2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

A Multi-spectral Dataset for Evaluating Motion Estimation Systems

Weichen Dai[1], Yu Zhang[1], Shenzhou Chen[2], Donglei Sun[3], and Da Kong[1]

2021 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9077-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRA48506.2021.9561906

***Abstract*— Visible images have been widely used for motion estimation. Thermal images, in contrast, are more challenging to be used in motion estimation since they typically have lower resolution, less texture, and more noise. In this paper, a novel dataset for evaluating the performance of multi-spectral motion estimation systems is presented. All the sequences are recorded from a handheld multi-spectral device. It consists of a standard visible-light camera, a long-wave infrared camera, an RGB-D camera, and an inertial measurement unit (IMU). The multi-spectral images, including both color and thermal images in full sensor resolution (640 $\times$ 480), are obtained from a standard and a long-wave infrared camera at 32Hz with hardware-synchronization. The depth images are captured by a Microsoft Kinect2 and can have benefits for learning cross-modalities stereo matching. For trajectory evaluation, accurate ground-truth camera poses obtained from a motion capture system are provided. In addition to the sequences with bright illumination, the dataset also contains dim, varying, and complex illumination scenes. The full dataset, including raw data and calibration data with detailed data format specifications, is publicly available.**

I. INTRODUCTION

In recent years, vision-based motion estimation methods such as visual odometry (VO) [1] and visual simultaneous localization and mapping (vSLAM) [2] have attracted full attention for their diverse applications. These methods have been investigated in great detail for standard cameras, which can take advantage of rich textures only in bright illumination. For example, in scenarios such as data center inspection, firefighting, and rescue, standard cameras cannot provide sufficient information due to inadequate color textures, smog cover, or dim illumination. Therefore, some different types of sensors are used to enhance the robustness of vision-based motion estimation methods. In environments with complex illumination conditions, adding long-wave infrared (LWIR) cameras can complement the texture with information from another spectrum. Hence, the multi-spectral setup with these two types of cameras can become a reliable information source for all-day vision or fog-penetrating localization.

To evaluate the performance of various multi-spectral SLAM and odometry methods involving multi-spectral sources, a complete dataset with ground truth is necessary. Compared with the dataset for stereo standard cameras [3]–[5], the availability of a hardware-synchronization multi-spectral dataset is minimal. Moreover, stereo matching between RGB images and thermal images remains a challenging problem due to the insufficiency of textures in the latter and the difference between modalities. Therefore, the dataset should also provide reference stereo correspondences between different modalities.

In this paper, a multi-spectral dataset is provided for evaluating multi-spectral motion estimation methods. The dataset includes a set of sequences in diverse illumination conditions obtained on a setup consisting of a standard camera, an LWIR camera, an IMU, and a Kinect2. In the dataset, each sequence contains the color, thermal, and depth images, as well as the ground-truth trajectory. The color and thermal images (640 $\times$ 480) are captured at 32Hz with hardware-synchronization. The dense depth images are provided by the Kinect2 camera at 30Hz, enabling the projection of one image onto the other. The acceleration and angular velocity are measured by an Xsens IMU for challenging sequences. The ground-truth poses are recorded from a motion capture system at 120Hz. All sensors are calibrated carefully for higher accuracy.

The whole dataset, including the raw and calibration data, calibration files, and tool codes, is available on

https://github.com/NGCLAB/multi-spectral-dataset.

The main contributions of this paper are as follows:

- A new hardware-synchronization multi-spectral dataset with groundtruth poses is provided for the evaluation of multi-spectral motion estimation systems.
- The additional depth images can be used to study stereo matching between the visible and LWIR spectra.

The rest of the paper is organized as follows. Related work is reviewed in Section II, and the platform is introduced in Section III. The details of time synchronization and evaluation are presented in Section IV and the dataset is described in V. Finally, the experiments conducted in this dataset are discussed in Section VI and conclusions are drawn in Section VII.

II. RELATED WORK

Several types of visual sensors have been used in motion estimation, including monocular standard cameras [6], [7], stereo standard cameras [8], event-cameras [9], and RGB-D cameras [10]. These motion estimation methods can be categorized into filter-based [11] and factor-graph-optimization-based methods [12]. Besides, methods [13], [14] with deep learning methods have attracted the interest

[1]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, {weichendai,zhangyu80,kongda1997}@zju.edu.cn
[2]Alibaba A.I. Labs, shenzhou.csz@alibaba-inc.com
[3]Centre for English Language Education, University of Nottingham Ningbo China, donglei.sun@nottingham.edu.cn
This work was supported by NSFC 62088101 Autonomous Intelligent Unmanned Systems, the National Natural Science Foundation of China (Grant No. 61673341), National Key R&D Program of China (2016YFD0200701-3), Double First Class University Plan (CN), the Project of State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No.ICT2021A10).

of the community in recent years. The majority of these methods focus on utilizing the information on the visible spectrum.

Since the performance of standard cameras can be significantly influenced by illumination, LWIR cameras with multiple configurations, such as omnidirectional thermal cameras [15] and stereo thermal cameras [16], have been explored. With a different sensor modality, LWIR cameras are illumination independent, but they have their shortcomings. As mentioned in [17], the notable shortcomings include high noise level, high dynamic range, and the unique non-uniformity correction (NUC) mechanism that causes image corruption. Due to these factors, the visual odometry using LWIR setups [18], [19] usually cannot provide accurate results as those methods based on visible light in most environments. Hence, the methods [20]–[22] that integrate the information from IMU and those [23], [24] that rely on deep learning algorithms present a more practical solution. Besides, multi-spectral methods using LWIR cameras as complementary sensors have attracted extensive attention [25], [26] for their potential to function in poorly illuminated environments.

For visible light sensors, there are several well-known datasets, where the illumination condition is usually stable. Depending on the application, the carriers may be handheld [27], cars [4], micro aerial vehicles [28], and underwater vehicles [29]. For the task of fusing multi-sensor information, most datasets provide synchronized sensor data in addition to visible images [30]. Meanwhile, since hardware synchronization of sensors is critical [31], some high quality datasets also designed hardware-synchronization devices. Moreover, motion capture systems such as VICON are used to obtain the ground truth in most datasets. In other datasets, the ground-truth trajectories are acquired through GPS or accurate 3D reconstruction.

Most of the existing multi-spectral datasets were generated for fundamental tasks such as detection [32], segmentation [33], tracking [34], stereo matching [35], and place recognition [36]. A publicly available dataset with ground truth is also critical for evaluating the performance of multi-spectral motion estimation methods. For the task of motion estimation, the KAIST multi-spectral dataset [37] was proposed for all-day vision tasks in outdoor environments covering a wide range from urban to residential regions. In addition to RGB information, agricultural robots [38], also measure light emissions in the near-infrared (NIR) spectrum for separating vegetation from the soil and other background data. Both datasets are designed for outdoor environments. However, in most indoor environments, since thermal images cannot provide rich textures, it is more challenging to design multi-spectral motion estimation. Although the Vivid dataset [39] provides multi-spectral images, it cannot provide data with hardware synchronization. As summarised in Table I, there is a lack of multi-spectral dataset with hardware synchronization for both indoor and outdoor environments. Moreover, since stereo matching is the foundation of stereo methods, providing depth data plays an essential role in augmenting the development and validation of cross-modality matching algorithms.

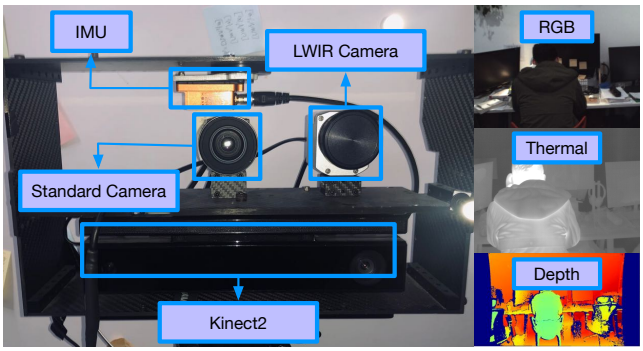


Fig. 1: The multi-spectral setup with Kinect2. *Left*: devices include an RGB camera, an LWIR camera, a Kinect2, an IMU, and the markers of motion capture systems. *Right*: example color image, thermal image, and depth image from Kinect2.

III. PLATFORM

The multi-spectral setup consists of a standard camera, an LWIR camera, an IMU, and a Kinect2, as shown in Fig. 1. To record accurate ground truth for the dataset, the OptiTrack motion capture system is used to provide 6D poses of the multi-spectral device using the markers on the setup.

The standard camera, LWIR camera, IMU, and Kinect2 are rigidly connected to a rack with OptiTrack markers. In addition to a printed chessboard, a unique chessboard whose surface is made of different materials is used to calibrate the system. Meanwhile, the parameters between the camera and the IMU are calibrated using kalibr [40] based on a grid of AprilTags [41], which is static in the scene. The necessary sensor information is summarized in Table II. In the following part, the key hardware component and how to calibrate it are briefly described.

A. The multi-spectral device

The multi-spectral device consists of a standard camera and an LWIR camera. The former is an ImageSource DFK 22BUC03. It uses a global shutter and captures 640×480 RGB images at 32 Hz. Moreover, this camera can be synchronized by an external trigger signal. The LWIR camera is an Optris PI 640, which produces 16-bit 640×480 thermal images and outputs a frame-sync trigger signal at 32Hz. This frame-sync trigger signal is set as the external trigger signal via a hardware connection to the standard camera. Therefore, this platform can provide synchronized color and thermal images at 32Hz when both cameras are set on trigger mode. The exposure time is set to the value less than the sensor synchronization period, which ensures that the captured images are at the same frequency.

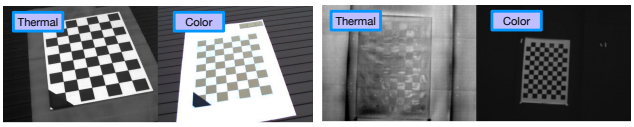
Due to the difference in information sources, the conventional printed board used to calibrate standard cameras cannot yield high contrast between textures in the thermal image. Therefore, a specialized equipment needs to be designed for calibration. There are two calibration methods: active [42], [43] and passive [19]. Active methods heat part

TABLE I: Comparison of multi-spectral datasets for motion estimation systems

Dataset	Year	Environment	Carrier	Spectrum	Baseline	Time sync	Ground truth
KAIST Odometry	2018	Urban	car	RGB @25Hz LWIR @25Hz	0.00m parallax-free	hw	OXTS RT 2002 pose @100Hz, acc. <2cm
Agricultural robot	2017	Terrain	car	RGB @30Hz NIR @30Hz	0.00m parallax-free	sw/hw	Leica RTK/Ublox EVK7-PPP position @10/4Hz, acc. <3/250cm
Vivid	2019	In-/outdoor	handheld	RGB @30Hz LWIR @20Hz	0.05m	sw	Cortex motion capture system /LeGO-LOAM pose @100Hz, acc. <1cm
Ours	2020	**In-/outdoor**	handheld	RGB @32Hz LWIR @32Hz	**0.14m**	**hw**	Optitrack motion capture system pose @120Hz, acc. <1cm

TABLE II: Overview of sensors in the setup

Sensor	Model	Rate	Characteristics
Standard color camera	ImageSource	32Hz	global shutter 640×480 RGB32
LWIR camera	Optris	32Hz	130 mK 640×480 16-bit thermal Spectral range 7.5-13 μm Non-uniformity correction
Depth camera	Kinect2	30Hz	Time of Flight 960×540 16-bit depth
IMU	Xsens	400Hz	3D accelerometer 3D gyroscope
Motion caption system	OptiTrack	120Hz	6-D pose Four cameras



(a) Our board (b) Printed board

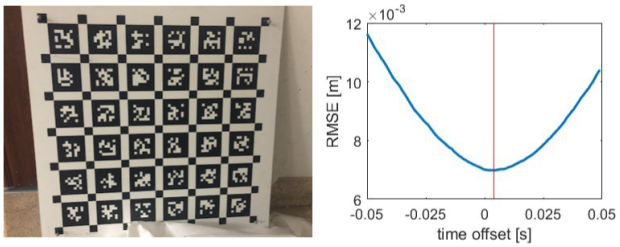
Fig. 2: Comparison of two calibration boards. A unique chessboard with the surface made of different materials can provide distinctive textures in both spectra. In contrast, the chessboard on the printed board cannot create distinct grids in the thermal camera.

of the calibration equipment to generate different thermal radiations. Passive methods do not depend on active heating or cooling.

Since active methods are inconvenient and complicated, a passive method is used for this setup. Exploiting the difference in the reflectivity of metallic and non-metallic materials in both the LWIR and visible spectra, an aluminum board with fiber squares on it is designed for calibration. This board provides distinct edges and excellent contrast in the sun, as shown in Fig. 2(a). During the calibration, the device followed a stop-capture-go manner to eliminate the error introduced by unsynchronized data. The intrinsic and extrinsic parameters of the multi-spectral device are obtained using MATLAB® tools [44].

B. Kinect2

To provide a reference for stereo matching between the visible and LWIR spectra, a Kinect2 is added to the setup.



(a) AprilTags board (b) Analysis of time delay

Fig. 3: (a) AprilTags board used for calibration and time synchronization. Four reflective markers are attached to the four small corner squares of the board. (b) The time delay between the motion capture system and the standard camera.

It integrates both the color and the depth camera. The depth images are 960×540, captured at 30Hz. Since Kinect2 follows the time of flight (ToF) measurement principle, the phase difference between the infrared illuminator and the infrared camera is used to compute the depth. The timestamp of the depth image is well aligned with the multi-spectral device by the software.

The infrared camera of Kinect2 is a near-infrared (NIR) camera. Since the NIR spectrum is next to the visible spectrum, both spectra share similar textures. Therefore, the printed chessboard can be directly used to calibrate these two cameras. The difficulty in image capture lies in that the NIR light emitter cannot be turned off, and hence additional effort is required. Meanwhile, to obtain clear NIR texture, the illumination should also be adjusted appropriately.

IV. TIME SYNCHRONIZATION

In the data sequences, every sensor should be synchronized with the standard camera, since incorrectly paired images will introduce error.

High-accuracy hardware-synchronization is achieved on the multi-spectral device using the frame-sync signal generated by the LWIR camera. Upon the rising pulse, the image from the standard camera is captured. An experiment was designed to check the performance of the synchronization. In this experiment, a ball was released in front of the board. As shown in Fig. 4, both cameras can capture the same scene with little time delay. The actual difference between the timestamps of those two images is less than 1ms. Hence, no modifications to the timestamps are required in the dataset.

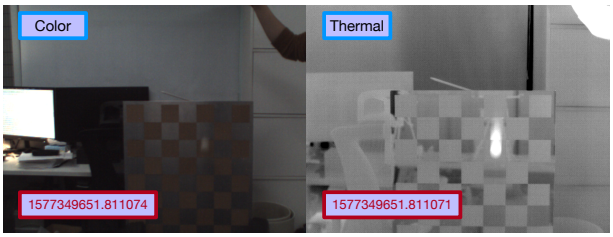


Fig. 4: Time-synchronization evaluation of multi-spectral images. In both images with the same timestamp, the position of the dropping ball indicates that two cameras can capture images almost simultaneously. Among most pairs, the thermal images are captured approximately 1ms later than the color images.

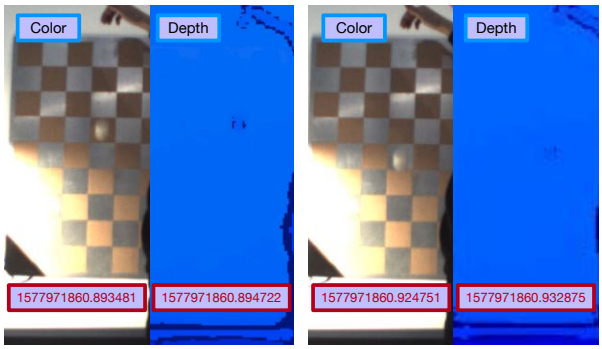


Fig. 5: Time-synchronization evaluation of Kinect2. Images from the standard camera and the depth camera. The timestamps in those red boxes indicate that the depth images are captured slightly later than the color images.

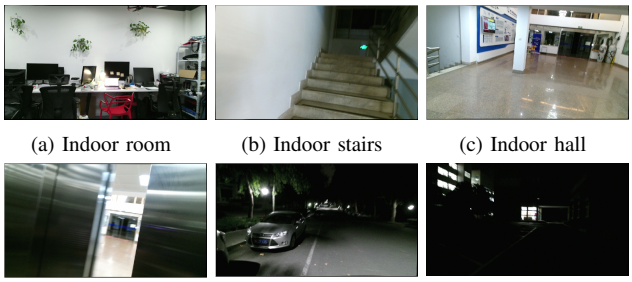
For the Kinect2 and the standard camera, the depth images are captured on average a little later than the color images, as shown in Fig. 5. The delay in the depth images is negligible, and no correction is required before data association.

The time delay between the motion capture system and the standard camera also needs to be determined before localization results can be evaluated. The time delay can be determined from the residuals of different time delays. As shown in Fig. 3(b), the poses from the motion capture system are approximately 4ms earlier than the images from the standard camera. The time delay is also trivial, and hence it is not necessary to modify the timestamps of the raw poses.

V. THE DATASET

The dataset includes both the evaluation sequences and the calibration data with a size of approximately 700GB. In the calibration data, both the raw data and our calibration results are provided. The dataset includes the following categories:

- Calibration:
  - *visible-LWIR*: the calibration data to compute the intrinsics and extrinsics of the multi-spectral device. The board shown in Fig. 2 is recorded in a stop-capture-go manner with changing viewpoints and small camera motion.
  - *visible-Kinect*: the calibration data to find the intrinsics and extrinsics of Kinect2 and to obtain the transformations between Kinect2 and the standard camera.
- Evaluation sequences: The sequences can be divided into two types.
  - *Testing and debugging scenes*: sequences intended to facilitate the development of novel multi-spectral algorithms with separated criteria, including motion, illumination, and person. For convenience, these sequences are divided into three categories according to the illumination:
    - ∗ *bright scenes*: captured in a bright room where the color images contain rich textures.
    - ∗ *scenes with varying illumination*: captured in a room with varying illumination, where the quality of the color images is not guaranteed.
    - ∗ *dark scenes*: captured in a dark room. The color images do not provide clear visible textures.
  - *challenging scenes*: sequences captured from more challenging environments, including texture-less, high-dynamic-range, and dark areas. Therefore, these sequences are very challenging for multi-spectral methods. To reduce the difficulty, the IMU information is provided.



(a) Indoor room (b) Indoor stairs (c) Indoor hall

(d) Indoor lift (e) Outdoor road (f) Outdoor building

Fig. 6: Data acquisition environments.

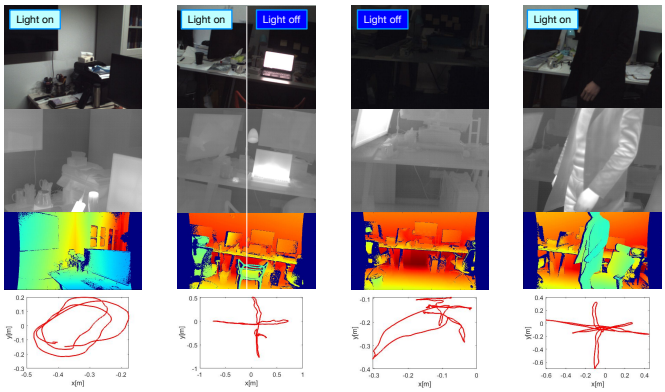
A. Sequences

The dataset was acquired in indoor office environments and outdoors at night, as shown in Fig. 6. The sequences are named according to these criteria: scene, illumination, motion, and the presence of a person in the image.

1) Illumination: Novel methods can be evaluated in environments with different illumination conditions as shown in Fig. 7. In addition to the sequences recorded in bright environments, the sequences with the suffix *-ic* in the name contain data with lights turned on and off randomly. Moreover, for testing the performance of multi-spectral methods in extreme illumination conditions, the sequences with suffix *-dim* contain data in dark scenes. In the challenging sequences, the images are captured under complex illumination conditions.

2) Motion: These sequences with *desk* are intended to evaluate methods with separated motions along and around the principal axes of the setup. There are five types of camera motion:

∗ *-halfsphere (hfsp)* denotes that the camera moves along the trajectory of a halfsphere with a diameter of 1m.



(a) (b) (c) (d)

Fig. 7: Example sequences. The images from the top to the bottom row are color images, thermal images, depth images, and ground-truth trajectories. The color of the depth images indicates the distance to the Kinect2. (a) desk2-circle. (b) desk1-halfsphere-ic. (c) desk1-rpy-dim2. (d) desk1-xyz-dy.

- *-xyz* denotes that the camera moves approximately along the x, y, and z axes.
- *-rpy* denotes that the camera only rotates with roll, pitch, and yaw motion.
- *-circle* denotes that the camera moves approximately around a circle.
- *-static* denotes that the camera is nearly motionless.

3) Person: Since the temperature of the objects in the office is almost uniform, the LWIR camera cannot provide rich textures. Since the temperature of the human body differs significantly from the environment, a person in the FoV of the camera can provide striking contrasts in textures in the thermal images. Hence, in the sequences with the suffix *-person*, there is a person sitting in front of a desk. Otherwise, the environment only contains static objects at room temperature.

4) Dynamic: In the sequences with the suffix *-dy*, a person is walking around in front of the setup to add disturbances. These data can be used to test the robustness of the multi-spectral system in dynamic environments since moving objects in the environment may lead to the failure of motion estimation.

5) Scene: The sequences recorded in the office room are labeled as *desk1* and *desk2*. *magistrale* denotes sequences featuring a walk in a university building. The sequences with the suffix *outdoor* are recorded around a university building. The suffix *lift* indicates that the camera was moved to different floors via elevator.

More information about each sequence can be found on the website.

B. Data format

Each sequence is saved as a ROS bag file.

1) ROS Bag Files: raw data is recorded in the following topics:

- /camera/image_raw: color images from the standard camera
- /optris/thermal_image: thermal images from the LWIR camera
- /camera/flag_state: flag states of the LWIR camera
- /kinect2/qhd/image_color_rect: color images from the Kinect2
- /kinect2/qhd/image_depth_rect: depth images from the Kinect2
- /kinect2/sd/image_ir_rect: NIR images from the Kinect2
- /imu/data: IMU data
- /vrpn_client_node/RigidBody/pose: raw poses from the motion capture system

The meaning of most of these topics is self-explanatory. All data use the time in the ROS system as the timestamp. The /flag_state of NUC is a type *enum* and contains the flag signal of the LWIR camera with the following fields: *FlagOpen, FlagClose, FlagOpening, FlagClosing*, and *Error*. The last topic contains raw poses stored as both a vector and a quaternion from the motion capture system.

VI. EXPERIENCES

This section presents an experimental evaluation with the proposed sequences to show how challenging this dataset can be. Since there is no open-source multi-spectral method, results were obtained from DSO (direct sparse odometry, direct method) [6] and ORB-SLAM3 (feature-based method) [8], which are two state-of-the-art monocular methods. The distinction between those two methods is that feature-based methods use the reprojection error of feature points. In contrast, direct methods exploit the photometric error of raw images directly. Besides, to show the benefit of LWIR, we evaluate both methods with a different spectrum.

As shown in Table III, these two methods show different performance. Since ORB-SLAM3 only uses corner features, the sequences captured in the office without rich textures may not have enough uniform distribution features for ORB-SLAM3 for robust matching. Therefore, the ORB-SLAM3 cannot perform very well in these sequences with good illumination. It also failed in all sequences with varying and dim illumination, although it has a relocalization module. On the other hand, DSO can utilize the edge and the texture in dark environments, showing a better performance in environments with poor illumination.

From the comparison between different spectrum we see that the method using LWIR shows better robustness but worse precision. This is because thermal images are independent from environment illumination but contain more noise from camera self-emission. For this reason, most uncooled LWIR cameras use NUC to eliminate the fixed noise, which may lead to fake tracking for the DSO method. Besides, the ORB feature is not designed for thermal images. Thus, the ORB feature descriptor works poorly.

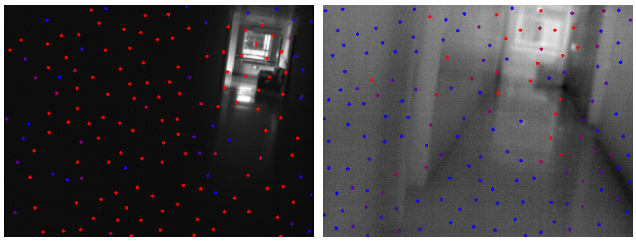
For the sequences with *magistrale* and *outdoor*, both methods failed in tracking, because in these sequences there are scenes with different challenges, such as texture-less, large-scale, low illumination, and high dynamic range scenes. Since these sequences contain the IMU information, the VINS-Mono [45] method was also tested. Similar to the

TABLE III: Comparison of the absolute trajectory error (ATE).

Sequence Name	DSO (RGB)	DSO (LWIR)	ORB-SLAM3 (RGB)	ORB-SLAM3 (LWIR)
Bright illumination				
desk1-halfsphere	0.1114	x	0.0186	x
desk1-rpy	0.0789	x	0.0929	x
desk1-static	x	x	x	x
desk1-xyz	0.0086	0.2050	0.0165	x
desk2-circle	0.0283	x	0.0214	x
desk2-halfsphere	0.0095	0.3106	0.0191	x
desk2-rpy	0.0747	x	0.0174	x
desk2-static	0.0018	x	0.005193	x
desk2-xyz	0.0065	0.005904	0.0098	x
desk1-circle-person	0.0138	x	0.017505	x
desk1-hfsp-person	0.0248	0.2909	0.0123	x
desk1-rpy-person	0.1254	x	0.0336	x
desk1-rpy-person-slow	0.0617	x	0.0550	x
desk1-static-person	x	x	x	x
desk1-xyz-person	0.0056	0.0280	0.0104	0.0484
desk1-halfsphere-dy	0.2553	x	0.1532	x
desk1-rpy-dy	0.0921	x	x	x
desk1-static-dy	x	x	x	x
desk1-xyz-dy	0.2677	x	0.0177	x
Varying illumination				
desk1-rpy-ic	x	x	0.0422	x
desk1-halfsphere-ic	x	0.2472	0.2003	x
desk1-static-ic	x	x	x	x
desk1-static-ic-lampon	x	x	x	x
desk1-xyz-ic	x	0.2858	x	x
desk2-circle-ic	x	x	x	x
desk2-rpy-ic	x	x	x	x
desk2-static-ic	x	0.005904	x	x
desk2-xyz-ic	x	0.2343	x	x
desk1-hfsp-ic-person	x	0.2558	x	x
desk1-rpy-ic-person	x	0.0851	x	x
desk1-static-ic-person	x	x	x	x
desk1-xyz-ic-person	x	x	x	x
Dim illumination				
desk1-halfsphere-dim	x	0.3792	x	x
desk1-halfsphere-dim2	0.3200	0.3730	x	x
desk1-rpy-dim	x	x	x	x
desk1-rpy-dim2	x	x	x	x
desk1-static-dim	x	x	x	x
desk1-static-dim2	x	x	x	x
desk1-xyz-dim	0.0130	0.2449	x	x
desk1-xyz-dim2	0.2838	0.3674	x	x
Complex environments				
magistale-*	x	x	x	x
outdoor-*	x	x	x	x

results of DSO and ORB-SLAM3, VINS cannot complete the entire motion estimation either, as shown in Fig. 8. However, the reasons for the failure of VINS-mono on color and thermal images are different. For color images, it is mainly due to the illumination, which can reduce the reliable features and led to the failure. For thermal images, although minor temperature variation can be captured, the LWIR camera still cannot provide high-quality texture information. As a result, it is difficult for VINS to obtain stable feature points on the thermal images to complete the estimation, even the initialization.

In summary, for this dataset the methods based on the standard camera cannot provide robust estimations when the environment's illumination becomes increasingly complex.



Tracking on color images Tracking on thermal images

Fig. 8: Tracking results from VINS-Mono. The red and blue points represent the reliable and unreliable points considered by VINS.

In contrast, cameras that do not rely on ambient illumination can provide information about the environment to perform localization tasks. However, from the accuracy comparison, we can see that the unique mechanism of the LWIR camera impairs its performance in bright illumination compared with the standard camera. Meanwhile, the failure in all the magistrale and outdoor environments suggests that the only way to achieve robust localization in complex environments is to combine several different sensors. Hence this dataset's challenge increases in the order of bright, dark(desk-dim), varying(desk-ic), and complex(magistrale, outdoor) environments.

VII. CONCLUSIONS

In this paper, a novel dataset was proposed for evaluating the performance of multi-spectral motion estimation methods. The dataset includes sequences captured from different scenes with a diverse set of illumination conditions. In the set-up, the multi-spectral cameras were in hardware-synchronization and well-calibrated. Highly accurate ground-truth poses were provided by a motion capture system for evaluation. Besides, depth images were provided for studying cross-modality stereo matching. Some commonly used robustness metrics were also proposed. This dataset is publicly available with both raw and calibration data. It is hoped that this dataset can facilitate the development of multi-spectral motion estimation.

Moreover, during the preparation of the dataset, some experience was gained. 1. United acquisition program. As far as possible, the drivers of all sensors should be integrated into one program instead of distributed in separate software. Through a unified acquisition procedure, the acquisition efficiency can be improved, while the unpredictable problems that occur during the acquisition process can be reduced. 2. Data check program. After the acquisition, the check program can quickly find whether there are acquisition problems, such as data loss, to improve the acquisition efficiency. We hope that our experience will be helpful to everyone.

REFERENCES

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry [tutorial]," *IEEE robotics & automation magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[2] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

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

[4] 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.

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

[6] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[7] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "Svo: Semidirect visual odometry for monocular and multicamera systems," *IEEE Transactions on Robotics*, vol. 33, no. 2, pp. 249–265, 2017.

[8] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual-inertial and multi-map slam," *arXiv preprint arXiv:2007.11898*, 2020.

[9] A. R. Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 994–1001, 2018.

[10] F. Endres, J. Hess, J. Sturm, D. Cremers, and W. Burgard, "3-d mapping with an rgb-d camera," *IEEE transactions on robotics*, vol. 30, no. 1, pp. 177–187, 2013.

[11] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse, "Monoslam: Real-time single camera slam," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[12] H. Strasdat, J. M. Montiel, and A. J. Davison, "Visual slam: why filter?" *Image and Vision Computing*, vol. 30, no. 2, pp. 65–77, 2012.

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

[14] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[15] E. Benli, R. L. Spidalieri, and Y. Motai, "Thermal multisensor fusion for collaborative robotics," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3784–3795, 2019.

[16] S. Mita, X. Yuquan, K. Ishimaru, and S. Nishino, "Robust 3d perception for any environment and any weather condition using thermal stereo," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2569–2574.

[17] K. Hajebi and J. S. Zelek, "Structure from infrared stereo images," in *Computer and Robot Vision, 2008. CRV'08. Canadian Conference on*. IEEE, 2008, pp. 105–112.

[18] P. V. K. Borges and S. Vidas, "Practical infrared visual odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 8, pp. 2205–2213, 2016.

[19] T. Mouats, N. Aouf, L. Chermak, and M. A. Richardson, "Thermal stereo odometry for uavs," *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6335–6347, 2015.

[20] S. Khattak, C. Papachristos, and K. Alexis, "Keyframe-based direct thermal–inertial odometry," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3563–3569.

[21] C. Papachristos, F. Mascarich, and K. Alexis, "Thermal-inertial localization for autonomous navigation of aerial robots through obscurants," in *2018 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2018, pp. 394–399.

[22] J. Poujol, C. A. Aguilera, E. Danos, B. X. Vintimilla, R. Toledo, and A. D. Sappa, "A visible-thermal fusion based monocular visual odometry," in *Robot 2015: Second Iberian Robotics Conference*. Springer, 2016, pp. 517–528.

[23] S. Zhao, P. Wang, H. Zhang, Z. Fang, and S. Scherer, "Tp-tio: A robust thermal-inertial odometry with deep thermalpoint," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020.

[24] M. R. U. Saputra, P. P. de Gusmao, C. X. Lu, Y. Almalioglu, S. Rosa, C. Chen, J. Wahlström, W. Wang, A. Markham, and N. Trigoni, "Deeptio: A deep thermal-inertial odometry with visual hallucination," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1672–1679, 2020.

[25] T. Mouats, N. Aouf, A. D. Sappa, C. Aguilera, and R. Toledo, "Multispectral stereo odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1210–1224, 2015.

[26] W. Dai, Y. Zhang, D. Sun, N. Hovakimyan, and P. Li, "Multi-spectral visual odometry without explicit stereo matching," in *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019, pp. 443–452.

[27] D. Caruso, J. Engel, and D. Cremers, "Large-scale direct slam for omnidirectional cameras," in *International Conference on Intelligent Robots and Systems (IROS)*, 9 2015.

[28] 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.

[29] M. Ferrera, V. Creuze, J. Moras, and P. Trouvé-Peloux, "Aqualoc: An underwater dataset for visual–inertial–pressure localization," *The International Journal of Robotics Research*, vol. 38, no. 14, pp. 1549–1559, 2019.

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

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

[32] S. Hwang, J. Park, N. Kim, Y. Choi, and I. So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1037–1045.

[33] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," *arXiv preprint arXiv:1909.10980*, 2019.

[34] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "Rgb-t object tracking: benchmark and baseline," *Pattern Recognition*, vol. 96, p. 106977, 2019.

[35] W. Treible, P. Saponaro, S. Sorensen, A. Kolagunda, M. O'Neal, B. Phelan, K. Sherbondy, and C. Kambhamettu, "Cats: A color and thermal stereo benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2961–2969.

[36] W. Maddern and S. Vidas, "Towards robust night and day place recognition using visible and thermal imaging," *RSS 2012: Beyond laser and vision: Alternative sensing techniques for robotic perception*, 2012.

[37] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "Kaist multi-spectral day/night data set for autonomous and assisted driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[38] N. Chebrolu, P. Lottes, A. Schaefer, W. Winterhalter, W. Burgard, and C. Stachniss, "Agricultural robot dataset for plant classification, localization and mapping on sugar beet fields," *The International Journal of Robotics Research*, 2017.

[39] A. J. Lee, Y. Cho, S. Yoon, Y. Shin, and A. Kim, "Vivid: Vision for visibility dataset."

[40] 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)*. IEEE, 2016, pp. 4304–4311.

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

[42] Y. W. K. Zoetgnandé, A.-J. Fougères, G. Cormier, and J.-L. Dillenseger, "Robust low resolution thermal stereo camera calibration," in *Eleventh International Conference on Machine Vision (ICMV 2018)*, vol. 11041. International Society for Optics and Photonics, 2019, p. 110411D.

[43] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "A mask-based approach for the geometric calibration of thermal-infrared cameras," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 6, pp. 1625–1635, 2012.

[44] 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.

[45] 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.