The VCU-RVI Benchmark: Evaluating Visual Inertial Odometry for Indoor Navigation Applications with an RGB-D Camera

He Zhang, Lingqiu Jin and Cang Ye, *Senior Member, IEEE*

***Abstract*— This paper presents VCU-RVI, a new visual inertial odometry (VIO) benchmark with a set of diverse data sequences in different indoor scenarios. The benchmark was captured using an Structure Core (SC) sensor, consisting of an RGB-D camera and an IMU. It provides aligned color and depth images with 640×480 resolution at 30 Hz. The camera's data is synchronized with the IMU's data at 100 Hz. Thirty-nine data sequences covering a total of $\sim$3.7 kilometers trajectory were recorded in various indoor environments by two experimental setups: hand-holding the SC sensor or installing it on a wheeled robot. For the data sequences from the handheld SC, some were recorded in our laboratory under three challenging conditions: fast sensor motion, radical illumination changing, and dynamic objects, and the rest were collected in various indoor spaces outside the laboratory in the East Engineering Building, including corridors, halls, and stairways, during long-distance navigation scenarios. For the data sequences captured using the wheeled robot, half of them were recorded with sufficient IMU excitation in the beginning of the sequence, to meet the need of testing the VIO methods with the requirement of sufficient motion conditions for initialization. We placed three bumpers on the floor of the lab to create an uneven terrain to make the robot motion 6-DOF. The sequences also include data collected from navigational courses with a long trajectory. For trajectory evaluation, a motion capture system is used to generate accurate pose data (at a rate of 120 Hz), which will be used as the ground truth. We conducted experiments to evaluate the state-of-the-art VIO algorithms using our benchmark. These algorithms together with the evaluation tools and the VCU-RVI dataset are made publicly available.**

I. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a very active research topic with wide range of applications such as field robotics, augmented reality and autonomous vehicles [1]. In recent decades, extensive researches have been conducted to develop visual SLAM (vSLAM) methods: including monocular [2], [3], stereo [4] , RGB-D [5], and visual-inertial approaches [6], [7]. Compared to the camera-only-based methods, adding an inertial measurement unit (IMU) provides scale observability and improves robustness. On one hand, IMU measurement provides accurate short-period motion estimation in case of rapid motion or texture-less environments. On the other hand, the camera complements the IMU with long-period pose estimation. Therefore, the visual inertial odometry (VIO) [6], [7] has become more and more popular in the robotics communities. To support the development of VIO, many benchmarks have been published for monocular [6] or stereo camera-based VIO [7]. However, as far as we know, there is no benchmark for evaluating the performance of an RGB-D camera based VIO. There is a thread of researches for VIO algorithms that fuse the RGB-D camera data with IMU for motion state estimation with applications to 3D mapping [8], body shape reconstruction [9], ground robots [10], and blind navigation [11]. A good benchmark is needed to support future research in developing RGB-D camera based VIO algorithms.

Fig. 1: Data collection environments in the East Engineering Building. Top: a corridor, stairway, and hall; Bottom: the laboratory.

In this paper, we present the VCU-RVI benchmark, a new dataset with thirty-nine diverse sequences of data collected in the East Engineering Building of the Virginia Commonwealth University (snapshots shown in Fig. 1). The data sequences were obtained from a Structure Core (SC) sensor produced by Occipital Inc. As depicted in Fig. 2, the sensor consists of a color camera, a Infrared (IR) stereo camera, and an IMU, making itself an RGB-D camera based visual-inertial system (RGB-D VINS). The data sequences were recorded by hand-holding or robot-mounting the SC (Fig. 3). Each sequence contains: 1) a synchronized 640×480 color and depth image stream at 30 Hz, 2) IMU data stream at 100 Hz, and 3) ground truth pose data generated by a motion capture system at 120 Hz. For the data recorded in the laboratory, the ground truth is available for the whole sequence, while for the others, the ground truth is available at the start and the end of each sequence. We have evaluated the state-of-the-art VIO methods on our benchmark. To reproduce the experimental results, the dataset and the selected VIO methods' implementations are available at: `https://github.com/rising-turtle/VCU_RVI_Benchmark`.

This work was supported by the NIBIB and the National Eye Institute of the NIH under award R01EB018117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. H. Zhang, L. Jin and C. Ye are with Computer Science Department, Virginia Commonwealth University, Richmond, VA 23284, USA. (e-mail: `hzhang8@vcu.edu`, `jinl@mymail.vcu.edu`, `cye@vcu.edu`).

TABLE I: Comparison of Benchmark with Visual and Inertial Data

Dataset	year	carrier	cameras	IMUs	ground truth	stats
TUM RGBD [12]	2012	handheld, wheeled robot	1 RGB (rolling shutter) and 1 Depth (synchronized and aligned with RGB) 640x400 @30Hz	NONE	motion capture pose @120 Hz, marker pose acc. $\sim$ 1mm	47 seqs, 0.59 km
ICL-NUIM [13]	2014	handheld	synthetic 1 RGB and 1 Depth 640x480 @30Hz	NONE	perfect synthetic ground truth pose	8 seqs, 0.055 km
Kitti Odometry [14]	2013	car	1 stereo RGB 2x1392x512 @10Hz, 1 stereo gray 2x1392x512 @10Hz	OXTS RT3003 3-axis acc/gyro @10Hz	OXTS RT3003 pose @10Hz, acc. $<$10cm	22 seqs, 39.2 km
EuRoc MAV [15]	2015	MAV	1 stereo gray 2x752x480 @20Hz	ADIS16488 3-axis acc/gyro @200Hz	laser tracker pose @20Hz, motion capture pose @100Hz, acc $\sim$ 1mm	11 seqs, 0.9 km
PennCOS - YVIO [16]	2017	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	fiducial markers pose @30Hz, acc $\sim$ 15cm	4 seqs, 0.6 km
TUM VI [17]	2018	handheld	1 stereo gray 2x1024x1024 @20Hz	BMI160 3-axis acc/gyro @200Hz	partial motion capture pose @120Hz, marker pose acc. $\sim$ 1mm	28 seqs, 20 km
VCU-RVI (proposed)	2020	handheld, wheeled robot	1 RGB (global shutter) and 1 Depth (synchronized and aligned with RGB) 640x480 @30Hz	BMI1055 3-axis acc/gyro @100Hz	partial motion capture pose @120Hz, LED marker acc $\sim$ 1mm	39 seqs, 3.7 km

II. Related Work

Several datasets have been published in the literature to boost the development of vSLAM and VIO algorithms for various applications. Here, as illustrated in Table I, we briefly review the most relevant datasets that use an RGB-D camera or a visual-inertial system.

RGB-D camera-based odometry and SLAM datasets: The TUM RGB-D dataset [12] is targeted to evaluate the performances of RGB-D camera-based vSLAM algorithms for indoor navigation applications. It provides 47 RGB-D sequences with a Kinect carried by hand-holding or a wheeled robot in different indoor scenarios: floor, desk, room, and office. Furthermore, to test the robustness of vSLAM methods, it records sequences under various lighting, texture, or dynamic conditions. For each sequence, the ground-truth pose trajectory is recorded by a motion capture system. Evaluation tools are provided in the benchmark to align SLAM trajectory with ground-truth to compute pose estimation error. The ICL-NUIM dataset [13] contains RGB-D sequences generated by synthetically simulating indoor environments. Perfect ground-truth poses and surface model are provided for evaluating visual odometry and surface reconstruction. The aim of the TUM RGB-D dataset or the ICL-NUIM dataset focuses on evaluation for vision-only odometry or SLAM and they do not have any inertial measurement.

Visual-inertial odometry and SLAM datasets: The popular KITTI [14] vision benchmark is targeted for autonomous navigation. It recorded camera frames from one stereo color camera and one gray camera installed on a ground vehicle for outdoor use. The ground truth trajectories are obtained through a precise GPS/IMU sensor suite with accuracy below 10 cm. The frequencies of the camera and the IMU data are both 10 Hz, sufficing to capture the motion of a car but not a hand-held device. The PennCOSYVIO [16] is a VIO benchmark with synchronized data from a stereo-VI (visual-inertial) sensor suite (a stereo camera and an IMU), two Project Tango handheld devices, and three GoPro Hero 4 cameras. These sensors are placed on a handheld setup and used to record sequences along a $\sim$ 150m long-path. The frequencies of the camera data and IMU data are 30Hz and 200 Hz, respectively, sufficing to record the motion of a handheld device. The EuRoC MAV dataset [15] aims to benchmark VIO approaches in the application of autonomous MAV (micro aerial vehicle). It consists of 11 sequences of data captured from a stereo-VI suite on a MAV. To test the robustness of VIO to motion blur or illumination variance, some data sequences are recorded under fast motion speed or changing illumination conditions. Compared to EuRoC MAV, the recent TUM VI [17] dataset increases the indoor scenario varieties by recording sequences in four indoor settings: room, corridor, hall, and slide. For trajectory evaluation, the ground truth poses at the start and the end of each sequence are provided from a motion capture system. The room dataset is captured from a fish-eye stereo-VI sensor suite on a handheld rig and it also contains ground truth poses throughout. Compared to these VI datasets, the VCU-RVI dataset composes of thirty-nine data sequences from an RGB-D-VI sensor suite (an RGB-D camera and an IMU). The RGB-D camera employs an active stereo imaging technologyto obtain depth measurements. Hence, it can provide more dense depth data than a stereo camera in textureless and poor illumination scenarios. The dense depth data is transformed into the coordinate system of the color camera and reprojected onto the color image to generate an aligned depth image. To our knowledge, this is the first benchmark containing the synchronized sequences from an RGB-D-VI sensor suite. Furthermore, the senor suite is carried by both hand or a wheeled robot in various indoor space, making the dataset suitable for evaluating VIO algorithms in different indoor navigation applications such as augmented reality [18], rescue ground vehicle [10], blind navigation [11], etc.

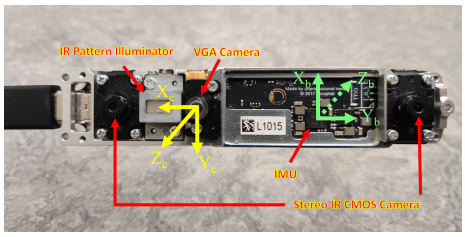

Fig. 2: Components of Structure Core: a color VGA camera, an IR stereo cameras, an IR projector and an IMU. The body/IMU and camera coordinate systems are denoted by $\{B\}$ $(X_bY_bZ_b)$ and $\{C\}$ $(X_cY_cZ_c)$, respectively. The initial $\{B\}$ is taken as the world coordinate system $\{W\}$. In this paper, the super scripts b and c describe a variable in $\{B\}$ and $\{C\}$, respectively. The transformation matrix between $\{B\}$ and $\{C\}$ is pre-calibrated and denoted as $T_c^b = [R_c^b;\ \mathbf{t}_c^b]$.

III. Sensor Setup

A. RGB-D Camera

The RGB-D camera we chose is the Structure Core (SC) as shown in Fig. 2. It has a color camera and an Infrared (IR) stereo camera. IR stereo camera is capable of determining the depth ranging from 0.4 to more than 5 meters for the points on the image plane. Moreover, it uses an IR laser projector to project a static pattern on the scene to facilitate stereo matching to improve the quality of its depth data. A disparity map from the stereo camera is used to generate the depth data through triangulation. Using the known displacement between the color camera and the IR stereo camera, the depth data is aligned with the color image. The SC can provide aligned color and depth imaging data (resolution: 640×480 pixels) at 30 fps. Also, it has a built-in IMU sensor that is able to provide both accelerometer and gyroscope measurements at a rate up to 800 fps. The integration of the IMU makes the SC an RGB-D VINS. The coordinate systems of the IMU and the camera are defined in Fig. 2.

B. IMU

The built-in IMU in SC is Bosch BMI055. It consists of a 12-bit tri-axial accelerometer and a 16-bit tri-axial gyroscope, providing a 6-axis inertial measurement. The resolution of the accelerometer and the gyroscope are $0.98mg$ and $0.004\ ^\circ/s$, respectively. The reported noise density for the accelerometer and the gyroscope are $150\mu g\frac{1}{\sqrt{Hz}}$ and $1.745\times10^{-4}\frac{rad}{s^2}\frac{1}{\sqrt{Hz}}$, respectively. In the dataset, their update rates were both set to 100 Hz.

C. Motion Capture System

The OptiTrack motion capture (MoCap) system is set up in a $6m\times6m$ laboratory space (Fig. 1). Our system utilizes 12 infrared Prime 17 cameras to record the positions of LED target (consisting of 5 IR LEDs mounted on a 3D-printed bracket as depicted in Fig. 3 left). As the SC is also mounted on the bracket, the ground truth poses can be computed from the positional data of the five LEDs. The MoCap system is accurately-calibrated with a mean error of 0.847mm. We set the MoCap to record its data at a frame rate of 120 Hz.

IV. Calibration

A. Camera Intrinsic Calibration

We recorded a sequence of data by slowly moving the SC in front of a 6×6 Aprilgrid with a size of $0.8\times0.8m$.

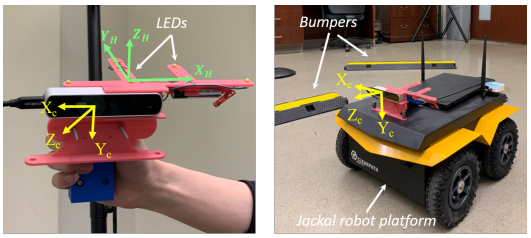

Fig. 3: Handheld/wheel-robot-installed SC for data collection

Then, we used the Kalibr toolbox [19] to calibrate the camera intrinsic parameters based on the pinhole camera model and radtan distortion model. The calibrated parameters are almost the same as what are provided by Occipital.

B. IMU Intrinsic Calibration

We calibrated the IMU by assuming that the IMU measurements (accelerations and angular velocities) are perturbed by a white noise with a standard deviation σ_w and a bias that is slowly changing due to random walk (an integration of a white noise with a standard deviation σ_b). To obtain calibration data, we fixed the SC on a vise and recorded the data of the gyroscope and the accelerometer at 100Hz for 5 hours. The computed noise density and random walk bias by Kalibr toolbox [19] are reported in Table II.

TABLE II: IMU Noise and Random Walk Bias

	Noise density	Random walk
accelerometer	$1.19\times10^{-4}\frac{m}{s^2}\frac{1}{\sqrt{Hz}}$	$5.35\times10^{-6}\frac{m}{s^3\sqrt{Hz}}$
gyroscope	$2.08\times10^{-4}\frac{rad}{s}$	$2.0\times10^{-7}\frac{rad}{s^2}\frac{1}{\sqrt{Hz}}$

C. Camera-IMU Extrinsic Calibration

We recorded a sequence of camera-IMU data (about 3 minutes long) and ran the Kalibr calibration tool [19] on the data to estimate the extrinsic transformation matrix (T_c^b) between the color camera and the IMU. We also measured the translation from the color camera to the IMU with a vernier scale. Compared to the calibration result, the extrinsic T_c^b provided by Occipital is more close to the manually measured translation. Therefore, we used the extrinsic transformation matrix provided by Occipital in our evaluation.

D. MoCap-IMU Extrinsic Calibration

In order to compare the VIO result with the ground truth in the MoCap coordinate system, the extrinsic transformation matrix (T_b^M) from the IMU to the MoCap is required. It is computed by $T_b^M = T_H^M T_c^H (T_c^b)^{-1}$. The transformation matrix from the LED-target to the Mocap coordinate system (T_H^M) and the camera-IMU extrinsic transformation matrix (T_c^b) are known. T_c^H is the transformation matrix from the camera to the LED-target coordinate system and it can be obtained through a hand-eye calibration approach [20] by the following steps: 1) recording a data sequence by moving the SC around a checkerboard in such a way that the checkerboard is in the SC' field of view (FOV), 2) computing the camera's poses by using the calibration method [21], 3) estimating T_c^H based on the camera poses and the LED-target's poses by using Daniilidis' method

[20]. This procedure has achieved a final pose residual of ~ 0.6703 mm.

E. MoCap-IMU Time Synchronization

For each data sequence, the timestamps between the MoCap and the IMU must be synchronized. To achieve this, we adopted the grid search approach [17] to achieve time synchronization. When capturing each data sequence, we kept the SC motionless on purpose for 3 seconds after a short-period of movement ($\sim$ 5 seconds) at the beginning. This created a distinctive pattern between the IMU-measured and the MoCap-generated motion data (i.e., angular rates) to align the two data sources and synchronize their times. In so doing, the data at the first 20 seconds are used for time-synchronization by the grid search. VIO pose estimation results on the time-synchronized data sequences will be compared with the ground truth for performance evaluation.

V. DATASET

Each data sequence in our dataset consists of a ground truth pose data provided by the Mocap and visual-inertial data obtained from the SC. There are two types of the data sequences that were recorded by hand-holding or robot-mounting the SC.

A. Handheld SC

According to the data recording conditions and environments, the data sequences are divided into the following categories:

- **lab**: a data sequence captured inside the laboratory with ground truth along the entire trajectory.
  - **lab-simple**: the SC moved at a normal speed in a static environment with constant illumination.
  - **lab-motion**: the SC moved at a high speed in a static environment with constant illumination. The mean and max of the rotation speed are about 35 and 120 degree per second, respectively.
  - **lab-light**: the SC moved at a normal speed in a static environment with varying illumination. For some sequences, the illumination condition changes from normal indoor lightness to complete darkness.
  - **lab-dynamic**: the SC moved at a normal speed in a dynamic environment with constant illumination. The dynamic objects include 1-2 persons, a chair, a wheeled robot, or a rollator.
- **corridor**: data sequences recorded by moving the SC along corridors.
- **hall**: data sequences recorded by moving the SC in corridors, over stairways, and in one/two halls.

B. Robot-mounted SC

For the data sequences recorded by using the wheeled robot, they were captured in the laboratory and along the corridors. The wheeled robot was controlled by a bluetooth gamepad to move linearly from point to point. Since the acceleration measurements of the IMU have little variance when it is controlled to move linearly [10], resulting in insufficient IMU excitement for VIO initialization [6], [10]. We used the following two ways to avoid the problem:

- **manual**: first hand-holding and rotating the SC for about five seconds to produce sufficient IMU excitement for a good system initialization and then placing the SC on the robot and driving the robot to move.
- **bumper**: first driving the robot to move across the bumper to produce sufficient variation in IMU measurements for initialization and then driving the robot around in the whole area.

C. Calibration

In addition to the above types, we also collected three data sequences for users to calibrate the camera-IMU and the camera-LED system.

- **camera**: sequence with slow motion for camera intrinsic calibration.
- **camera-imu**: sequence with fast motion for camera-imu extrinsic calibration.
- **hand-eye**: sequence with slow motion for camera-marker extrinsic calibration.

The **camera** and the **camera-imu** were recorded by moving the SC around an array of AprilTags and keeping the AprilTags within the SC's FOV during the entire data collection process. The **hand-eye** was captured by moving the SC around a checkerboard and maintaining the checkerboard within the SC's FOV throughout the data sequence.

VI. EVALUATION

A. Evaluation Metric

We employ the absolute position error (APE) to evaluate the pose estimation accuracy of a VIO method. The APE is the root mean squared difference between the position estimation $\hat{p}_i$ and the ground-truth 3D position p_i, aligned with an transformation matrix T.

$$e_{ape} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \|T\hat{p}_i - p_i\|^2} \tag{1}$$

where N is the total number of estimated positions that have corresponding ground truth data. The transformation matrix T is computed by an ICP algorithm [22] using the first 20 seconds data to satisfy the following condition:

$$\mathbf{T}_e > \sqrt{\frac{1}{M}\sum_{i=h+1}^{h+M} \|T\hat{p}_i - p_i\|^2} \tag{2}$$

where $M = 100$, $T = ICP(p_i, \hat{p}_i), i = h+1, ..., h+M$, threshold $\mathbf{T}_e$ is set as $10mm$, and h is the last data point for the 3 seconds motionless period. The index of h is provided along with each data sequence. For the data sequences collected by the robot with bumpers, travelling over bumpers may lead to robot body vibration that may cause blurred images and increase IMU measurement noise [10] and thus degrade the VIO's pose estimation accuracy. In this case, the transformation matrix T computed by only using the data of the first 20 seconds is inaccurate. Therefore, for data

sequences **lab-bumper***, we used all pose data to compute the transformation matrix T for alignment.

TABLE III: APE in meters of the evaluated methods for handheld data

Data Sequence	[6]	[10]	[23]	Length [m]
lab-simple1	0.142	0.137	**0.109**	15
lab-simple2	0.175	0.431	**0.151**	17
lab-simple3	0.236	**0.116**	0.123	23
lab-motion1	**0.526**	0.581	0.532	66
lab-motion2	0.699	0.66	**0.565**	73
lab-motion3	0.452	0.361	**0.253**	64
lab-motion4	0.54	**0.488**	0.55	62
lab-motion5	0.373	0.378	**0.372**	80
lab-motion6	0.706	0.725	**0.592**	84
lab-light1	X	X	X	49
lab-light2	**0.4**	0.48	X	69
lab-light3	0.384	X	**0.378**	42
lab-light4	0.29	3.11	**0.27**	39
lab-light5	X	**1.12**	X	53
lab-light6	**1.06**	X	X	53
lab-dynamic1	0.287	**0.216**	0.235	33
lab-dynamic2	0.55	0.324	**0.215**	32
lab-dynamic3	0.653	0.52	**0.471**	40
lab-dynamic4	**1.7**	2.5	3.6	44
lab-dynamic5	0.469	0.177	**0.073**	21
corridor1	4.39	5.13	**3.23**	210
corridor2	**1.61**	1.81	3.23	206
corridor3	**3.97**	6.81	5.23	210
corridor4	4.33	**1.95**	2.2	210
hall1	4.8	2.34	**1.51**	251
hall2	12.71	7.98	**3.23**	309
hall3	X	X	**6.52**	357

TABLE IV: APE in meters of the evaluated methods for robot data

Data Sequence	[6]	[10]	[23]	Length [m]
lab-manual1	X	1.174	**0.179**	28
lab-manual2	0.402	0.308	**0.191**	25
lab-manual3	0.971	X	**0.134**	43
lab-bumper1	X	X	X	21
lab-bumper2	0.384	X	**0.372**	30
lab-bumper3	X	**0.278**	X	17
lab-bumper4	X	**0.297**	0.458	23
lab-bumper5	X	0.269	**0.193**	30
corridor-manual1	X	21.42	**21.39**	206
corridor-manual2	X	X	X	210
corridor-bumper1	X	X	**10.79**	206
corridor-bumper2	X	X	X	210

B. Results

To show that the benchmark is suitable for evaluating VIO methods, we provide the results of three methods: VINS-Mono [6], VINS-RGBD [10], and DUI-VIO [23]. According to the benchmark comparison study in [24], VINS-Mono [6] achieves the most accurate result due to its robust initialization strategy. Therefore, we use VINS-Mono as one method for evaluation to show that the dataset can also be used to benchmark monocular camera based VIO approaches. VINS-RGBD extends VINS-Mono by adding depth information to obtain the absolute scale in the initialization process. Furthermore, in the optimization process, VINS-RGBD sets the inverse depth of a visual feature as a constant by assuming the inverse depth measurements for visual features are accurate enough. DUI-VIO extends VINS-RGBD by exploiting the uncertainty information of the depth data. Both VINS-RGBD and DUI-VIO rely on the data from a RGB-D camera and an IMU, and they differ in the strategies to use the depth measurements to improve pose estimation accuracy.

The results are summarized in Tables III and IV. The trajectories estimated by the three methods are compared with the ground truth in Fig. 4. For the handheld data in Table III, all methods can ran successfully on almost all the sequences. In the Table, a "X" indicates that the method failed in initialization or resulted in a large APE ($>$50 meters) and the smallest APE in each row is bolded. All the methods performed mostly well for lab-easy and lab-motion sequences, but they failed on some of the lab-light data sequences. This was due to that the sudden illumination change caused the images lose almost all of the visual features. Lack of sufficient visual constraints made the IMU bias estimation less accurate, resulting in a diverged trajectory. For the lab-dynamic sequence, the accuracy of these methods varies significantly. For example, DUI-VIO generates the smallest APE for lab-dynamic5 but the largest APE for lab-dynamic4. On the contrary, VINS-Mono results in the smallest APE in lab-dynamic4 yet the largest APE in lab-dynamic5. All three methods achieved much worse results on lab-dynamic4 than on the other sequences. It seems that different dynamic situations (e.g. one/two moving people, or moving chair, robot) may affects the VIO methods differently. For the corridor and hall sequences with a longer trajectory, VINS-Mono, VINS-RGBD, and DUI-VIO have generated results with a larger APE ($>$4 meters) in the 2nd, 3rd, and 4th sequences, respectively. The trajectories estimated by VINS-Mono and VINS-RGBD also diverge at some points during processing the hall3 sequence.

For the robot data in Table IV, VINS-Mono, VINS-RGBD, and DUI-VIO failed in the 4th, 6th, and 9th sequences, respectively. Also for the corridor sequences, all methods have a very large APE ($>$10 meters). The degraded performance of VINS-Mono for the wheeled robot data is because that the motion of the wheeled robot is quite different from that generated by the hand. The general movement of a wheeled robot, such as moving along straight lines or circular arcs, makes the VINS fail to observe the scale information [25]. This explains why the trajectories estimated by VINS-Mono drift in the 9th sequences. The degraded performance of VINS-RGBD and DUI-VIO might be caused by the noisy IMU measurements as the results of the body vibration when the robot moves on the even terrain [10]. Therefore, to improve its pose estimation performance for a wheeled robots, a VIO should take into account the scale observability [25], noise filtering [10], etc.

The evaluation case study verifies that our dataset is suitable for benchmarking RGB-D/monocular camera based VIO approaches. The result shows that these three selected algorithms produce a significant drift on data sequences collected under challenging conditions, such as drastic lighting variation, dynamic operating environment, rough sensor motion caused by the robot's moving on uneven terrain, etc.

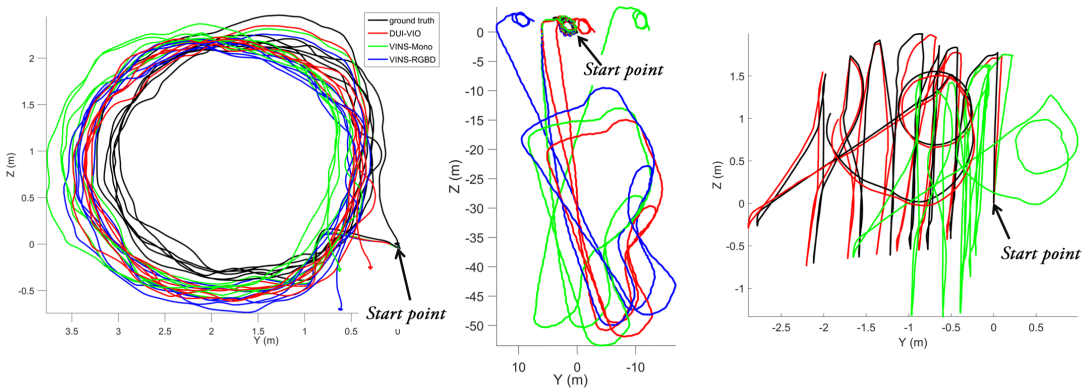

Fig. 4: Trajectories estimated by the three methods for lab-motion3 (left), hall2 (middle), and lab-manual2 (right)

Therefore, the dataset serves its purpose well as a benchmark to evaluate the performances of VIO algorithms under both normal and extreme conditions for the future research.

VII. Conclusion

In this paper, we present VCU-RVI benchmark to evaluate the VIO methods for indoor navigation applications. The dataset can be used to test a monocular or RGB-D camera based VIO methods. It includes thirty-nine data sequences recording a VINS' data in both normal cases and three challenging cases: fast motion, radical illumination change, and dynamic objects. It contains different indoor scenarios including laboratory, corridor, stairway, and hall. The data sequences were recorded by hand-holding or robot-mouting the VINS to evaluate VIO performance for different applications. The ground truth is provided and manually synchronized with the IMU's data for each sequence. To make the dataset easy to use, three VIO methods and the tool to generate the APE result, along with the data sequences, are provided in our benchmark.

References

[1] M. Ben-Ari and F. Mondada, "Robots and their applications," in *Elements of Robotics*. Springer, 2018, pp. 1–20.

[2] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[4] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2015, pp. 1935–1942.

[5] I. Dryanovski, R. G. Valenti, and J. Xiao, "Fast visual odometry and mapping from rgb-d data," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, 2013, pp. 2305–2310.

[6] T. Qin, P. Li, and S. Shen, "Vins-mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018.

[7] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visual-inertial slam using nonlinear optimization," in *Proceedings of Robotis Science and Systems*, 2013.

[8] T. Laidlow, M. Bloesch, W. Li, and S. Leutenegger, "Dense rgb-d-inertial slam with map deformations," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2017, pp. 6741–6748.

[9] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu, "Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus," in *Proc. of ECCV*, 2018, pp. 384–400.

[10] Z. Shan, R. Li, and S. Schwertfeger, "Rgbd-inertial trajectory estimation and mapping for ground robots," *Sensors*, vol. 19, no. 10, p. 2251, 2019.

[11] H. Zhang and C. Ye, "A visual positioning system for indoor blind navigation," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2020.

[12] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of IEEE/RSJ IROS*, 2012, pp. 573–580.

[13] A. Handa, T. Whelan, J. McDonald, and A. J. Davison, "A benchmark for rgb-d visual odometry, 3d reconstruction and slam," in *Proc. of IEEE Int. Conf. on Robotics and automation*, 2014, pp. 1524–1531.

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[15] 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*, vol. 35, no. 10, pp. 1157–1163, 2016.

[16] B. Pfrommer, N. Sanket, K. Daniilidis, and J. Cleveland, "Penncosyvio: A challenging visual inertial odometry benchmark," in *Proc. of Int. Conf. on Robotics and Automation*, 2017, pp. 3847–3854.

[17] D. Schubert, T. Goll, N. Demmel, V. Usenko, J. Stückler, and D. Cremers, "The tum vi benchmark for evaluating visual-inertial odometry," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2018, pp. 1680–1687.

[18] P. Li, T. Qin, B. Hu, F. Zhu, and S. Shen, "Monocular visual-inertial state estimation for mobile augmented reality," in *Proc. of IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2017, pp. 11–21.

[19] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2013, pp. 1280–1286.

[20] K. Daniilidis, "Hand-eye calibration using dual quaternions," *The International Journal of Robotics Research*, vol. 18, no. 3, pp. 286–298, 1999.

[21] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[22] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.

[23] H. Zhang and C. Ye, "DUI-VIO: depth uncertainty incorporated visual inertial odometry based on an rgb-d camera," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2020.

[24] J. Delmerico and D. Scaramuzza, "A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots," in *Proc. of IEEE ICRA*, 2018, pp. 2502–2509.

[25] K. J. Wu, C. X. Guo, G. Georgiou, and S. I. Roumeliotis, "Vins on wheels," in *Proc. of IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 5155–5162.