Robust Vision-based Autonomous Navigation, Mapping and Landing for MAVs at Night

Shreyansh Daftry[1], Manash Das[2], Jeff Delaune[1], Cristina Sorice[1], Robert Hewitt[1], Shreetej Reddy[3], Daniel Lytle[4], Elvin Gu[5], and Larry Matthies[1]

[1] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, USA
[2] Indian Institute of Technology, Kharagpur, India
[3] University of Pennsylvania, Philadelphia, USA
[4] University of Southern California, Los Angeles, USA
[5] Carnegie Mellon University, Pittsburgh, USA

Abstract. This paper is about vision-based autonomous flight of MAVs at night. Despite it being dark almost half of the time, most of the work to date has addressed only daytime operations. Enabling autonomous nighttime operation of MAVs with low SWaP on-board sensing capabilities is still an open problem in current robotics research. In this paper, we take a step in this direction and introduce a robust vision-based perception system using thermal-infrared cameras. We present this in the context of safe autonomous landing on rooftop-like structures, and demonstrate the efficacy of our proposed system through extensive real-world flight experiments in outdoor environments at night.

Keywords: Micro Aerial Vehicles, Thermal-infrared, Night Operation, Vision-based Navigation, 3D Mapping, Autonomous Landing

1 Introduction

Micro Aerial Vehicles (MAVs) have built a formidable résumé by making themselves useful in a number of important applications, from disaster scene surveillance and package delivery to robots used in aerial imaging, architecture and construction. As these aerial robots aspire for long-term autonomous operation, the ability to navigate safely becomes critical. While vision-based methods for autonomous navigation of MAVs have been well studied in general [9,7,24], most of these approaches have addressed only daytime operation under normal illumination conditions using standard visible-spectrum cameras. Enabling operations at night would significantly enhance the tactical value of such a system.

In this paper, we propose to use thermal-infrared (TIR) cameras to enable vision-based autonomous operations for MAVs at night. The nature of TIR modality makes them highly robust to low-illumination conditions (including total darkness, as shown in Figure 1) and other environmental effects such as the presence of fog, smoke and dust. Furthermore, COTS, uncooled, TIR cameras based on micro-bolometer focal planes are now small enough to be practical payloads on resource-constrained MAVs.

Fig. 1. An instance of a scene captured with visible-spectrum camera during the day (left), at night (right) and with a TIR camera at night (right). It can be seen that the nature of TIR modality enables the MAV to have more robust night vision.

Despite the aforementioned advantages, these cameras also have several characteristics that are challenging for vision algorithms - limited resolution, low SNR, rolling shutter and motion-blur distortion, etc. Hence, the performance of traditional computer vision techniques developed for electro-optical imagery may not directly translate to the thermal domain. However, to date there have been very few attempts to study the use of thermal-infrared modality for visual navigation and mapping. Even fewer have validated them experimentally in the context of autonomous robotics through real-world experiments. Our goal and primary contribution in this work was to develop a robust end-to-end visual navigation, mapping and autonomous landing system, using a passive monocular thermal-infrared camera as its primary sensing modality.

2 Related Work

Vision in low-illumination conditions. The problem of autonomous navigation and mapping in low-illumination conditions has been widely discussed in the robotics community, in the context of unmanned ground vehicles or automated driving. Most of these approaches have either focused on the use of active sensors such as lidar [1,16], or the use of active illuminators [17,10]. Other approaches [20,18] have used consumer cameras at night, but rely on strong priors from the urban environment. These sensors and methods thus do not scale to MAVs that are constrained to low SWaP and operate in unstructured environments.

Vision with thermal-infrared. In the last few years, there have been several applications of thermal cameras in machine vision - ranging from object recognition and human activity detection [15] to visual odometry and SLAM [25,2]. See [14] for a comprehensive review. In particular, thermal-infrared modality has proven to be very useful in challenging environments, such as in the presence of obscurants [5,21] and darkness [6], where vision-based systems would otherwise fail [4]. The work of [19] is most closely related to ours, where a thermal stereo-pair is used onboard a MAV for visual navigation.

3 Technical Approach

In this section, we describe our proposed approach for robust autonomous flight at night. The primary focus of this paper is going to be the perception system, which is responsible for estimating the 3D scene structure, using a stream of thermal images and the associated pose of the vehicle. Figure 2 provides a schematic overview of the framework. The core components includes state-estimation and monocular dense 3D mapping. In addition, we also discuss the technical approach for our application of safe autonomous landing on rooftop-like structures.

3.1 State Estimation

Our state-estimation pipeline is based on VINS-MONO [23] - a tightly coupled visual-inertial odometry algorithm. It is an optimization-based SLAM framework that takes as input calibrated thermal images and inertial data, and produces high-accuracy pose estimates of the MAV in metric scale. There were various reasons to use this particular algorithm - it runs in real-time onboard the MAV, achieves state-of-the-art performance across several benchmark datasets in the visible domain [8] and has a formulation to compensate for rolling shutter distortions. In our experience, large rolling shutter distortions can be detrimental to VIO performance during high-speed flights. It is to be noted that while we evaluate the performance of our state-estimation pipeline, as discussed in Section 4.2, a comprehensive benchmarking of different VIO algorithms in the thermal-infrared domain is beyond the scope of this paper.

3.2 Monocular Dense Reconstruction.

Structure from Motion (SfM) approaches [26] can be used to reconstruct 3D scene geometry from a moving monocular camera. However, such SfM approaches produce sparse maps that are not suited for collision-free navigation: typically the resulting point cloud of 3D features is highly noisy, and more crucially, the absence of visual features in regions with low texture does not necessarily imply free space. In contrast, dense approaches to 3D reconstruction use camera motion and variable-baseline stereo for depth map estimation. Here, we use a dense approach based on the REMODE (REgularized MOnocular Depth Estimation) algorithm [22]. The motivation behind using it is two fold: First, it performs the depth estimation with direct pixel intensities instead of indirect feature descriptors, making it more robust for low-contrast TIR images. Second, the highly parallelized formulation allows fast onboard computation on a GPU.

In the following we give a brief overview of the REMODE algorithm. The algorithm computes a dense depth map for selected reference views where the depth computation for a single pixel is formalized as a Bayesian estimation problem. A depth-filter is initialized for all pixels in every newly selected reference image $\mathbf{I}_r$ and every subsequent image $\mathbf{I}_k$ is then used to perform a recursive Bayesian update step of the depth estimates. Given a new observation $\{\mathbf{I}_k, \mathbf{T}_{wk}\}$, where $\mathbf{T}_{wr} \in SE(3)$ describe the pose of reference frame $\mathbf{r}$ relative to world frame

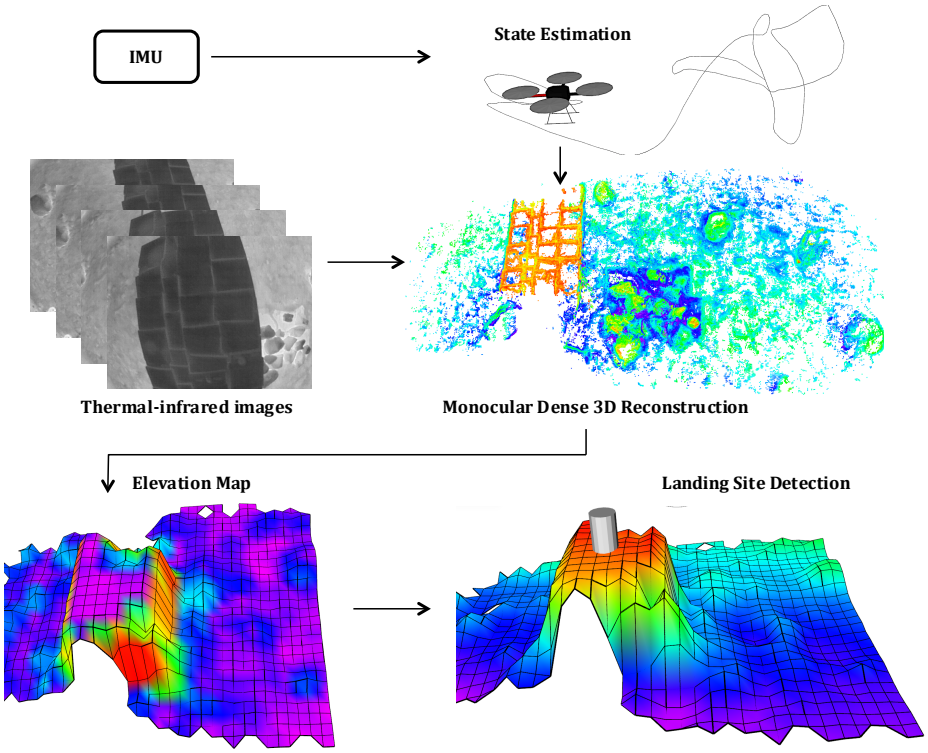

Fig. 2. System overview. The key components of the perception pipeline shown above include visual-inertial state estimation, 3D mapping and safe landing site detection

$\mathbf{w}$, the 95% depth-confidence interval $[d_i^{\min}, d_i^{\max}]$ of the depth filter corresponding to pixel i into the image $\mathbf{I}_k$ and find a segment of the epipolar line l. Using a zero-mean sum of squared differences (ZMSSD) score on a 8×8 patch, we then search the pixel $\mathbf{u}'_i$ on the epipolar line l that has the highest correlation with the reference pixel $\mathbf{u}_i$. A depth measurement d_i^k corresponding to a pixel i in the image $\mathbf{I}_k$ is computed by triangulating the corresponding points $\mathbf{u}_i$ and $\mathbf{u}'_i$ from the views r and k respectively. After the depth map converged, we enforce its smoothness by applying a total-variation based image filter [3].

In practise, both the camera intrinsic matrix and the pose estimates from visual-inertial state estimation are not accurate enough. This results in inaccurate estimation of the fundamental matrix $\mathbf{F}$, and as a result depth-filters do not converge. In contrast, $\mathbf{F}$ derived from only the image correspondences are more accurate as they inherently capture the epipolar geometry. Thus, we use a direct optimization framework based on DSO [11], that computes the relative camera pose between two observations $\mathbf{T}_{rk}$, such that the intensity difference of

the corresponding points $\mathbf{u}_i$ and $\mathbf{u}'_i$ is minimum using the cost function E_i

$$E_i = \sum_{i \in \mathcal{N}_p} \left\| \mathbf{I}_r[\mathbf{u}_i] - \mathbf{I}_k[\mathbf{u}'_i] \right\|_\gamma \tag{1}$$

$$\tag{2}$$

To compensate for automatic gain changes between frames, an affine brightness transfer function given by $e^{-a_k}(\mathbf{I}_k - b_k)$ is added to Equation 1. Thus, resulting in the following cost function:

$$E_i = \sum_{i \in \mathcal{N}_p} w_{\mathbf{u}_i} \left\| (\mathbf{I}_r[\mathbf{u}_i] - b_r) - \frac{e^{a_r}}{e^{a_r}} (\mathbf{I}_k[\mathbf{u}'_i] - b_k) \right\|_\gamma \tag{3}$$

where $w_{\mathbf{u}_i}$ is a gradient-dependent weighting given by:

$$w_{\mathbf{u}_i} = \frac{c^2}{c^2 + \left\| \nabla \mathbf{I}(\mathbf{u}) \right\|_2^2} \tag{4}$$

The relative camera pose $\mathbf{T}_{rk}$ here is initialized using the pose from the state-estimation pipeline. Given a depth measurement per pixel, we simply use the camera model to obtain the corresponding 3D point.

3.3 Vision-based Safe Autonomous Landing

While the above described perception pipeline is applicable for vision-based navigation and obstacle avoidance in general, we study this specifically in the context of autonomous landing. To land safely and autonomously on different structures is an essential capability for many scenarios - perch-and-stare missions, to conserve or recharge batteries, etc. At minimum, this requires an ability to find safe landing spots through a semantic interpretation of the environment.

The first step towards this is to generate a 2.5D interpretation of the terrain. We use a recently developed robot-centric elevation mapping framework [12]. The map is stored as a 2-dimensional grid, where the height in each cell is modeled as a normal distribution. As the robot moves and generates instantaneous point clouds, the height estimates for each cell is updated using a recursive Bayesian update step. The goal of the original work was to develop a local map representation that serves foot-step planning for walking robots over and around obstacles. However, we find that the local two-dimensional elevation map is an efficient on-board map representation for MAVs [13] that are flying outdoors - it allows us to keep a safe distance to the surface and to detect and approach suitable landing spots. By tightly coupling the local map to the robots pose, the framework can efficiently deal with drift in the pose estimate.

Once we have the elevation map, we are now in a position to determine the criteria for safe landing spots. We define a suitable landing site as: (1) approximately planar and level, (2) sufficiently large to permit MAV ingress, landing, and egress; and (3) free of obstacles to ensure vehicle safety. A cost function is defined using the above, and used to reason about a suitable landing site; then, we use motion planning and control to navigate to the desired location.

Fig. 3. (a) Quadrotor platform, (b) Test setup during the day, and (c) at night.

4 Experiments and Results

In this section we analyze the qualitative and quantitative performance of our proposed system on publicly available benchmark datasets, and through real-world flight experiments on a MAV at night.

4.1 Experimental Setup

Hardware Platform. We use a Luminier QAV400 quad-rotor platform equipped with a down-ward facing thermal-infrared camera (We do experiments with both FLIR Boson and A65) and a MPU-9250 IMU, as shown in Figure 3(a). Our flight computer is a Nvidia Jetson TX2 board, which incorporates a quad-core ARM processor and an embedded GPU with 256 CUDA cores.

Mission Scenario. In all our tests, we consider the perch-and-stare scenario, where the quad-rotor is tasked to explore a given area using a set of pre-defined waypoints, find a suitable elevated platform for surveillance and autonomously land on it. We created an analog test setup at JPL's Marsyard with artificial obstacles and a landing platform, as shown in Figure 3.

System Overview. The software perception pipeline runs onboard the MAV. We use the ground-station only for communication and mission-level task planning. On average, the quadrotor was flying $2-3$ meters above the surface and used an elevation map of size 5x5 meters, with a resolution of 10 centimeters per cell. During each flight, the quadrotor was restrained using a light-weight tether. Its to be noted that the tether is only for compliance to federal regulations and does not limit the feasibility of a free flight.

4.2 Performance Evaluation

Several flight tests were conducted at different times of the night - ranging from dusk to midnight, to evaluate the robustness of the proposed method across varying illumination, thermal conditions and flight speed. Figure 4 shows the system performance during one of the flight sequence. From takeoff to landing,

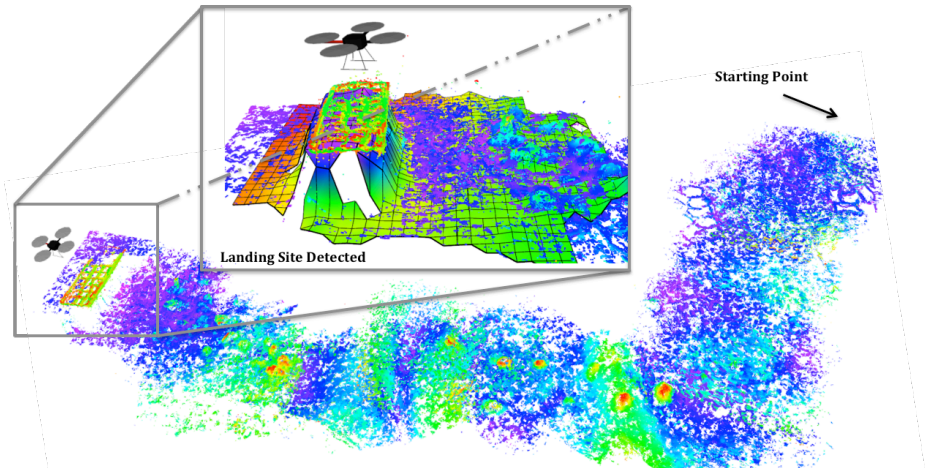

Fig. 4. Qualitative results from a flight test showing the 3D reconstruction of the scene (textured with thermal intensity), elevation map and the detected landing site.

the state-estimation module was able to successfully track the pose of the vehicle. Furthermore, we visualize the 3D reconstruction of the scene generated during the flight. Note: The point cloud is textured using the thermal intensity images. It can be seen that even though the quadrotor was flying in total darkness, over a terrain with low texture, it is still able to accurately reconstruct the scene. This reconstruction is both dense and semantically rich; several important properties regarding the terrain can easily be inferred - the shape and size of rocks, thermal properties of the different terrains such as bedrock vs. sand pits, etc.

Overall, the MAV was able to complete each mission successfully - navigate, map the environment and find a suitable landing site. Furthermore, to quantitatively evaluate the performance of our system, we setup an outdoor motion-capture arena to obtain ground-truth for all our flight experiments. The average position drift of the pose estimation w.r.t ground truth in x, y and z axis was $0.1112m$, $0.02556m$ and $0.04978m$ respectively, which is around 1% of the distance traveled. Similarly, an average orientation drift of 3.171 deg was observed across all the flight tests.

Computationally, the state estimation pipeline runs at 30 Hz and consumes 2 cores on CPU, while the monocular dense reconstruction runs at 1 Hz and uses about 35% of the GPU, on average. The elevation mapping and landing package uses another core, and the fourth core is reserved for the camera driver, communication, and control. It was observed that an update rate of 1 Hz is sufficient to always maintain a dense elevation map below the MAV. However, when moving in a straight line, the local elevation map behind the MAV is more populated than in the front. In the future, we will modify the MAV to have a slightly forward facing camera or consider using a multi-camera setup in order to have better distribution.

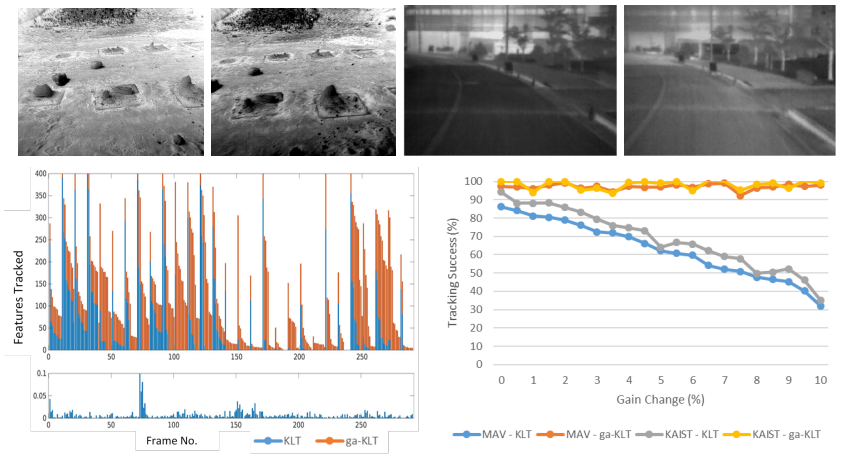

Fig. 5. Analysis of degraded performance due to AGC. (Top) Nearby frames from an image sequence illustrating the significant change in appearance as a result of AGC. (Bottom-Left) Features tracked and the corresponding gain change for one of the daytime flight sequence, (Bottom-Right) Tracking success as a function of gain-change.

4.3 Experimental Insights

A key insight obtained in this work is that applying vision-based techniques developed for visible-spectrum directly to thermal-infrared domain requires careful design considerations to account for the nuances of thermal-imaging. In particular, accounting for photometric effects like gain-change and rolling-shutter distortion can improve the robustness of perception tasks.

Automatic Gain Change. In the context of vision-based methods with thermal-images, one of the most important challenge we identified was with respect to rapid automatic gain change (AGC). As hot objects move in and out of the TIR camera's field of view, large regions of pixels saturate, and the sensor adjusts gain to darken the image in an attempt to avoid saturation. The resulting change in intensity can be dramatic from one frame to the next (See Figure 5), invalidating the constant brightness assumption enforced in most vision-front end solutions. To quantify the effect of such rapid gain, we compare the feature tracking performance of standard KLT tracker to a gain-adaptive version (using the open-source code from [27]). Note: These experiments were conducted with thermal images from the FLIR A65 camera; the effects of AGC in FLIR Boson camera was minimal and hence we did not see similar improvements in tracking performance by compensating for AGC change.

Rolling shutter Distortion. We qualitatively evaluated our state-estimation pipeline with- and without rolling shutter compensation. It was observed that while this can be neglected for MAV flights in slow dynamics, rolling distortions can be significant during aggressive motions; leading to state-estimation failure. In future work, we will also extend our dense reconstruction pipeline to compensate for rolling shutter effects.

5 Conclusion

In this paper, we take a step towards enabling autonomous MAV flight at night. We present a real-time perception system that employs thermal-infrared cameras for sensing, models its nuances during algorithm design and selection, and is able to perform robust state-estimation and monocular dense 3D mapping. Quantitatively and qualitatively, we show improvements in performance over simply re-using visible-spectrum algorithms, through extensive real-world experiments. We have demonstrated our system in the context of safe autonomous landing on rooftop-like structures. However, our method and findings apply to any aspect of experimental robotics where an autonomous system has to operate at night.

6 Acknowledgement

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

References

1. Barfoot, T.D., McManus, C., Anderson, S., Dong, H., Beerepoot, E., Tong, C.H., Furgale, P., Gammell, J.D., Enright, J.: Into darkness: Visual navigation based on a lidar-intensity-image pipeline. In: Robotics Research, pp. 487–504. Springer (2016)
2. Borges, P.V.K., Vidas, S.: Practical infrared visual odometry. IEEE Transactions on Intelligent Transportation Systems **17**(8), 2205–2213 (2016)
3. Bresson, X., Esedolu, S., Vandergheynst, P., Thiran, J.P., Osher, S.: Fast global minimization of the active contour/snake model. Journal of Mathematical Imaging and vision **28**(2), 151–167 (2007)
4. Brunner, C., Peynot, T., Vidal-Calleja, T., Underwood, J.: Selective combination of visual and thermal imaging for resilient localization in adverse conditions: Day and night, smoke and fire. Journal of Field Robotics **30**(4), 641–666 (2013)
5. Burian, F., Kocmanova, P., Zalud, L.: Robot mapping with range camera, ccd cameras and thermal imagers. In: Methods and Models in Automation and Robotics (MMAR), 2014 19th International Conference On, pp. 200–205 (2014)
6. Chen, L., Sun, L., Yang, T., Fan, L., Huang, K., Xuanyuan, Z.: Rgb-t slam: A flexible slam framework by combining appearance and thermal information. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on, pp. 5682–5687 (2017)
7. Daftry, S., Zeng, S., Khan, A., Dey, D., Melik-Barkhudarov, N., Bagnell, J.A., Hebert, M.: Robust monocular flight in cluttered outdoor environments. arXiv preprint arXiv:1604.04779 (2016)
8. Delmerico, J., Scaramuzza, D.: A benchmark comparison of monocular visual-inertial odometry algorithms for flying robots. Memory **10**, 20 (2018)
9. Dey, D., Shankar, K.S., Zeng, S., Mehta, R., Agcayazi, M.T., Eriksen, C., Daftry, S., Hebert, M., Bagnell, J.A.: Vision and learning for deliberative monocular cluttered flight. In: In Proceedings of the International Conference on Field and Service Robotics (FSR) (2015)

10. Dubbelman, G., van der Mark, W., van den Heuvel, J.C., Groen, F.C.: Obstacle detection during day and night conditions using stereo vision. In: Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on, pp. 109–116 (2007)
11. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE transactions on pattern analysis and machine intelligence **40**(3), 611–625 (2018)
12. Fankhauser, P., Bloesch, M., Gehring, C., Hutter, M., Siegwart, R.: Robot-centric elevation mapping with uncertainty estimates. In: Mobile Service Robotics, pp. 433–440. World Scientific (2014)
13. Forster, C., Faessler, M., Fontana, F., Werlberger, M., Scaramuzza, D.: Continuous on-board monocular-vision-based elevation mapping applied to autonomous landing of micro aerial vehicles. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on, pp. 111–118 (2015)
14. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Machine vision and applications **25**(1), 245–262 (2014)
15. Han, J., Bhanu, B.: Human activity recognition in thermal infrared imagery. In: Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on, pp. 17–17 (2005)
16. Husain, A., Jones, H., Kannan, B., Wong, U., Pimentel, T., Tang, S., Daftry, S., Huber, S., Whittaker, W.L.: Mapping planetary caves with an autonomous, heterogeneous robot team. In: Aerospace Conference, 2013 IEEE, pp. 1–13 (2013)
17. Mascarich, F., Khattak, S., Papachristos, C., Alexis, K.: A multi-modal mapping unit for autonomous exploration and mapping of underground tunnels. In: 2018 IEEE Aerospace Conference, pp. 1–7 (2018)
18. Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on, pp. 1643–1649. IEEE (2012)
19. Mouats, T., Aouf, N., Chermak, L., Richardson, M.A.: Thermal stereo odometry for uavs. IEEE Sensors Journal **15**(11), 6335–6347 (2015)
20. Nelson, P., Churchill, W., Posner, I., Newman, P.: From dusk till dawn: Localisation at night using artificial light sources. In: Robotics and Automation (ICRA), 2015 IEEE International Conference on, pp. 5245–5252 (2015)
21. Papachristos, C., Mascarich, F., Alexis, K.: Thermal-inertial localization for autonomous navigation of aerial robots through obscurants. In: 2018 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 394–399 (2018)
22. Pizzoli, M., Forster, C., Scaramuzza, D.: Remode: Probabilistic, monocular dense reconstruction in real time. In: Robotics and Automation (ICRA), 2014 IEEE International Conference on, pp. 2609–2616. IEEE (2014)
23. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. arXiv preprint arXiv:1708.03852 (2017)
24. Shen, S., Michael, N., Kumar, V.: Autonomous indoor 3d exploration with a micro-aerial vehicle. In: Robotics and Automation (ICRA), 2012 IEEE International Conference on, pp. 9–15. IEEE (2012)
25. Vidas, S., Sridharan, S.: Hand-held monocular slam in thermal-infrared. In: Control Automation Robotics & Vision (ICARCV), 2012 12th International Conference on, pp. 859–864 (2012)
26. Wu, C.: Towards linear-time incremental structure from motion. In: 3D Vision-3DV 2013, 2013 International Conference on, pp. 127–134. IEEE (2013)
27. Zach, C., Gallup, D., Frahm, J.M.: Fast gain-adaptive klt tracking on the gpu. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp. 1–7. IEEE (2008)