2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Towards Long Term SLAM on Thermal Imagery

Keil, C, Gupta, A., Kaveti, P., Singh, H.
Institute of Experiential Robotics. Northeastern University, Boston MA
{keil.c, gupta.anik, kaveti.p, ha.singh}@northeastern.edu

***Abstract*— Visual SLAM with thermal imagery remains a difficult problem for many state of the art (SOTA) algorithms. Compared with visible spectrum imagery, thermal imagery generally has lower contrast, higher noise, and tends to have lower resolution, making for challenging front-end data association. Thermal imagery also presents a difficult problem for long term relocalization and map reuse, because the relative temperatures of objects in thermal imagery tend to change dramatically from day to night. Feature descriptors typically used for relocalization in SLAM are unable to maintain consistency over these diurnal changes. We show that learned feature descriptors can be used within existing bag of word based localization schemes to dramatically improve place recognition across large temporal gaps in thermal imagery. In order to demonstrate the effectiveness of our trained vocabulary, we have developed a baseline SLAM system, integrating learned features and matching into a classical SLAM algorithm. Our system demonstrates good local tracking on challenging thermal imagery, and relocalization that overcomes dramatic day to night thermal appearance changes. Our code and datasets are available here: https://github.com/neufieldrobotics/IRSLAM_Baseline**

I. Introduction

Simultaneous Localization and Mapping (SLAM) allows robots to track their movement, and build knowledge about their environment. Visual SLAM systems, which use camera imagery as their primary sensing modality, are pivotal for a wide variety of robotics applications, but often struggle in environments with poor visibility or significant illumination changes, such as those encountered in nocturnal or adverse weather conditions. Long-Wave Infrared (LWIR) imagery, commonly referred to as thermal imaging, emerges as a promising solution to provide visibility in dark, dust-filled, or smoke-filled environments without lighting. In addition, thermal cameras can offer significant power and weight savings when compared with LIDAR, and can offer improved visibility in autonomous driving, or other scenarios where lighting is available. Unfortunately, temperature driven appearance changes in outdoor thermal imagery that manifest over even a few hours pose unique challenges, particularly in feature extraction and localization under varied environmental conditions.

Existing feature-based methods [15] [17] [23], are notably less effective with infrared (IR) imagery. This ineffectiveness is due to reduced and inconsistent feature extraction in the short term, and inverting image gradients caused by the variations in LWIR energy across different objects in the long term. We show that inconsistent feature extraction causes the ORB [27] based place recognition schemes used in almost all SOTA visual SLAM systems [3] [26] [37] to be ineffective over temporal gaps of only a few hours. Leading feature-based SLAM systems, such as ORB-SLAM3 [3], encounter significant difficulties. In contrast, state-of-the-art flow-based methods like DROID-SLAM [34] provide reasonable local tracking results, but lack an easily exploitable mapping/place recognition model conducive to relocalization within diurnal LWIR datasets. Other flow-based frameworks, including VINS-FUSION [26] and Basalt [37], rely on BRIEF/ORB features [2] for loop closure detection or relocalization, thus faltering with IR imagery.

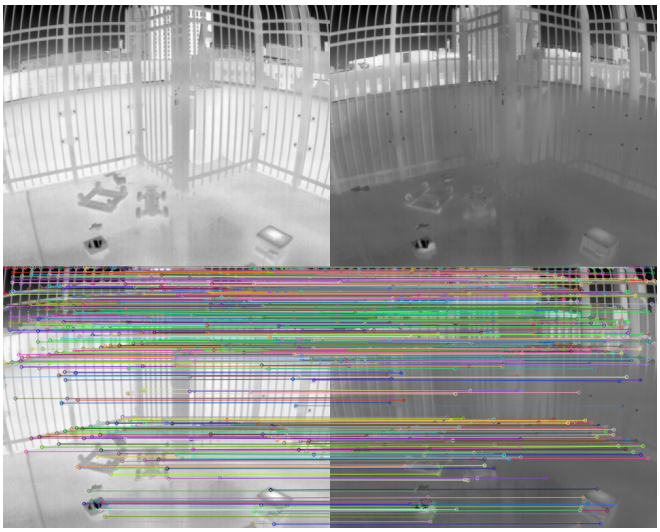


Fig. 1: Long Wave Infrared (thermal) imagery poses a significant challenge for place recognition due to dramatic appearance changes over the course of a day. At the top we show a pair of images taken with a static camera approximately 12 hours apart. At the bottom we show matches that are recoverable using the Gluestick feature matching pipeline.

This work endeavors to facilitate all-day autonomy for robotic systems employing LWIR cameras as their primary sensor. Our approach initially revisits classical feature extraction techniques, identifying issues with their suitability for LWIR imagery. Subsequently, we advocate for the utilization of Gluestick [24] as a learning-based method for extracting and matching features resistant to illumination changes. Lastly, we integrate this learned feature descriptor within MCSLAM [16] to assess its visual SLAM performance. For loop closure and relocalization testing, we develop a Bag of Words (BoW) vocabulary employing SuperPoint

features from LWIR images captured at various times of the day, utilizing data gathered in urban, outdoor settings via handheld or vehicle-mounted cameras.

To rigorously evaluate our method, and compare against other SLAM systems, we collected comprehensive test datasets over 24-hour periods using both static and mobile IR cameras with RTK GPS ground truth. These datasets highlight the inadequacies of most existing data collections, particularly in capturing the dynamic illumination conditions inherent to outdoor environments. Our experiments indicate that our Gluestick augmented variant of MCSLAM, adeptly tracks features and achieves relocalization between day and night imagery.

Our contributions can be summarized as follows:

- We present an extensive dataset collected with FLIR Boson thermal cameras. This includes a set of twenty four hour outdoor timelapses with static cameras, sequences with nearly identical camera trajectories captured in the same locations during the day and night, with GPS ground truth, and a large set of trajectories with no ground truth, which can be used for training BoW models, or for unsupervised training.
- We train a BoW vocabulary using Superpoint [9] features, which we make publicly available, to show effective loop closure and Visual Place Recognition (VPR) across day-night datasets.
- We propose an effective feature based visual SLAM baseline using MCSLAM [16] with Superpoint [9] features and the GlueStick [24] matcher which demonstrates strong local tracking, and includes our BoW vocabulary, allowing us to save a map during the day and accurately relocalize at night.

II. Related Work

A. Feature Matching in Thermal Images

Feature matching is a cornerstone of effective SLAM, providing the critical linkage between successive frames and across sessions. ORB [27], SIFT [21], and SURF [1] are widely used, hand-tuned feature detection algorithms. Performance evaluation of these features on thermal images in terms of detection, repeatability, and matching shows poor performance due to the non uniformity of noise and low contrast [23]. Learning-based features [39] [33] [9] [35] show a significant improvement in feature detection and matching, offering enhanced robustness and accuracy over classical features. Methods such as [22] [42], were specifically designed for thermal imagery. [42] proposed augmenting the SuperPoint model with a specialized noise filter for thermal imagery while [22] leverages cross-spectral data to extract thermal features. These descriptors were trained and evaluated for short time scales. However, thermal images between day and night have dramatic intensity differences (as shown in Fig. 1), including intensity inversions, and feature matching usually fails in these cases. In our approach, we propose to use Gluestick [24] which uses a Cross-Attention based matching scheme to robustly match images in challenging scenarios like low overlap, different lighting conditions etc.

B. Thermal Simultaneous Localization and Mapping (SLAM)

Visual SLAM is a well-explored research area with several versatile and robust methods, including sparse featured-based frameworks [3], multi-sensor systems [16], and direct flow-based methods [26] [37]. These popular classical SLAM systems require good lighting conditions and high overlap imagery to work and thus can fail in more challenging scenarios. Compared with visible light spectrum images, there is relatively little work in the SLAM space on thermal imagery. To overcome the limitations of thermal data, recent works have looked at fusing thermal imagery with visible spectrum imagery [25], LIDAR [32] and IMU data [38] [4] [18] [7]. [38] proposed an edge based feature tracking approach for improving thermal inertial odometry, while [29], and [15] proposed incorporation of 14 or 16 bit thermal images for direct thermal inertial solutions. Notably, the research is mostly biased towards odometry.

Effective place recognition and loop closure are pivotal for ensuring the consistency and reliability of SLAM systems, especially for long-term and multi-session mapping between day and night when the images look very different. Recent work on global image descriptors [43], [30] have shown impressive results in large real world environments, but their use has mostly been limited to offline structure from motion (SFM) applications. Bag of Words (BoW) approaches have been instrumental in achieving efficient loop closure detection with significantly lower computational requirements when compared with deep learning approaches. DBoW2 [12], which was first released in 2012, is still used by almost all state-of-the-art systems that employ place recognition [3] [26] [37] [31] [16]. Several recent studies have blended BoW vocabularies and learned features, using SuperPoint or similar features in BoW schemes for improved robustness. These have variously incorporated novel verification metrics [40], examined methods for turning SuperPoint into a binary descriptor [6], and used learned matching for verification [28]. Within the domain of thermal imagery, [29] uses the learned global descriptor approach, and [15] uses BoW, but neither of them explores day-to-night relocalization. [20] uses a generative image translation technique to approach the specific problem of day-to-night thermal relocalization, but does not show any results in a SLAM framework.

C. SLAM Datasets

SLAM datasets, such as the KITTI dataset [13], are critical for the development and comparison of SLAM systems. In comparison to visual spectrum or LIDAR datasets, SLAM focused thermal datasets are limited. [5] and [19] provide monocular thermal trajectories captured outdoors at different times. [41] provides unsynchronized stereo thermal imagery collected in the same locations at different times. We chose to collect our own dataset so as to have better control over the data collection procedure, and have better loop closure coverage.

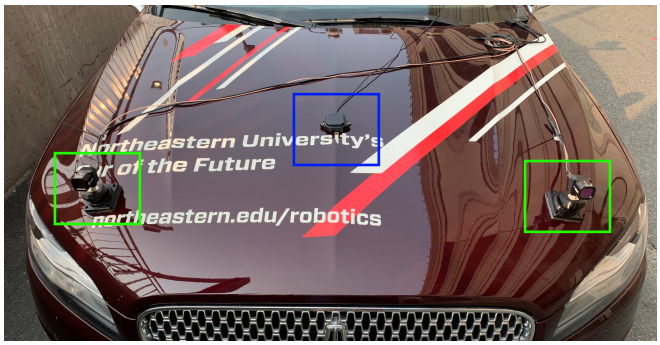


Fig. 2: Data Collection setup showing two FLIR Boson ADK cameras (green), and the RTK GPS antenna (blue).

With this context established, we organize the remainder of the paper as follows. In the next section we describe our data collection for training, analysis, and bench-marking our SLAM system. Following that we outline our method for training a bag of words vocabulary, and SLAM integration. Finally we discuss our validation experiments and conclusions.

III. Dataset

We collected a varied dataset of thermal imagery to evaluate Day-Night relocalization and SLAM performance, which we now publish. Here we describe our data collection procedure and the different formats we collected.

A. Collection Types

We collected three main types of data. First, we collected a test set of 24 hour outdoor timelapses with static cameras, in which there are very few dynamic objects. We provide ten scenes with images taken every ten minutes, showing buildings in a semi-urban environment. We use these scenes to benchmark methods where a pixel-level accurate ground truth is useful. Next, we collected a large set of monocular sequences with handheld and vehicle mounted cameras, with no ground truth trajectory information, which we use to augment our BoW training. Finally, we collected three sets of matched day and night trajectories of various sizes with stereo and side facing cameras. For these trajectories, we followed a pre-determined route during the day and then again at night, taking care to be as consistent as possible between attempts. We collect at least two loops at each time so that place recognition evaluations can be carried out for day-to-day/night-to-night and also for day-to-night, and visa versa. The features of these datasets are summarized in table I. For convenience we refer to the sequences by the names given in table I. For these sets we provide ground truth position data with RTK GPS, making for an easily verifiable set for benchmarking day-to-night and night-to-day loop closure. Images from the "KRI" sequence are shown in figure 3. In our subsequent evaluation we use the "KRI" loop as our test set and retain the others for training.

Name	Description
Garage Roof	100m baseline loops taken on top of a parking garage, with few dynamic objects.
Carter Field	350m baseline loops on a public road in Boston, with mobile cars and pedestrians.
KRI	350m basline loops at the Kostas Research Institute, with few dynamic objects.

TABLE I: SLAM Evaluation Datasets.

B. Hardware Setup

All of our images were captured using FLIR Boson ADK cameras, with a resolution of 512x640, and a horizontal field of view of 75 degrees. In the case of the paired day-night trajectories, we use an unsynchronized stereo pair with a baseline of 1.1m mounted on a Lincoln MKZ car. The stereo pair and RTK GPS antenna placement can be see in figure 2. We capture frames at 30fps for paired datasets, and at 60fps in the monocular trajectories. The system time for camera timestamps was synchronized with GPS time at the start of each trajectory.

C. Fixed Field Correction

Thermal cameras periodically need to recalibrate their image sensors to account for internal heat buildup. This is referred to as fixed field correction (FFC), or non-uniformity correction (NUC) in some publications [41]. During FFC, imagery is lost for a few frames, making tracking more difficult. FFC events occur every few minutes throughout each trajectory.

D. Calibration

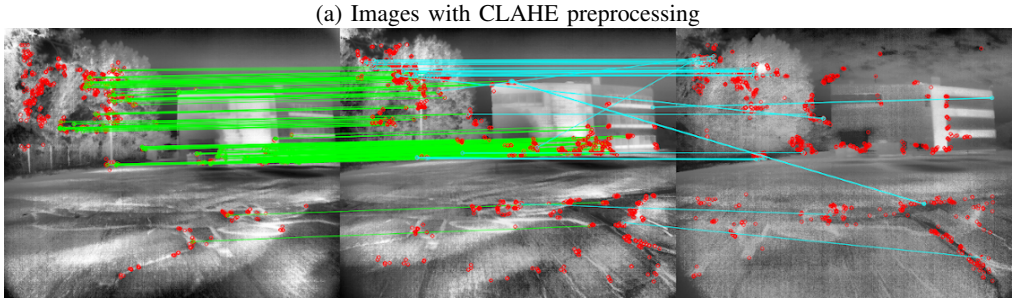
Traditional camera calibration targets show little to no contrast between white and black regions when viewed with a thermal camera. We use a wooden calibration target with a checkerboard pattern formed from copper tape. When the calibration target is heated with an external source the wood appears white, and the copper squares, which reflect the ambient environment, appear dark, providing enough contrast for calibration. We use Kalibr [11] to estimate the intrinsic coefficients for all cameras and the extrinsic parameters for our stereo pair. The extrinsic transform of the GPS antenna was manually measured.

IV. Method

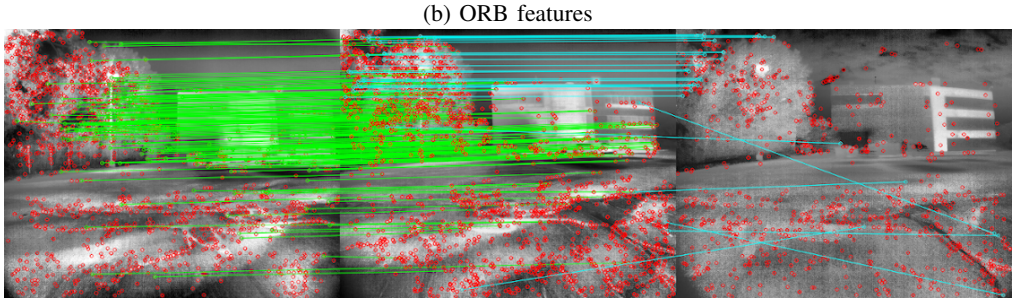
A. Image Preprocessing

Raw IR images output by the Boson camera have very poor contrast. We apply Contrast Limited Adaptive Histogram Equalization (CLAHE) [10], which increases contrast thereby increasing keypoint extraction, at the cost of amplifying noise. There is a trade-off between increasing the number of keypoints extracted, and increasing the the noise in the location of the extracted keypoints, especially for ORB features. We empirically determined a set of CLAHE parameters that perform well on our imagery in the context of SLAM. Note that noise in the IR images is characterized by a grid like overlay called fixed pattern noise [42]. An example of processed images can be seen in figure 3a.

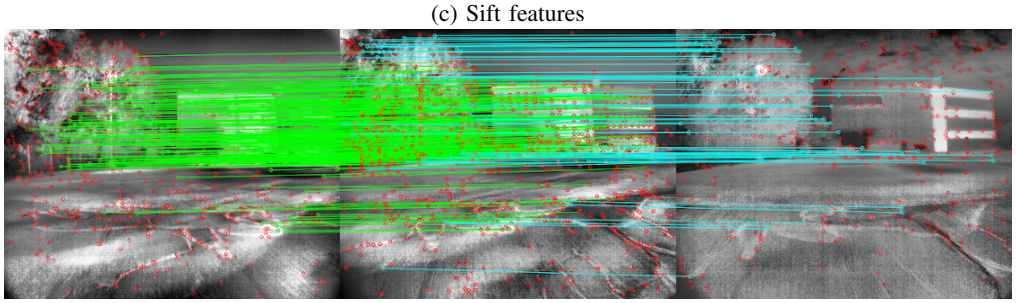
(a) Images with CLAHE preprocessing



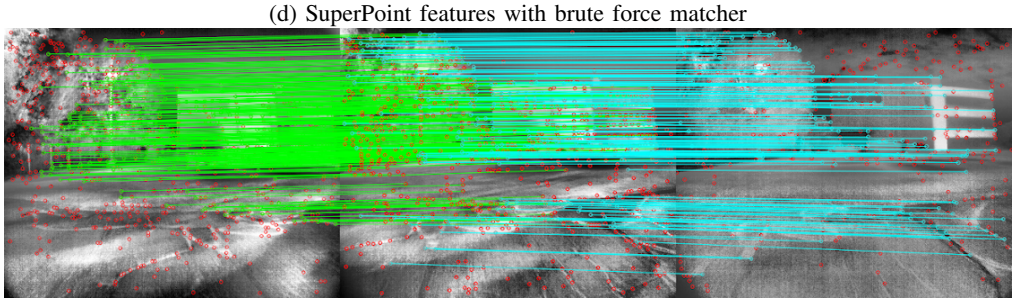
(b) ORB features



(c) Sift features



(d) SuperPoint features with brute force matcher



(e) SuperPoint features with Gluestick matcher

Fig. 3: We show a qualitative example of the number of matched features across images taken during the day (left pair) and then the same scene at night (right pair). At the top we show preprocessed images, below that we show ORB, Sift, SuperPoint and SP+Gluestick. Features are matched with brute force matching and are filtered for geometric consistency using a RANSAC fundamental matrix estimation. SP features with the Gluestick matcher outperform all other methods and are notable better at matching features in the foreground, which is important for parallax in feature based pose estimation.

B. Gluestick

Our use of Gluestick [24] is motivated by the qualitative analysis show in figure 3. We show a pair of images captured at the same time during the day with a small camera translation, and a third image from the same location captured at night, with feature matches for ORB, SIFT, SuperPoint, and Gluestick. We can clearly see that the learned features perform much better in both the day-day and day-night scenarios. Gluestick is notably better at matching features in the foreground, which is important for accurately measuring camera translation. Figure 4 shows a quantitative comparison for the same features using our timelapse dataset to evaluate matches across a twenty four hour period. We again see that Gluestick significantly outperforms the other methods. ORB and Sift features struggle particularly with day to night matching, because they rely on the corner gradient orientation to for orientation invariance. Changing temperature gradients over time result in completely different feature orientations and descriptors, or in features dropping below detection thresholds entirely. The learned features do not explicitly use a feature orientation, have significantly greater numerical complexity, and are trained to be robust to noise, resulting in better performance.

With the above analyis in mind, we use Gluestick, to generate and match SuperPoint features for training our BoW vocabularies, and also for front-end SLAM data association. Gluestick notably also matches lines from image pairs, which are represented by the SuperPoint descriptors at the line endpoints. We observed that the matched lines are often qualitatively correct, but the locations of endpoints are obviously variable from frame to frame. We use the matched line features for place recognition, but do not use them to estimate camera motion. Initially we attempted to use Superpoint features alone, matched using classical techniques, in a modified version of ORB SLAM3, but generally found that this system had poor performance and would lose tracking frequently. Gluestick produces better matches, and introduces far fewer incorrect matches, yielding much better performance in our final SLAM pipeline. In our experiments we achieve good results using the pretrained weights for Gluestick. In the future, retraining for IR imagery is possible with a more sophisticated data collection scheme to generate sufficiently diverse IR imagery with depth and pose ground truth.

C. Place Recognition

We extend DBoW2 [12] to build a vocabulary for Superpoint features extracted from IR images, with the aim of achieving good day-night/night-day place recognition. To train the vocabulary we match pairs of images from sequences in our training sets using Gluestick and then use only the features that are successfully matched across the pairs as training input for the vocabulary. We train with approximately 38,000 pairs of images from a mix of standalone sequences, and paired day-night sequences captured across several months. Omitting the paired trajectories, using only

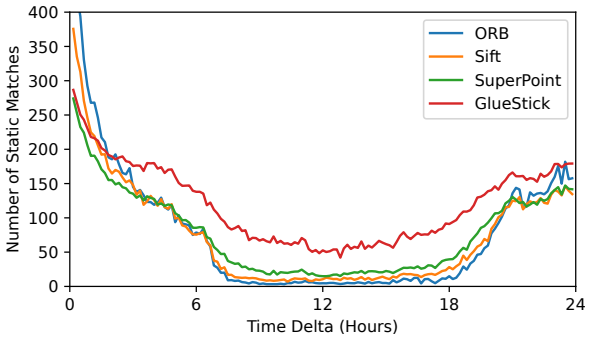


Fig. 4: The mean number of correct matches for different feature types from 10 scenes in our static timelapse dataset. We match features from the first image from each scene to every other image taken over 24 hours at 10 minute intervals. Proposed matches are taken as correct if the pixel disparity is less than three pixels. For ORB, Sift, and SuperPoint we use a brute force matcher.

day or only night images resulted in poor temporal generalization, suggesting that online vocabulary building methods would not work well here. Note that we do not explicitly match across day and night images, and only use sequential image pairs from within the same sequence. We are able to achieve compelling results without handling day to night matches in a principled way. In the future we can improve on this by using GPS information to select day-night pairs, and explicitly train a vocabulary with day to night feature matches. We empirically determined that a vocabulary with five levels and a branching factor of 10 performs best. This is 10x smaller than the vocabulary used by ORB-SLAM. Larger vocabularies exhibit poor performance on day to night matches in our testing, because the increased number of words results in corresponding day and night features being assigned to different words.

The main advantage of using binary features for place recognition in a BoW scheme is that they are extremely fast to compute. In our implementation, the L2 distance computation for comparing Superpoint's floating point valued features takes roughly 80 times as long to compute as the Hamming distance used for brief features. However, even with this performance constraint, the single threaded DBoW2 implementation running on a modern laptop can add Superpoint features to a large database and search for match candidates at over 50 frames per second, which is easily fast enough for real world SLAM loop closure. In a typical BoW scheme the majority of the computation time for adding an image or querying a database is dedicated to computing the BoW feature vector for an image. Once a BoW feature vector has been computed, the actual lookup time to search a database is essentially the same regardless of the original feature type (assuming a similar number of features per image, and distribution of those features in a database) because the comparisons are on words in the vocabulary, which have the same datatype for ORB or SuperPoint, or any other feature.

D. SLAM Pipeline

We chose to use Multi Camera SLAM [16], (MCSLAM) as the basis of our SLAM system. MCSLAM is a flexible feature based framework, partially based on ORB-SLAM, which supports camera setups beyond the typical stereo-pairs supported by most popular SLAM systems. In this work we limit our analysis to stereo imagery, however due to the low resolution of IR cameras and general difficulty of working in IR imagery, this choice also allows us to consider arrays of cameras with overlapping and non overlapping fields of view for a more robust system in the future, or examine combined arrays of RGB and IR cameras.

In our implementation we replace ORB features with Superpoint descriptors, and improve most of the matching framework with Gluestick. The open source implementation of Gluestick works as an end to end pipeline, taking a pair of images and outputting a prediction which contains all of the extracted features and matches. We made some minor changes to the model so that instead of working in one step, we can individually extract features for each image, and then separately match those features with as many other sets of features as needed. In the MCSLAM workflow, images that are taken at the same time are referred to as a multi-camera frame. In our adaptation, for each multi-camera frame (always stereo pair in this work) we extract features for each image and then use Gluestick to find matches across the cameras. These matches become 3D landmarks, while unmatched features become monocular landmarks. We then associate the current multi-camera frame with the previous one by matching each camera's current image features with the previous set of features for that camera. The resulting set of associations between the current and previous frames can then be used by the MCSLAM backend to establish the relationship to the 3D map. MCSLAM currently supports a BoW loop closure mechanism based on DBoW2 and DLoopDetector [12], which allows us to use our SuperPoint vocabulary to achieve relocalization. We use the stereo features matched in each frame, rather than the full set, to help limit the search to stronger features.

V. Experiments

A. Place Recognition

Here we demonstrate our place recognition system in isolation from the full slam system in order to systematically demonstrate it's effectiveness. First, we test loop closure with a minimal temporal gap on the KRI test dataset. Running a separate experiment for day and night, we split the trajectories, using the first loop to build a BoW database and then search that database with queries from the second loop. The descriptors used to build and query the database are matched features from sequential frames in the sequence. We test the database using 100 uniformly distributed images and take the best scoring candidate as a loop closure, after rejecting false positives by requiring a minimum number of matches be RANSAC inliers of a fundamental matrix estimation. We set the false positive rejection threshold to the minimum required to eliminate all false positives, as is typically done in relocalization experiments [12]. We compare against ORB features using the vocabulary distributed with ORB-SLAM, and ground truth the experiment using GPS. We are able to detect an image with strong overlap in 100% of the cases for both day and night, with ORB performing nearly as well. See the results in table II.

	Ours	Orb Vocabulary
Day-Day	100%	100%
Night-Night	100%	97%
Search Day DB with Night Images	93%	12%
Search Night DB with Day Images	91%	10%

TABLE II: Recall at 100% precision for our IR SuperPoint vocabulary and an ORB vocabulary, using our KRI test dataset.

Next, to demonstrate results for a significant temporal window, we perform the same analysis but search for 100 frames from the KRI day set against a database of KRI night images, and visa versa. We are able to show 91% and 93% success respectively, with no false positives. The same experiment with ORB features shows only 12% and 10% correct matches respectively (see table II). In general, ORB features from thermal imagery are not viable for place recognition across a significant temporal gap, making map reuse implausible. Gluestick shows promise.

B. SLAM

Actually re-localizing relative to a previous map is significantly more challenging than finding loop closure candidates from a set of images. There is significant added complexity from a software engineering standpoint, and the SLAM system must be able to recover an accurate 3D pose from feature matches. With this in mind, we test our augmented MCSLAM in two ways. First we analyze our front-end tracking on IR trajectories. Second, we examine relocalization across the day-night temporal gap. In all cases we use the KRI trajectories as the basis of our comparison, because we used the other trajectories to train our BoW model.

1) Front End Tracking: Figure 5 shows a qualitative comparison of our results on the KRI day dataset, relative to the GPS ground truth. We do not show comparisons against any feature based SLAM methods, because MCSLAM [16] and ORB SLAM [3] are both only able to track in stereo for short periods. We observed that ORB features can be matched from frame to frame, but over a sequence noisy keypoint extraction makes it difficult to build a 3D map. The results from the Day trajectory show that we are able to generate reasonably accurate trajectories, implying similarly accurate maps. We are not able to maintain tracking through the entire trajectory due to one difficult point where there are minimal features near enough to the camera for tracking, highlighting the difficulty of using IR data for SLAM.

2) Relocalization: To demonstrate relocalization within a map, we use the saved map from our daytime trajectory shown in figure 5, and attempt to relocalize at every time step in the night trajectory. Relocalization for a single

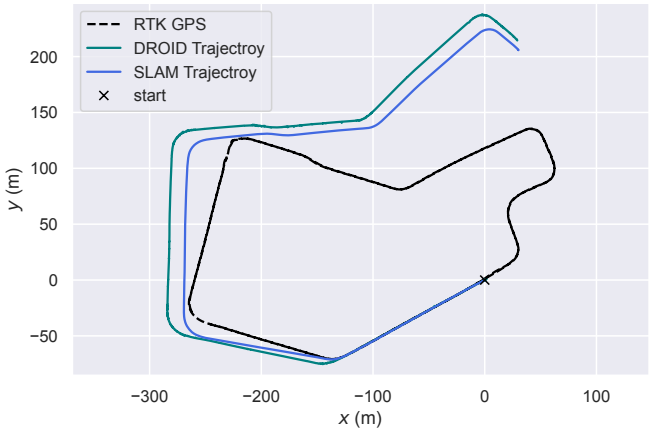


Fig. 5: Day trajectory for the KRI dataset. ORBSLAM and ORB-MCSLAM are only able to track for small sections of the map. We show a comparison with Droid SLAM, even though it is not directly relevant to our work because it shows similar tracking performance. There is significant accumulated drift, but scale and qualitative features are correct. The trajectory ends due to a very difficult low texture region. Note that Droid SLAM is able to track longer but we have stopped it at the same point for visual clarity.

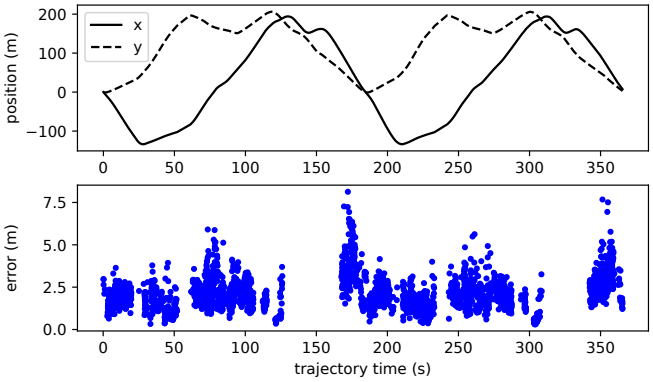


Fig. 6: Estimated relocalization error computed from the KRI night sequence relocalized into the KRI day map. The upper plot shows the ground truth trajectory x and y positions (east-west and north-south). The bottom plot shows the magnitude of the estimated error at corresponding points in the trajectory. Note that the trajectory contains two approximately identical loops, and that the large gaps in the error plot correspond to the section of the trajectory that is not included in the day map. We can see that the trend of relicalization error is consistent across the two cycles. The same results are shown overlain on the map in figure 7

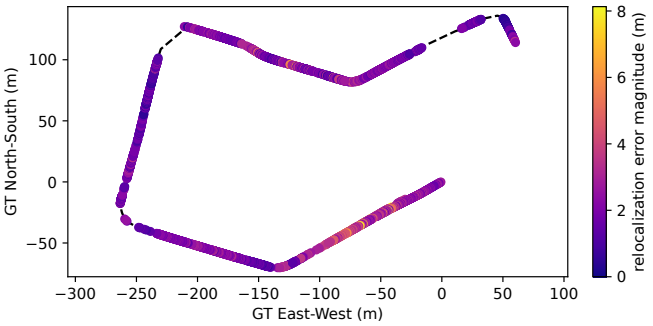


Fig. 7: Error magnitude heatmap of the night trajectory relocalized into the day map. At certain points in the trajectory, the average error spikes. This generally happens because in these regions the strongest features tend to be on distant structures.

stereo frame is achieved by searching a DBoW2 database for similar images, removing false positives with an island identification procedure similar to the implementation in DLoopDetector [12], matching the features from the map images with the query images using Gluestick, associating those image features with 3D points in the saved map, and then computing a perspective n-point solution with GTSAM [8] to estimate the camera's pose in the map. We threshold a minimum number of RANSAC inliers to reject low quality estimations, and reject a small number relocalization poses that are more than 10m from the pose of the map keyframe returned by our BoW search. Our map has a noticeable drift over time, so we estimate the relocalization error by locally aligning the map to its GPS ground truth, applying the alignment transformation to our relocalized position, and computing the error between the transformed, relocalized position and the GPS ground truth for the relocalization frame. The local alignment is based on the Umeyama alignment [36] implementation in EVO [14]. The resulting error is show in figure 6. In the majority of cases we are able to relocalize with less than 3m of error. We can achieve single shot relocalization over most of the map, with low error relative to the size of the map, and make the general observation that larger relocalization error is not attributed to incorrect BoW candidate selection, but to noisy pose estimation resulting from the majority of visual features being on distant objects in some areas of the map.

VI. Conclusion and Future Work

We have shown that challenging loop closure, enabling map reuse, is possible within long term IR datasets. We are able to achieve this using a BoW system, making it suitable for relatively simple incorporation into existing SLAM systems. Our baseline SLAM system is able to generate maps using Gluestick for data association, and outperforms feature based SLAM systems that use binary descriptors. One avenue for future work would be to better optimize the system for memory use and speed. Gluestick is not a perfect drop in replacement for the efficient matching scheme used in MCSLAM, or other feature based methods. Improvements and optimizations could be made with regards to matching across more than two frames, and matching between cameras with known extrinsic parameters. In the future we will be looking at collecting larger, more diverse datasets that include other sensing modalities for comparison, enabling us to retrain or fine-tune feature extraction and matching, build better vocabularies, and conduct more thorough analysis.

REFERENCES

[1] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006.

[2] Michael Calonder, Vincent Lepetit, and Pascal Fua. Brief: Binary robust independent elementary features. 12 2011.

[3] Carlos Campos, Richard Elvira, Juan J. Gómez Rodríguez, José M. M. Montiel, and Juan D. Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[4] Xingxin Chen, Weichen Dai, Jiajun Jiang, Bin He, and Yu Zhang. Thermal-depth odometry in challenging illumination conditions. *IEEE Robotics and Automation Letters*, 2023.

[5] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[6] Joan P. Company-Corcoles, Emilio Garcia-Fidalgo, and Alberto Ortiz. Appearance-based loop closure detection combining lines and learned points for low-textured environments. *Auton. Robots*, 46(3):451–467, mar 2022.

[7] Jeff Delaune, Robert Hewitt, Laura Lytle, Cristina Sorice, Rohan Thakker, and Larry Matthies. Thermal-inertial odometry for autonomous flight throughout the night. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1122–1128. IEEE, 2019.

[8] Frank Dellaert and GTSAM Contributors. borglab/gtsam, May 2022.

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018.

[10] R. Eustice, O. Pizarro, H. Singh, and J. Howland. Uwit: underwater image toolbox for optical image processing and mosaicking in matlab. In *Proceedings of the 2002 Interntional Symposium on Underwater Technology (Cat. No.02EX556)*, pages 141–145, 2002.

[11] Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013.

[12] Dorian Gálvez-López and J. D. Tardós. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, October 2012.

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[14] Michael Grupp. evo: Python package for the evaluation of odometry and slam. `https://github.com/MichaelGrupp/evo`, 2017.

[15] Jiajun Jiang, Xingxin Chen, Weichen Dai, Zelin Gao, and Yu Zhang. Thermal-inertial slam for the environments with challenging illumination. *IEEE Robotics and Automation Letters*, 7(4):8767–8774, 2022.

[16] Pushyami Kaveti, Shankara Narayanan Vaidyanathan, Arvind Thamil Chelvan, and Hanumant Singh. Design and evaluation of a generic visual slam framework for multi camera systems. *IEEE Robotics and Automation Letters*, 8(11):7368–7375, 2023.

[17] Shehryar Khattak, Frank Mascarich, Tung Dang, Christos Papachristos, and Kostas Alexis. Robust thermal-inertial localization for aerial robots: A case for direct methods. In *2019 International Conference on Unmanned Aircraft Systems (ICUAS)*, pages 1061–1068. IEEE, 2019.

[18] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Keyframe-based thermal–inertial odometry. *Journal of Field Robotics*, 37(4):552–579, 2020.

[19] Alex Junho Lee, Younggun Cho, Young-sik Shin, Ayoung Kim, and Hyun Myung. Vivid++ : Vision for visibility dataset. *IEEE Robotics and Automation Letters*, 7(3):6282–6289, 2022.

[20] Dong-Guw Lee, Hyeonjae Gil, Seungsang Yun, Jeongyun Kim, and Ayoung Kim. Night-to-day thermal image translation for deep thermal place recognition. *Intelligent Service Robotics*, 16(4):403–413, 2023.

[21] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

[22] Yawen Lu and Guoyu Lu. Superthermal: Matching thermal as visible through thermal feature exploration. *IEEE Robotics and Automation Letters*, 6(2):2690–2697, 2021.

[23] Tarek Mouats, Nabil Aouf, David Nam, and Stephen Vidas. Performance evaluation of feature detectors and descriptors beyond the visible. *Journal of Intelligent & Robotic Systems*, 92:33–63, 2018.

[24] Rémi* Pautrat, Iago* Suárez, Yifan Yu, Marc Pollefeys, and Viktor Larsson. GlueStick: Robust image matching by sticking points and lines together. In *International Conference on Computer Vision (ICCV)*, 2023.

[25] Liang Qin, Chang Wu, Xiaotong Kong, Yuan You, and Zhiqi Zhao. Bvt-slam: A binocular visible-thermal sensors slam system in low-light environments. *IEEE Sensors Journal*, pages 1–1, 2023.

[26] Tong Qin, Peiliang Li, and Shaojie Shen. Vins-mono: A robust and versatile monocular visual-inertial state estimator. *IEEE Transactions on Robotics*, 34(4):1004–1020, 2018.

[27] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011.

[28] Ali Samadzadeh and Ahmad Nickabadi. Srvio: Super robust visual inertial odometry for dynamic environments and challenging loop-closure conditions. *IEEE Transactions on Robotics*, 39(4):2878–2891, 2023.

[29] Muhamad Risqi U. Saputra, Chris Xiaoxuan Lu, Pedro Porto B. de Gusmao, Bing Wang, Andrew Markham, and Niki Trigoni. Graph-based thermal–inertial slam with probabilistic neural networks. *IEEE Transactions on Robotics*, 38(3):1875–1893, 2022.

[30] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12708–12717, 2019.

[31] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[32] Young-Sik Shin and Ayoung Kim. Sparse depth enhanced direct thermal-infrared slam beyond the visible spectrum. *IEEE Robotics and Automation Letters*, 4(3):2918–2925, 2019.

[33] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE international conference on computer vision*, pages 118–126, 2015.

[34] Zachary Teed and Jia Deng. DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras. *Advances in neural information processing systems*, 2021.

[35] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. *Advances in Neural Information Processing Systems*, 33:14254–14265, 2020.

[36] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.

[37] V. Usenko, N. Demmel, D. Schubert, J. Stueckler, and D. Cremers. Visual-inertial mapping with non-linear factor recovery. *IEEE Robotics and Automation Letters (RA-L) & Int. Conference on Intelligent Robotics and Automation (ICRA)*, 5(2):422–429, 2020.

[38] Yu Wang, Haoyao Chen, Yufeng Liu, and Shiwu Zhang. Edge-based monocular thermal-inertial odometry in visually degraded environments. *IEEE Robotics and Automation Letters*, 8(4):2078–2085, 2023.

[39] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European conference on computer vision*, pages 467–483. Springer, 2016.

[40] Haosong Yue, Jinyu Miao, Yue Yu, Weihai Chen, and Changyun Wen. Robust loop closure detection based on bag of superpoints and graph verification. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3787–3793, 2019.

[41] Seungsang Yun, Minwoo Jung, Jeongyun Kim, Sangwoo Jung, Younghun Cho, Myung-Hwan Jeon, Giseop Kim, and Ayoung Kim. Sthereo: Stereo thermal dataset for research in odometry and mapping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3857–3864, 2022.

[42] Shibo Zhao, Peng Wang, Hengrui Zhang, Zheng Fang, and Sebastian Scherer. Tp-tio: A robust thermal-inertial odometry with deep thermalpoint. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4505–4512, 2020.

[43] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19370–19380, 2023.