2017 IEEE International Conference on Robotics and Automation (ICRA)
Singapore, May 29 - June 3, 2017

RGB-T SLAM: A Flexible SLAM Framework By Combining Appearance and Thermal Information

Long Chen, Libo Sun, Teng Yang, Lei Fan, Kai Huang and Zhe Xuanyuan

***Abstract*— Visual SLAM in low illumination scenes remains a considerably challenging task since the available amount of appearance information frequently stays insufficient. To tackle with this problem, we propose a novel SLAM framework by using both appearance information and thermal information, which possesses illumination-free recognizable contents, in a flexible manner. The key idea is to continuously update a RGB-T map, which contains both RGB and thermal map points to implement location and mapping. More specifically, in our SLAM system, we detect features in both RGB and thermal images and combine them together to update the RGB-T map and implement simultaneous location and mapping. Both quantitative and qualitative results demonstrate the effectiveness of our framework, especially under low illumination environments.**

I. INTRODUCTION

Vision-based Simultaneous Location and Mapping (SLAM) is becoming increasingly important for robotics, since it only needs low cost optical cameras. In recent years, a lot of researches, such as [1, 2], have proposed to build different SLAM systems in different environments. Depending on the types of camera sensors, current vision-based SLAM methods can be divided into binocular SLAM and monocular SLAM. In this paper, our works are based on monocular SLAM.

Most of the current existing SLAM approaches(including binocular and monocular) are based on feature points and the required geometrical information is computed by using the correlation matching between these feature points. These feature-based works can be found in a keyframe-based nonlinear optimization framework [3] or a filter-based framework [4]. Recently, direct approaches have become popular, such as [5]. They compute depth maps in an incremental fashion, and track camera poses by using direct image alignment.

However, all of these visual methods, no matter feature-based or direct, have a common drawback - they can work only in normal illumination environments. When facing a low illumination environment, these current state-of-the-art visual methods will lose their efficacy, as they can not obtain enough RGB features or textures.

For low illumination environments where not enough RGB information can be obtained, thermal sensors can be a good information source. This thermal information, which does not directly depend on the intensity of illumination, provides a useful approach to implement location and reconstruction. In [6], a infrared sensor based SLAM was proposed, however, it only uses thermal information and does not consider objects which do not have obvious thermal textures, making the system lose generality.

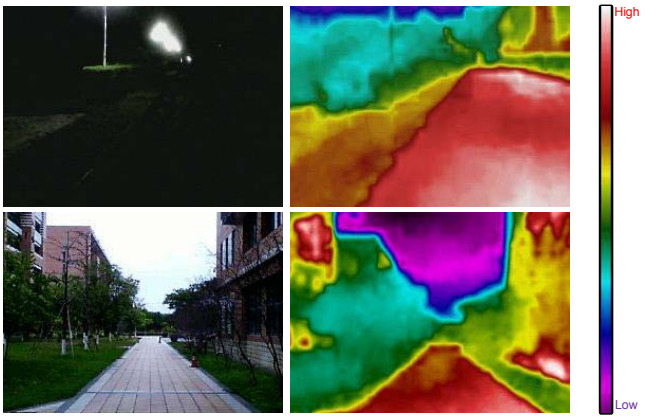


Fig. 1. RGB-T SLAM: Our approach combines environments RGB and thermal information together to build an accurate and robust SLAM system in both low illumination and general environments. Top: RGB image and thermal image in low illumination environment (night). Bottom: RGB and thermal image in general environment (daytime).

Unlike optical cameras which can provide RGB images only, current two-mode cameras can provide both RGB image and thermal image for the same scene. Pixels in RGB and thermal images have a direct correspondence, which is corrected by the hardware of the cameras. Being aware of this, we propose a RGB-T SLAM, which is based on feature-based ORB-SLAM [7] framework. By fusing RGB and thermal information together, the proposed RGB-T SLAM exhibit a much more robust and accurate performance than any other visual methods in the literature. Our main contributions are:

- We propose a novel SLAM framework by using both the appearance information and the thermal information in a flexible way. Be different from most existing visual SLAM methods, which rely heavily on illumination conditions, our RGB-T SLAM is suitable for much more wide scenarios including different illumination environments.
- We propose a buffer-based initialization method with dual superiorities in speed and effectiveness, comparing with state-of-the-art SLAM system, such as [5, 7]. The key insight is that instead of using only one previous frame as reference frame, the current frame will keep

L. Sun, L. Chen, T. Yang, L. Fan and K. Huang are with the School of Data and Computer Science, Sun Yat-sen University, China.
Z. Xuanyuan is with the Beijing Normal University - Hong Kong Baptist University United International College, China.

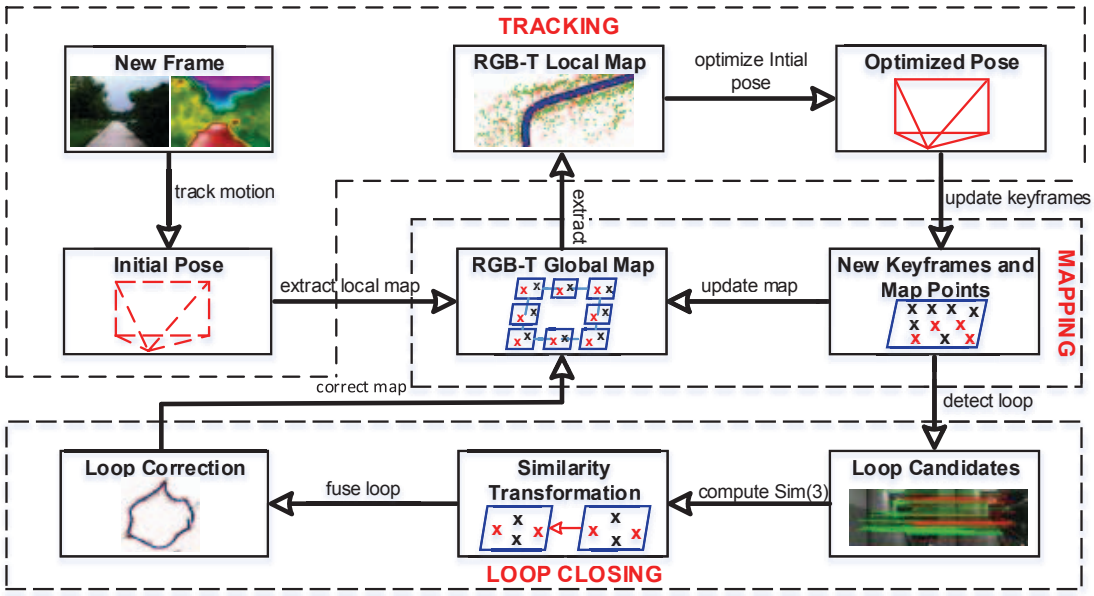


Fig. 2. RGB-T SLAM system overview.

searching its reference frame in a buffer to attempt initializing, until the initialization is succeed.

- We combine RGB and thermal information together in feature matching, which can remedy the poor feature matching performances of descriptor only methods.

The remainder of this paper is organized as follows: Section II introduces some recent works. Section III describes some fundamental knowledge about RGB-T SLAM. Section IV presents how new frames are tracked by using RGB-T map. In Section V and Section VI, we describe mapping and loop closing separately. Finally, in Section VII, we present an evaluation of our system, and give a brief conclusion in Section VIII.

II. Related Work

A. Vision-based SLAM

In recent years, camera sensor based SLAM are widely used in SLAM systems. Depending on the types of sensors, These SLAM systems can be divided into - Monocular SLAM, Stereo SLAM and RGB-D SLAM.

Compared with monocular and stereo cameras, RGB-D cameras need not compute pixels depth, as they can obtain each pixel's depth directly. This benefit provides a convenient way to build a SLAM system. In [8], a RGB-D SLAM system was proposed, which can use a RGB-D camera to create accurate 3D point clouds in indoor environment. However, as RGB-D sensors have a very limited effective distance, these RGB-D camera based SLAM systems can not work normally in large scale outdoor environments, obviously restrict their range of application.

Stereo cameras could compute depth directly and guarantee reliable accuracy in several meters range, which make them widely used in current SLAM system. In [9], they proposed a representative stereo SLAM, in which they use the fixed base line to compute pixels depth and these obtained depth information will be used for location and reconstruction. However, although the fixed baseline between two cameras in stereo SLAM system provide a convenient way to compute depth and estimate camera pose, it has a limited range at which they can provide reliable measurements. When facing a large scale environment, the stereo SLAM system will lose its efficacy.

Monocular SLAM, a hot research topic in recent SLAM works [3, 5, 7], adapts only one camera to implement location and reconstruction. Compared with RGB-D SLAM and stereo SLAM, monocular SLAM gets rid of the scale limitation, which enables it to seamlessly switch between differently scaled environments. This benefit makes monocular SLAM keep its flexibility in different scene, but it reduces the robustness of SLAM system, since it can hardly obtain enough information to support the SLAM system when facing severe rotation and translation.

B. The Application of Thermal Information

The thermal-infrared camera which receives surface radiation from environments can then estimate the temperature by Stefan-Boltzmann Equation accurately. Thermal-infrared cameras have been widely researched in areas, such as 3D temperature mapping [10, 11] and monocular SLAM [6].

3D mapping of the objects, which adapts temperature information to color the surface of a 3D model, can help us to monitor the condition of targets. To achieve this goal, the method in [11] combines Kinect and a thermal-infrared camera to map while estimating the pose of depth camera.

In [6], Stephan Vidas et al. proposed a method using monocular thermal-infrared camera, which enables robotics to locate themselves when common vision sensors per-

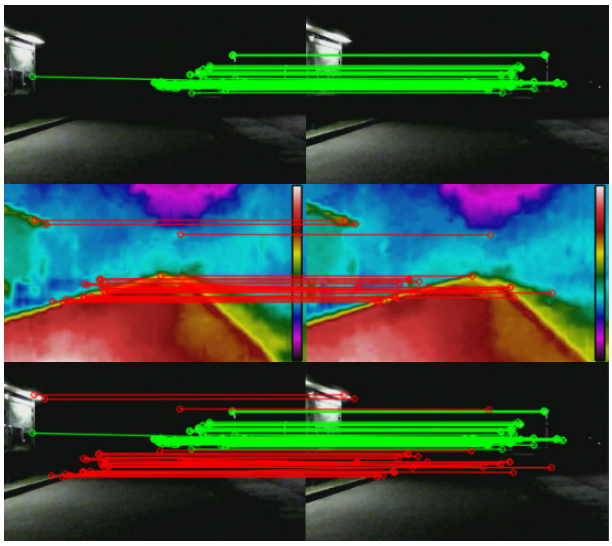


Fig. 3. Feature Combination Example: Two images' features are combined in RGB image. Top: Matched features in RGB image. Middle: Matched features in thermal image. Bottom: Features' combination result. Note that some features which can't be detected and matched in the dark area of RGB image can be detect and matched in thermal image.

form poorly or fail. They find that feature detector can be performed on the thermal map once lowering the sensitivity threshold, and then they utilize GFTT [12] and FAST [13] detectors for their approach. During tracking, they derive Lucas-Kanade [14] optical flow method, but when the thermal-infrared camera corrects the non-uniformity and an apparent delay incurs, SURF [15] features are extracted between recent consecutive frames instead.

However, thermal-infrared cameras also have some drawbacks, such as poor texture. To supplement these drawbacks a regular camera which accepts visible-spectrum can be combined to guide and map for robotics simultaneously.

III. System Overview

Our system (see Fig. 2), can be divided into three main components: Tracking, Mapping and Loop Closing. These three components accomplish new camera pose tracking, sparse point map reconstruction and loop detection and optimization respectively.

A. Feature Detection and Matching

For feature detection, we choose to adopt Fast [13] corner detector to extract features in both RGB image and thermal image, while different thresholds are adopted to meet textures' difference in two images. However, compared with using single ORB descriptor [16] to obtain feature matching results, We combine RGB and thermal information together in a refinement to remedy the poor matching performance of descriptor only method.

During the process of refinement, features and their surroundings' RGB and thermal information are considered. If the RGB-T difference between two features exceed a threshold, they will be culled. The refinement difference function is defined as

$$Dist_{i,j} = \alpha \sum_{n=1}^{R_g} (G_i - G_j) + \beta \sum_{n=1}^{R_t} (T_i - T_j) \qquad (1)$$

Where α and β are two constants, which represent RGB and thermal information's proportion. R_g and R_t are two radius in which RGB and thermal information are considered. G_i and G_j are RGB channels' mixed values of two features and T_i and T_j represent two features' thermal values.

B. RGB-T Information Fusion

We introduce a brief design in which we effectively solve the fusion problem of RGB and thermal information. Similar to RGB-D, we can measure each RGB image pixel's thermal information and obtain these thermal information associated thermal image at the same time. It means that we can get two images for the same scene, and each pixel in RGB image, it has a directly corresponding pixel which locates at the same position in thermal image.

This fusion combines features together, which enables our system to obtain more robust performances in different environments, since much more information can be obtained. A fusion result is shown in Fig. 3.

C. RGB-T Map Initialization

To start the system, a buffer based small range movements are performed to implement the initialization.

In the process of initialization, when a new frame F_i comes, we try to find its reference frame F_j to compose two potential initialization frames to achieve a successful initialization. If a successful initialization is found, the new frame and its reference frame will be created as first two keyframes. Otherwise, the initialization buffer will be updated using the new frame. An example of two initialization frames' selection is shown in Fig. 4.

Only enough matches are found between F_i and F_j, can they be chosen to perform initialization process. The initialization process can be decoupled into three main steps:

1) Compute a fundamental matrix F_{ij} and a homography H_{ij} between two frames:

$$(X_i)^T F_{ij} X_j = 0 \qquad H_{ij} X_j = X_i \qquad (2)$$

Where $X_i \leftrightarrow X_j$ is the matches between F_i and F_j. To select a specific assumption model, a score S_F for fundamental matrix and a score S_H for homography are computed. If not enough inliers are found in two models, the process will be reset and a new reference frame will be chosen.

2) Assess scores of two models and select a specific model:

$$P_F = S_F / (S_F + S_H) \qquad (3)$$

Where P_F is the assessment percentage of fundamental matrix. If P_F is above a certain threshold, the fundamental matrix model will be selected. Otherwise, the homography will be selected.

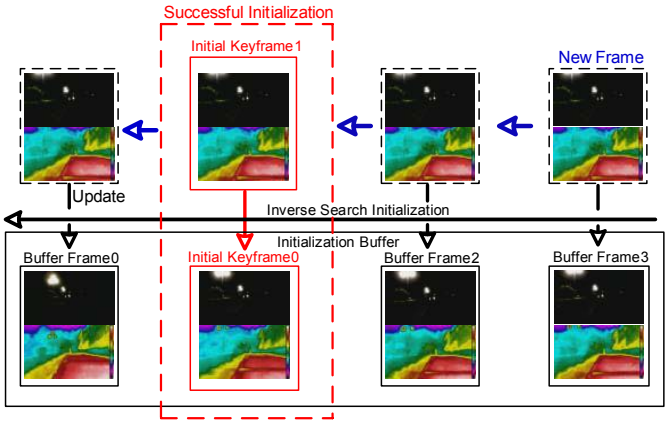


Fig. 4. Initial Keyframe Creation: When a new frame comes, it will be searched in the initialization buffer to try to find a successful initialization. If a successful initialization can not be found, the existing buffer will be updated by the new frame.

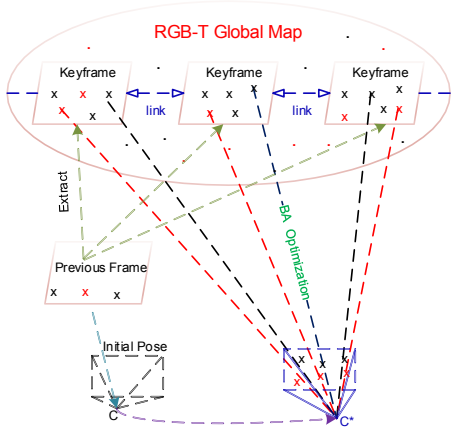


Fig. 5. Initial Pose Estimation: A velocity model is used for giving a predicted camera position and a motion-based BA is performed to get initial Pose.

3) Recover motion and structure from the selected model to find a reconstruction, and refine the reconstruction using the method of Bundle Adjustment [17].

IV. Tracking

The tracking component is performed to obtain camera poses when new frames come. It is divided into two separated steps - initial pose estimation and RGB-T map based pose optimization.

A. Initial Pose Estimation

We introduce a motion model combined with a motion-based BA to obtain initial camera poses. The description of this estimation is shown in Fig. 5. The velocity of previous motion is used to give a predicted pose of current frame at first. If we get a predicted camera pose, the map points contained in previous frame F_j will be searched around their predicted position in current frame F_i by using a projection search.

During projection search, features in both RGB and thermal image are searched and matched separately at first. These matched features are then combined in the fusion process, which is described in Section III-B.

Once enough matches are obtained, initial current pose will be obtained by using a motion-based Bundle Adjustment. During the process of motion-based BA, positions of matched features in previous frame coordinate must be transformed into current frame coordinate. The transformation function is defined as

$$(x_{i,j}, y_{i,j}, z_{i,j})^T = R_{ij} X_j + t_{ij} \tag{4}$$

Where X_j represents points' 3D position in previous frame F_j. $(x_{i,j}, y_{i,j}, z_{i,j})^T$ is the predicted 3D position of X_j in current frame F_i. $R_{i,w}$ and $t_{i,w}$ are rotation and translation matrix, which come from the velocity of previous motion.

Features' position observation error $e_{i,j}$ between F_i and F_j can be defined as

$$e_{i,j} = \lambda_{ij} - \lambda^*(x_{i,j}, y_{i,j}, z_{i,j}) \tag{5}$$

Where λ_{ij} is features' 2D position in F_i. λ^* is the the projection function, which is given by

$$\lambda^*(x_{i,j}, y_{i,j}, z_{i,j}) = \begin{bmatrix} \frac{x_{i,j}}{z_{i,j}} f_x + c_u \\ \frac{y_{i,j}}{z_{i,j}} f_y + c_v \end{bmatrix} \tag{6}$$

Where $(f_x,\ f_y)$ and $(c_u,\ c_v)$ represent cameras' focal length and principle point location.

The motion-based BA will be performed by minimizing error function. The error function is given by

$$E = \sum_{i,j} h_c(e_{i,j}^T \Lambda_{i,j}^{-1} e_{i,j}) \tag{7}$$

where h_c is the Huber robust cost function. $\Lambda_{i,j}^{-1}$ is the covariance matrix.

B. RGB-T Map Based Pose Optimization

After we get an estimated initial camera pose, we extract a set of map points from the RGB-T global map and project these included map points into current frame. The map point set is extracted as follows:

1) Find all the keyframes which share common map points with current frame. Compose these keyframes as a keyframe set K_s.
2) Include all the keyframes which have a neighbor in K_s. The new keyframe set is extended to $K_s{}^*$.
3) Include all the map point in $K_s{}^*$. The obtained map points set is defined as M_s.

All the map points in M_s will be projected and searched in current frame. Finally, these searched map points will be used for performing BA again to get the optimized camera pose.

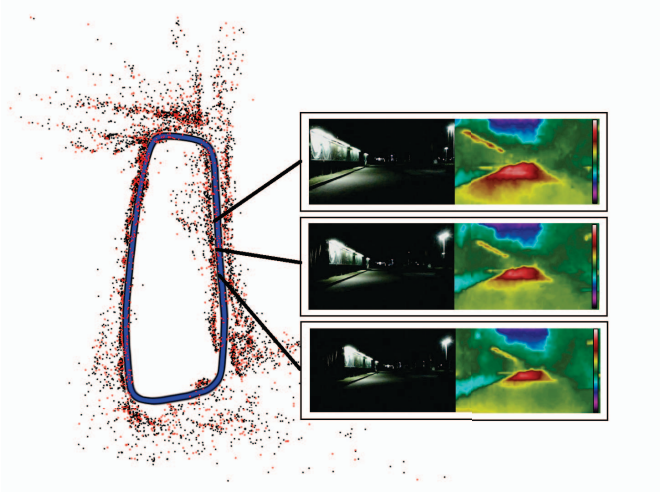


Fig. 6. Trajectory and Point Cloud: Blue line is trajectory. Black and red are RGB and thermal map points.

V. Mapping

The mapping thread is performed to update keyframes and map points in global map. These updated keyframes and map points will be used for tracking and loop closing.

A. Keyframe Update

When a new frame is successfully tracked, it will be treated as a potential keyframe and a keyframe selection function will be performed to decide whether it should be created as a keyframe. If a new frame is selected as keyframe, it will be added to keyframe database and its links between other keyframes will also be updated.

In order to keep a compact reconstruction, the mapping thread tries to cull redundant keyframes. Those keyframes whose 85% of the map points can be seen in at least other three keyframes will be culled.

B. Map Point Update

To maintain a succinct but effective global map point set, only these features which are matched in keyframes will be chosen to create as map points.

After a new keyframe is added to keyframe database, we extract it linked other keyframes from the database to compose a keyframe set KF. For each keyframe KF_j in KF, we try to find matched features between KF_j and current new keyframe K_i. These matched features will be used to create new map points.

Similar with the culling of keyframes, we also try to cull the map points judged in bad condition. After new map points are created, their visibility among keyframes will be checked to find and cull bad map points.

VI. LOOP CLOSING

We introduce a loop closing component to detect and close loops. After new keyframes are created, the loop closing thread tries to find loops and perform optimization.

A. Loop Detection

Compared with these vision-based SLAM systems which consider environments' RGB information only when detect loops, we introduce environments thermal information into loop detection. A DBoW2 [18] based bags of words place recognition is introduced in our system to find loop candidates. During loop detection, both environment's RGB and thermal similarity are considered. The similarity function is defined as

$$S_{total} = \alpha S_{rgb} + \beta S_t \tag{8}$$

Where S_{rgb} and S_t are the bag of words vector similarity of RGB vector and thermal vector. α and β are two constants, which depend on environments' RGB and thermal information.

B. Loop Correction

To close a loop, we first compute the similarity transformation between current keyframe K_i and loop candidate keyframe K_c, and after we get the similarity transformation, we begin to correct the loop.

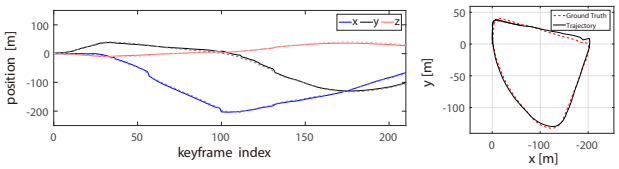
During the process of loop correction, the current keyfame pose is corrected with the similarity transformation S_{ic} at first and then this correction will be propagated to all of the current keyframe's neighbors. To effectively close the loop, a pose graph optimization will be performed to distribute the loop closing error.

VII. Experiments

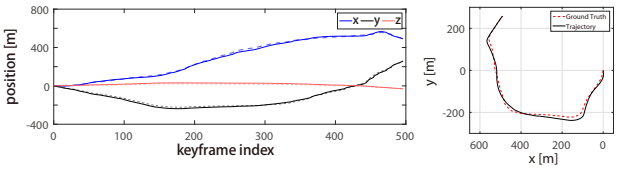
To give a thorough evaluation of the propose RGB-T SLAM, we have performed extensive experiments, including different illumination and dimension environments in six datasets. The experimental datasets are produced by a camera and a RTK-GPS. The camera is used for producing RGB images and thermal images, and the GPS which can achieve cm-level accuracy is adopted to provide ground truth information. A location and mapping example of RGB-T SLAM is shown in Fig. 6. The qualitative comparisons of our trajectories and ground truth are shown in Fig. 7 .

We also compare our RGB-T SLAM with three other advanced vision-based monocular SLAM system in our eight datasets. For these vision-based SLAM, only RGB images are used. In order to compare RGB-T SLAM, ORB-SLAM [7], LSD-SLAM [5] and PTAM [3] with the ground truth, we align the keyframe trajectories using a similarity transformation, as scale is unknown, and measure the absolute trajectory error.

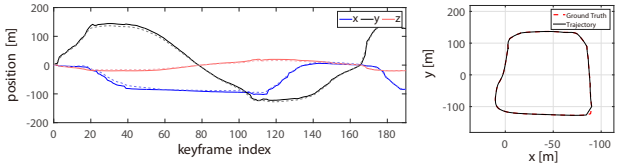
A detailed comparison result is shown in Table I. As can be seen from the table, in general environments (abundant illumination), the proposed RGB-T SLAM have almost equal accuracy performance with current state-of-the-art methods. However, as for low illumination environments, in which these current vision-based SLAM lose their abilities, our system can still keep its robustness and accuracy.



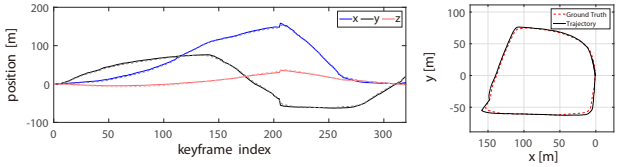
(b) Sequence 01: Dormitory block in the afternoon.



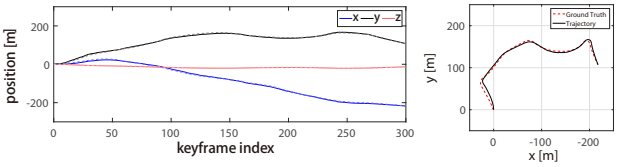
(c) Sequence 02: School campus in the afternoon.



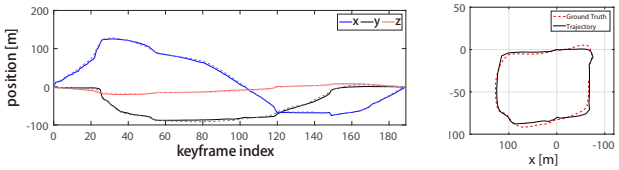
(d) Sequence 03: Transformer substation in the morning.



(e) Sequence 04: Gymnasium in the late afternoon.



(f) Sequence 05: Dormitory block at night.



(g)Sequence 06: Transformer substation at night

Fig. 7. Trajectory Comparison: Estimated camera trajectory and ground truth (dashed) for sequences 01 - 06.

VIII. Conclusion

In this paper we proposed a novel and robust RGB-T SLAM, which combines environment's RGB and thermal information together, and can work in both low illumination and general environment. In contrast to current state-of-the-art vision-based SLAM methods, our system keep a comparable accuracy in general environment. For low illumination environments in which these vision-based SLAM can not work, the proposed RGB-T SLAM system can still keep its accuracy and robustness. In our experiments, we show that the accurate and robust performance of our approach is comparable to other current SLAM systems.

TABLE I
KEYFRAME LOCALIZATION ERROR COMPARISON IN DIFFERENT SEQUENCES.

Absolute KeyFrame Trajectory RMSE (m)				
Sequence	RGB-T SLAM	ORB-SLAM	PTAM	LSD-SLAM
01	4.31	X	5.23	9.72
02	7.58	7.53	7.93	X
03	7.21	X	X	9.93
04	3.19	–	–	–
05	4.16	–	–	–
06	7.49	–	–	–

Experiment comparison results for RGB-T SLAM, ORB-SLAM, PTAM and LSD-SLAM. The trajectories have been aligned with 7DoF with the ground truth. X means the tracking is lost at some point and - means a significant portion of the sequence is not processed by the system or the system can not normally work.

References

[1] A. Concha, G. Loianno, V. Kumar, and J. Civera, "Visual-inertial direct slam," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, May 2016, pp. 1331–1338.

[2] J. Engel, J. Sturm, and D. Cremers, "Camera-based navigation of a low-cost quadrocopter," vol. 57, no. 1, pp. 2815–2821, 2012.

[3] B. G. Klein and D. Murray, "Parallel tracking and mapping for small," in *AR workspaces, International Symposium on Mixed and Augmented Reality*, 2012.

[4] M. Li and A. I. Mourikis, "High-precision, consistent ekf-based visualinertial odometry," *International Journal of Robotics Research*, vol. 32, no. 6, pp. 690–711, 2013.

[5] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," September 2014.

[6] S. Vidas and S. Sridharan, "Hand-held monocular slam in thermal-infrared," in *Control Automation Robotics Vision (ICARCV), 2012 12th International Conference on*, Dec 2012, pp. 859–864.

[7] 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.

[8] 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, 2014.

[9] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme," in *IEEE Intelligent Vehicles Symposium*, 2010, pp. 486–492.

[10] S. Vidas, P. Moghadam, and S. Sridharan, "Real-time mobile 3d temperature mapping," *IEEE Sensors Journal*, vol. 15, no. 2, pp. 1145–1152, 2015.

[11] S. Vidas, P. Moghadam, and M. Bosse, "3d thermal mapping of building interiors using an rgb-d and thermal camera," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2311–2318.

[12] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition(CVPR)*, Jun 1994, pp. 593–600.

[13] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *European conference on computer vision*. Springer, 2006, pp. 430–443.

[14] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, p. 4, 2001.

[15] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[16] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, Nov 2011, pp. 2564–2571.

[17] B. Triggs, P. F. Mclauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment a modern synthesis," *Vision Algorithms Theory and Practice*, 2000.

[18] D. Galvez-Lpez and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.