\,, \tag{7}$$

$$\mathbf{E}_i = \left(\hat{\mathbf{T}}_i^{-1}\hat{\mathbf{T}}_{i+\Delta}\right)^{-1}\left(\mathbf{T}_i^{-1}\mathbf{T}_{i+\Delta}\right)\,, \tag{8}$$

where $\mathrm{trans}(\cdot)$ takes the 3D translational component of a pose. This error measures how accurate pose changes are in a small time interval Δ. The set of frame indices $I_{\text{gt},\Delta}$ is the same as I_{gt}, but we have to take out Δ poses at the end of each tracked segment.

B. Results

To verify that the dataset is suitable for benchmarking visual-inertial odometry systems, we provide the results of several state-of-the-art methods that have open-source implementations. Unless specified otherwise, the methods are used with default parameters on quarter resolution images (512x512 pixels). We found that most of the algorithms have default parameters tuned to images with VGA resolution, which makes their performance better on sub-sampled datasets, while full resolution data might be useful for future research.

We provide evaluations for ROVIO [21], OKVIS [6] and VINS-Mono [22]. The results are summarised in table III and table IV and a visualization for some sequences is presented in fig. 6. All systems are able to track most of the sequences until the end, surprisingly, even the sequences with complete absence of visual features for some parts of the trajectory (slides). However, sometimes the estimators diverge at some point during the sequence, which results in erratic translation or rapid drift. We call a sequence diverged, if the ATE based on just the end-segment is larger[3] than $2\,\mathrm{m}$, which is indicated by underlines in table IV. The ATE values are still informative, as most often divergence happens towards the end (values larger than $1000\,\mathrm{m}$ are shown as "X").

OKVIS and VINS-Mono perform mostly well, but struggle for some of the longer outdoor sequences. ROVIO is more prone to drift and diverges on several sequences, which might be explained by it's use of a Kalman filter compared to computationally more demanding non-linear least squares optimization employed by OKVIS and VINS-Mono. VINS-Mono diverges on most of the outdoor sequences, but typically only after the camera returns to the motion capture room and switches from mainly forward motion to fast rotations. This might indicate a drift in accelerometer bias estimates.

The evaluation shows that even the best performing algorithms have significant drift in long (magistrale, outdoors) and visually challenging (slides) sequences. This means that the dataset is challenging enough to be used as a benchmark for further research in visual-inertial odometry algorithms.

VII. CONCLUSION

In this paper, we proposed a novel dataset with a diverse set of sequences in different scenes for evaluating visual-inertial odometry. It contains high resolution images with high dynamic range and vignette calibration, hardware synchronized with 3-axis accelerometer and gyro measurements. For evaluation, the dataset contains accurate pose ground truth at high frequency at the start and end of the sequences.

[3]For all evaluated systems, median values over all sequences for ATE based on just the start-segment are less than $0.1\,\mathrm{m}$ and less than $0.5\,\mathrm{m}$ for just the end-segment.

TABLE IV
RMSE ATE IN M OF THE EVALUATED METHODS

Sequence	OKVIS	ROVIO	VINS	length [m]
corridor1	**0.33**	0.47	0.63	305
corridor2	**0.47**	0.75	0.95	322
corridor3	**0.57**	0.85	1.56	300
corridor4	0.26	**0.13**	0.25	114
corridor5	**0.39**	2.09	0.77	270
magistrale1	3.49	4.52	**2.19**	918
magistrale2	**2.73**	13.43	3.11	561
magistrale3	1.22	14.80	**0.40**	566
magistrale4	**0.77**	39.73	5.12	688
magistrale5	1.62	3.47	**0.85**	458
magistrale6	3.91	X	**2.29**	771
outdoors1	X	101.95	**74.96**	2656
outdoors2	73.86	**21.67**	133.46	1601
outdoors3	32.38	**26.10**	36.99	1531
outdoors4	19.51	X	**16.46**	928
outdoors5	**13.12**	54.32	130.63	1168
outdoors6	**96.51**	149.14	133.60	2045
outdoors7	**13.61**	49.01	21.90	1748
outdoors8	**16.31**	36.03	83.36	986
room1	**0.06**	0.16	0.07	146
room2	0.11	0.33	**0.07**	142
room3	**0.07**	0.15	0.11	135
room4	**0.03**	0.09	0.04	68
room5	**0.07**	0.12	0.20	131
room6	**0.04**	0.05	0.08	67
slides1	0.86	13.73	**0.68**	289
slides2	2.15	**0.81**	0.84	299
slides3	2.58	4.68	**0.69**	383

We perform hand-eye calibration on calibration sequences and time-offset estimation on all sequences to have ground truth data geometrically and temporally aligned with the IMU. In addition, we provide sequences to calibrate IMU white noise and random walk and vignetting of the camera. The dataset is publicly available with raw and calibrated data.

We also use our benchmark to evaluate the performance of state-of-the-art monocular and stereo visual-inertial methods. Our results demonstrate several open challenges for such approaches. Hence, our benchmark can be useful for the research community for evaluating visual-inertial odometry approaches in future research.

ACKNOWLEDGMENT

This work was partially supported through the grant "For3D" by the Bavarian Research Foundation and through the grant CR 250/9-2 "Mapping on Demand" by the German Research Foundation.

REFERENCES

[1] H. Jin, P. Favaro, and S. Soatto, "Real-time 3D motion and structure of point features: a front-end system for vision-based control and interaction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Trans. Pattern Anal. Machine Intell.*, 2018.

[3] R. Wang, M. Schwörer, and D. Cremers, "Stereo DSO: Large-scale direct sparse visual odometry with stereo cameras," in *International Conference on Computer Vision (ICCV)*, 2017.

[4] C. F. Olson, L. H. Matthies, H. Schoppers, and M. W. Maimone, "Robust stereo ego-motion for long distance navigation," in *IEEE Int. Conference on Computer Vision and Pattern Recognition (CVPR)*, 2000.

[5] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[6] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visualinertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[7] V. Usenko, J. Engel, J. Stueckler, and D. Cremers, "Direct Visual-Inertial Odometry with Stereo Cameras," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

[8] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, 2016.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[10] J.-L. Blanco-Claraco, F.-A. Moreno-Duenas, and J. Gonzalez-Jimenez, "The Malaga urban dataset: High-rate stereo and LiDAR in a realistic urban scenario," *The International Journal of Robotics Research*, vol. 33, no. 2, pp. 207–214, 2014.

[11] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and lidar dataset," *International Journal of Robotics Research (IJRR)*, vol. 35, no. 9, pp. 1023–1035, 2015.

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

[13] A. L. Majdik, C. Till, and D. Scaramuzza, "The Zurich urban micro aerial vehicle dataset," *The International Journal of Robotics Research (IJRR)*, vol. 36, no. 3, pp. 269–273, 2017.

[14] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 573–580.

[15] J. Engel, V. Usenko, and D. Cremers, "A photometrically calibrated benchmark for monocular visual odometry," *arXiv preprint arXiv:1607.02555*, 2016.

[16] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[17] E. Olson, "Apriltag: A robust and flexible multi-purpose fiducial system," *University of Michigan, Tech. Rep*, 2010.

[18] J. Rehder, J. Nikolic, T. Schneider, T. Hinzmann, and R. Siegwart, "Extending kalibr: Calibrating the extrinsics of multiple imus and of individual axes," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 4304–4311.

[19] "IEEE standard specification format guide and test procedure for single-axis interferometric fiber optic gyros," *IEEE Std 952-1997*, pp. 1–84, Feb 1998.

[20] P. Furgale, J. Rehder, and R. Siegwart, "Unified Temporal and Spatial Calibration for Multi-sensor Systems," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 1280–1286.

[21] M. Bloesch, M. Burri, S. Omari, M. Hutter, and R. Siegwart, "Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback," *The International Journal of Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[22] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *arXiv preprint arXiv:1708.03852*, 2017.