Hand-held Monocular SLAM in Thermal-infrared

Stephen Vidas, Sridha Sridharan
Image and Video Research Lab, SAIVT Research Program, QUT
2 George St, Brisbane, Australia
Email: {stephen.vidas, s.sridharan}@qut.edu.au

***Abstract*—Thermal-infrared imagery is relatively robust to many of the failure conditions of visual and laser-based SLAM systems, such as fog, dust and smoke. The ability to use thermal-infrared video for localization is therefore highly appealing for many applications. However, operating in thermal-infrared is beyond the capacity of existing SLAM implementations. This paper presents the first known monocular SLAM system designed and tested for hand-held use in the thermal-infrared modality. The implementation includes a flexible feature detection layer able to achieve robust feature tracking in high-noise, low-texture thermal images. A novel approach for structure initialization is also presented. The system is robust to irregular motion and capable of handling the unique mechanical shutter interruptions common to thermal-infrared cameras. The evaluation demonstrates promising performance of the algorithm in several environments.**

I. Introduction

The nature of thermal-infrared images makes them highly robust to changing lighting conditions and other environmental effects such as the presence of fog, smoke and dust [1] [2]. Vision-systems which utilize this modality can therefore operate effectively in difficult outdoor settings, or even in total darkness [3]. Thermal-infrared Simultaneous Localization And Mapping (SLAM) could be used to help guide and localize robots and other vehicles when conditions cause other sensors to fail, or perform poorly.

However, working in the thermal modality brings with it a range of challenges [4]. In the context of a video-based SLAM system, the most important challenges identified were:

- Limited resolution
- Low SNR (Signal-Noise Ratio)
- Data interruptions
- Poor spatial distribution of texture

The first three listed challenges have the effect of making it difficult to achieve a large number of reliably tracked features between frames in a thermal-infrared video sequence. The fourth listed challenge extends beyond the feature tracking problem and into the problem of SLAM. Only objects which are powered, metabolizing or have undergone a non-uniform exposure to thermal radiation tend to produce good textures for feature tracking. In most environments, there may only be a few of these objects in view at any one time, and they are rarely distributed well over the camera's field of view. This poses challenges for accurately initializing structure, particularly when the features lie on a planar surface such as common in many machines and man-made surfaces.

The core innovations of this paper seek to adapt modern SLAM algorithms to be effective in thermal-infrared. In the authors' experiments, no existing SLAM systems available to the computer vision and robotics communities have been able to perform the task. Proposed methods for addressing the challenges associated with the feature tracking problem are outlined in Section III. These include the layered implementation of multiple feature detectors, and feature track refinement and filtering. Additionally, a homography guided search is proposed for the online handling of data interruptions due to the unique thermal mechanics of the camera.

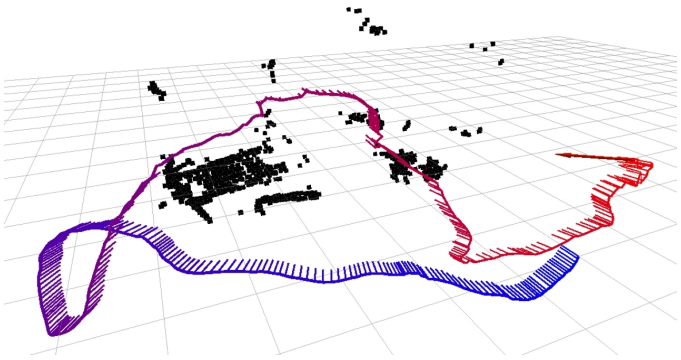

Fig. 1. An example reconstruction of the same scene as shown in Fig. 3. The feature-rich laptop keyboard can be discerned at the left of the point cloud.

The proposed implementation also takes several steps to achieve stable initialization and triangulation in thermal-infrared, which proved to be more difficult than in the visible-spectrum. Section IV-A presents a novel approach for selecting frames with a high probability of accurately intializing structure. The approach combines several metrics as a weighted product, which enables a single threshold to be varied to control for desired confidence. Section IV-B then discusses the proposed technique for ensuring reliable triangulation by subselecting views based on degree of separation.

An evaluation of the system's performance for the problem of real-time monocular SLAM in thermal-infrared is performed in Section V. The approach is inspired by the PTAM (Parallel Tracking and Mapping) [5] and VSLAM[1] implementations. The system has been implemented in ROS[2] and a driver for the Miricle 307K Thermal-IR camera is shared online[3]. Fig. 1 shows an example of the performance of the proposed approach.

[1] http://ros.org/wiki/vslam
[2] http://www.ros.org/wiki/
[3] http://code.google.com/p/thermalvis-ros-pkg/

II. Image Front-end

For the experiments, a Thermoteknix Miricle 307K thermal-infrared camera was used. This consists of a long-wave uncooled microbolometer detector sensitive in the 7-14μm range. The camera has a resolution of 640 by 480 pixels and is tested to see objects in the range of -20 to 150 degrees Celsius. It has a NEDT (Noise-Equivalent Differential Temperature) of 85mK.

14-bit monochromatic images are streamed from the camera at 25 frames per second. The intensity range is typically much smaller than 14-bits, and so the histogram can be normalized to a conventional 8-bit image to simplify processing. For this paper, the normalization procedure involved averaging the minimum and maximum intensity for a frame, and then normalizing the range spanning from 256 less than this average value to 256 greater than this value to the 8-bit interval of [0, 255]. This scaling (corresponding to a quantization factor of 2) ensures that texture is preserved in both high-contrast and low-contrast environments. In addition, temporal smoothing is used to ensure that the normalization mean does not shift by more than one level relative to the preceding frame.

A driver for streaming from this model camera utilizing the FFMPEG[4] libraries was written for compatibility with ROS (Robotics Operating System)[5], and can be found online[6]. The spatial calibration of thermal-infrared cameras is significantly more difficult than for conventional cameras, and is crucial for the good performance of any SLAM algorithm. The method of [6] was used to calibrate the camera, and remove the effects of lens distortion. This method is specifically tailored for convenience in the thermal-infrared modality, and an implementation can be found online[7].

III. 2D Feature Tracking

In the interests of simplicity and parallelization, 2D feature tracking has been completely decoupled from the other components of the system. The feature tracking layer requires the specification of one or more feature detectors (such as GFTT [7] or FAST [8]) as well as a minimum acceptable response for each detector and a maximum desired number of tracks (N_d). An evaluation of feature detectors in thermal-infrared found that often satisfactory features can be found and matched between images in this modality if the sensitivity thresholds are lowered below what would normally be used in a visible-spectrum system [9]. For this paper, generally only the GFTT detector was used with a minimum Hessian response of 0.08.

Many uncooled bolometer thermal-infrared cameras incur a periodic delay in operation known as a NUC (Non-Uniformity Correction). This is required to re-calibrate the pixel values radiometrically, preventing intensity drift and the accumulation of image noise. These delays can freeze the camera output for longer than one second, causing tracking problems if the camera is still in motion.

[4]http://ffmpeg.org/
[5]http://www.ros.org/wiki/
[6]http://code.google.com/p/thermalvis-ros-pkg/
[7]http://code.google.com/p/mm-calibrator/

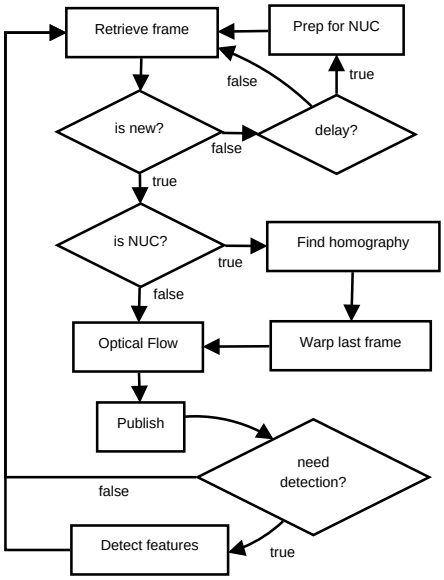

Fig. 2. The processes of the feature tracking mode.

The feature tracking node retrieves frames from the camera driver at the maximum rate, and follows the process shown in Fig. 2. The four checks involve determining the following:

- if the received frame is new;
- if there has been a sufficient delay since the last new frame to suggest that a NUC is occurring;
- if the new frame is the first one after a NUC; and
- if conditions warrant a new detection.

A. Detection

At launch, the first detector is run with the strongest acceptable features up to a maximum of N_d features retained. For each subsequent detector, only up to the strongest N_d acceptable features which are sufficiently separated (at least 5 pixels) from those already detected in the frame are retained and included in the tracking phase. This avoids redundant features which can over-emphasize certain regions and slow down the system.

For all newly detected features, the position of the feature is sub-pixel optimized by iterating to the radial saddle point. This refinement attempts to find the location q such that the expression in Eq. 1 is minimized for set of points p_i in the immediate neighbourhood of q.

$$\epsilon_i = DI_{p_i} \cdot (q - p_i) \tag{1}$$

Here, DI_{p_i} is the image gradient at one of the points p_i.

Fig. 3 shows an example of features tracked in a thermal-infrared video over several frames.

The two major conditions which call for the detector/s to be run again are:

- a low number of active tracks for a detector; or
- a specified amount of time elapsing between detections.

The threshold for the number of tracks is defined as 80% of the number of features retained for that detector after the

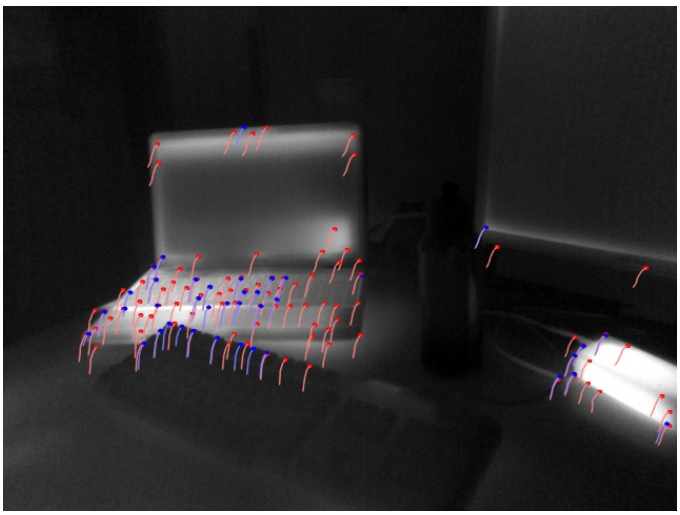

Fig. 3. Features tracked across 5 frames in sequence (1). Red features were detected with GFTT [7] and blue features with FAST [8].

previous detection. Each time a redetection is performed, only the strongest acceptable features from each detector are added to the active tracks, without the number of tracks for each detector exceeding N_d. Again, newly detected features are not considered if they are localized to within 3 pixels of an existing tracked feature (of any detector).

B. Tracking

For each new image, a sparse iterative version of the Lucas-Kanade optical flow algorithm [10], as implemented by OpenCV[8] is used. For consecutive images, an assumption of constant luminosity is generally valid, however when there is significant difference, the first image can simply be histogram-normalized to match the second. In normal cases, the initial estimates for the next locations of the tracked features are simply their positions in the previous frame.

The problem of NUC operations can be anticipated by checking for a long delay since the most recent unique frame. When this occurs, regular tracking is temporarily suspended, and SURF [11] features (with descriptors) are extracted from the last unique frame. As soon as the first new frame arrives, another set of SURF features are extracted and matching is performed between the two images on either side of the NUC operation. A homography H is then estimated to model the motion between the two frames. The first image is spatially warped together with its associated features to register it with the second image. This is done through the relationship shown in Eq. 2.

$$x' = \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = Hx \qquad (2)$$

[8] http://opencv.org/

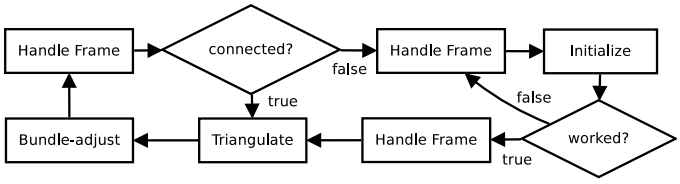

Fig. 4. The processes of the SLAM node.

Here, x' represents the new co-ordinates of the pixel or feature after warping, while x represents the original co-ordinates.

The histogram of this new warped image is then expanded and shifted in order to closely match that of the second image. The conventional Lucas-Kanade optical flow algorithm [10] is then used between this warped image and the new image, using the warped co-ordinates as the initial estimates for the new feature locations.

The effectiveness of this approach relies on the assumption that a homography is sufficiently accurate in modelling the camera motion over the NUC operation to provide reasonable estimates on new feature locations. Empirical testing found this to be a valid assumption in normal circumstances. The optical flow algorithm is generally capable of accommodating for the small effects of the 3D structure that invalidate the assumption of a pure homography relationship. However more sophisticated methods may need to be employed if the motion of the camera throughout the NUC process is severe.

IV. Localization and Mapping

The localization and mapping node includes structure initialization, pose estimation and on-line bundle-adjustment. This component continuously polls the feature tracking node checking for new tracks and projections, and follows the process shown in Fig. 4.

To summarize, each time an update is received from the feature tracking node, the system checks if the frame is sufficiently connected to any existing structure. If not, it waits until it receives a second frame, before attempting initialization which continues as more frames arrive. If an existing structure exists, the system attempts to triangulate any new points that occur in sufficiently well-distributed views, before estimating the current pose and adjusting a subset of the sequence in a single bundle-adjustment routine.

A. Initialization

When there is no established 3D structure in the system, a novel approach for selecting good keyframes for intializing the structure is used. The approach involves testing pairs of keyframes and assigning a keyframe score k dependent on a number of factors. As soon as a pair of frames that achieve a score above a specified threshold is found, both frames are selected and are used for obtaining an initial estimate of scene geometry.

The keyframe score k is obtained by normalizing a number of metrics to the range 0 to 1, and combining them in the formula shown in Eq. 3

$$k = \sqrt[n]{\prod_{i=1}^{n} S_i} \tag{3}$$

Here n metrics are used, and S_i is the normalized ith metric. The five metrics used in the proposed keyframe selection approach were:

- GRIC ratio
- Convergence error
- Proportion of points in front of cameras
- Translation score
- Angle

To calculate the metrics, the standard approach of calculating the Fundamental matrix based on tracked points, decomposing it into four possible transforms and selecting the best one was used [12]. The GRIC ratio was defined as F-GRIC divided by H-GRIC, as explained in [13]. When the F-GRIC score is lower, this suggests more 3D structure to the cameras' relative poses, rather than just simple rotation. The convergence error was then calculated by using sparse bundle adjustment[9] for 10 iterations. The bundle-adjusted camera poses were then used to calculate the points-in-front, translation and angle scores. The translation score was defined as the sum of the absolute translations in X and Y divided by the absolute translation in Z. This metric was chosen to help measure the degree of lateral motion of the cameras relative to each other, rather than straight forwards or backwards motion. The angle was simply the angle in degrees between the faced directions of each camera.

An asymmetric gaussian distribution was used to normalize each metric s, with a normalized score of 1 representing the best possible value for the metric, and a score of 0 the worst. The asymmetric gaussian mapping $S(s)$ is simply defined by two gaussian functions centered at the same mean (representing the ideal value for that metric), where each distribution is activated only for either values below or above the mean, and is shown in Eq. 4 [14].

$$S(s) = H(-\mu)N(s, \mu, \sigma_L) + H(\mu)N(s, \mu, \sigma_H) \tag{4}$$

where $H(\mu)$ is the heaviside step function and $N(s, \mu, \sigma)$ is a normal distribution with a peak of 1.0 centered at μ and with a variance of σ.

A total of approximately 640 trials using 5 different training sequences were used to estimate the parameters of these asymmetric gaussian mappings, by recording the values for each metric corresponding to each pair of frames that generated correct initializations. For each metric, the value μ (representing the ideal value) was calculated simply as the mean of the results corresponding to correct initializations. This was with the exception of the "points in front" score for which μ was assumed to be 1.00 (the case where all reconstructed points were in front of both cameras). Table I shows the learned parameters for the models used to calculate the normalized scores for each metric.

[9] http://www.ros.org/wiki/sba

TABLE I
IDEAL KEYFRAME-PAIR MODEL PARAMETERS

Metric	μ	σ_L	σ_H
Convergence	**1.67**	1.25	4.34
GRIC ratio	**1.32**	0.18	0.10
Points in front	**1.00**	0.08	-
Translation score	**27.3**	13.6	35.7
Angle	**8.64**	2.67	9.70

A minimum k value of 0.75 was used for the proposed system. This was found to avoid degenerate conditions in most cases. However for particularly difficult environments such as those with low or poorly distributed textural content this may be too strict. In these cases a lower threshold may need to be used for any chance of successfully initializing the system.

Once a valid pair of keyframes has been selected to initialize structure it is further optimized using sparse bundle adjustment. Using the now triangulated and refined 3D locations of points tracked between the two keyframes, a RANSAC-based PnP estimation is performed to estimate all of the intermediate poses. Bundle adjustment is then performed over this entire sequence. The structure is now considered "initialized" and the algorithm is ready to move into continuous operation and accept further frames.

B. Continuous Operation

As each new frame is received, a set of indices which represent the "active subset" of all previously received frames is selected. These indices represent a subset of cameras whose poses have already been estimated. In the proposed implementation, a maximum of 30 frames are used. The method for selecting keyframes is shown in Algorithm 1:

Algorithm 1 Keyframe selection

procedure INDICES(m) ▷ m: current frame
  $N \leftarrow min(m, N_{opt})$ ▷ N_{opt}: desired keyframes
  for $x \leftarrow 0, f-1$ **do** ▷ f: flowback
    $i_{N-x} = m - x$
  end for
  for $x \leftarrow f, N-1$ **do**
    if $(i_{(N-x+1)}/K) < (N-x)$ **then** ▷ K: spacing
      $i_{N-x} \leftarrow m - x$
    else
      $i_{N-x} \leftarrow i_{N-x+1} - 1 - ((i_{N-x+1} - 1) \bmod K)$
    end if
  end for
  return i ▷ **i**: keyframe indices
end procedure

This ensures that the desired number of cameras is selected, but is spread over a larger proportion of the frame history if

possible. For the experiments, f was set at 3 and K was set to 5.

With each new frame, untriangulated tracks that are viewed by a sufficient number of the active cameras are then checked over for those which can be triangulated. Triangulatable tracks are defined by those which are viewed by at least 10 pairs of active cameras whose estimated relative translations are above a threshold. The threshold is simply chosen to be 0.2, which is relative to the scale which is automatically normalized to separate the two intialization frames by a distance of 1.0.

Once points have been triangulated, the pose of the new frame is putatively estimated using the estimated velocity of the camera between the previous two frames. This pose is then added to the bundle-adjustment system, and a new set of keyframes selected. Bundle-adjustment is then performed with these new indices, with all cameras except for the most recent f fixed, and all points free to move.

V. Evaluation

A. Evaluation Sequences

Five sequences were used for the evaluation, which were different for those used for the training of the initialization parameters. The following is a brief description of the nature of these indexed test sequences:

1) The operator stood still and moved the camera my hand at relatively slow speeds facing a typical office desktop. Objects included a desktop monitor, power supply and laptop computer, of which the glossy finish resulted in substantial reflections.
2) A similar sequence to (1), but with fewer salient objects and less reflectivity.
3) An outdoor sequence, with the camera mounted to the rear of a bicycle and moving at a walking pace. Thermal contrast was good in this sequence, however, the motion of the bicycle meant that on several occasions the field of view changed considerably due solely to rotation.
4) An indoor sequence with a walking operator holding the camera at approximately chest height. Thermal contrast was fairly poor in this sequence. Salient objects were mostly ceiling-lights and desktop computers.
5) Another bicycle-mounted outdoor sequence. Motion and contrast were very good.

B. Processing time

Presently, the code has not been optimized, however real-time constraints can still be met easily on a modestly powerful computer (a quad-core 2.20GHz Intel Core i7 was used for the evaluation). Parallel processing has been utilized to enable certain processes to run simultaneously. Table II shows the processing time typically taken by each stage of the SLAM node of the system, which is the critical component in terms of balancing accuracy with efficiency.

The implementation of handling the tracks from the feature tracking node could be improved further; however the present implementation on average executes fast enough for the algorithm to run at 25 frames per second. Triangulation and putative pose estimation are completed very quickly. The bundle adjustment step can be completed effectively within approximately 20ms per frame. Full utilization of parallel processors has not yet been fully achieved for this step.

TABLE II
Processing time.

Process	Mean time (ms)	σ (ms)
Track handling	18.0	103.3
Triangulation	0.8	0.7
Putative pose	6.0	10.8
Bundle adjustment	18.6	12.2

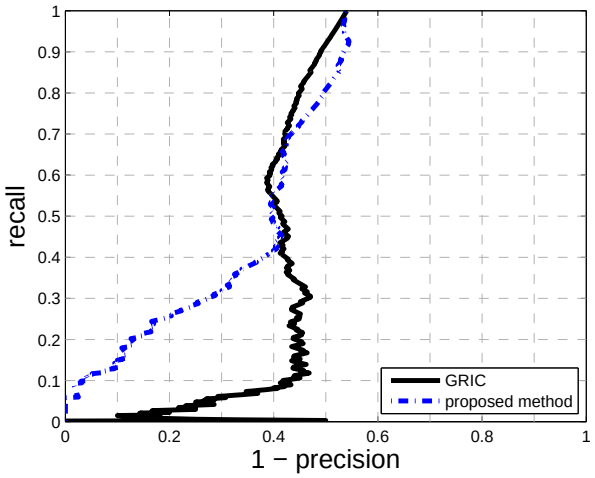

Fig. 5. Precision-recall curve for initial keyframe selection.

C. Initialization Performance

In order to evaluate the performance of the keyframe selection algorithm, the 5 testing sequences introduced in Section V-A were used. A total of 640 pairs of frames from these sequences were randomly selected and tested, with a pass awarded if they were able to be used to effectively initialize the system for ongoing operation. Each of the pairs was also scored using the proposed algorithm. Figure 5 shows precision-recall curves for each of the sequences, which demonstrate what proportion of keyframe pairs were effective as increasingly lower-scored keyframes were considered. From the figure it can be seen that the proposed approach is significantly better at ranking good keyframe pairs compared with the standard approach (using only GRIC). Weaker keyframe pairs are not sorted as effectively, however, in most applications only the higher-scoring pairs need to be considered.

D. Accuracy and stability

Reprojection error was used as the metric to judge the stability and performance of the system during ongoing operation. This considers the mean error of every ray from every point projecting onto each camera, without any points or projections being removed from the system. In the authors' experience, an

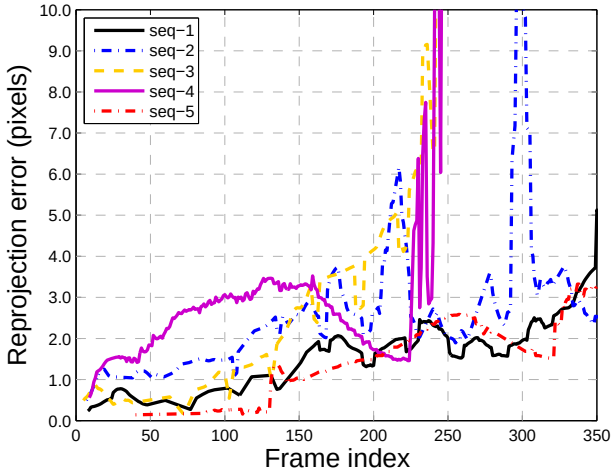

Fig. 6. Reprojection errors for the 5 test sequences.

error of below 1.50 generally corresponds to a good estimate of motion and a strong estimate of the structure of the scene. As the error increases up to approximately 3.00, the motion estimate may be usable for some purposes, but beyond this point the system becomes unreliable. Fig. 6 shows the change in reprojection error as the frame count increases in each sequence.

What follows is a brief discussion of the challenges of each sequence, which in some cases became failure conditions:

1) Late in the sequence, reflections of the moving human operator in the laptop monitor introduced a significant number of "false" points which were tracked, and which produced bad triangulations.
2) Results became increasingly inaccurate as more quick rotational motion was introduced, drastically reducing the number of triangulated points in view.
3) Well-tracked and triangulated features in this sequence were chiefly associated with a single structure. As the structure receded further into the background, features were tracked with decreasing precision and thus the reprojection error began to increase.
4) Breakdown in this sequence occurred during a period of mostly rotational motion, with the low thermal contrast of the scene limiting the number of available and reliable features to triangulate from.
5) A large tree with many strong features formed between the leaves and the comparitively "cool" backdrop (the sky) dominated this sequence, for which performance was relatively stable.

VI. Conclusion

An implementation capable of performing real-time monocular SLAM on thermal-infrared video has been presented. To the author's knowledge, this is the first system demonstrated to be capable of operating effectively in this difficult modality. The system has two key novelties: a normalized-product keyframe scoring algorithm for selecting keyframes for structure initialization, and a homography-guided optical-flow methodology for handling NUC (Non-Uniformity Correction interruptions).

The system is capable of performing in real-time without the aid of any other sensors. However, in its present form it is vulnerable to a number of failure conditions. Planned future work includes developing a more effective method for determining which keyframes should be used for ongoing system optimization. This will prevent circumstances where there is no longer sufficient structure in the subsystem for effective bundle-adjustment.

An anticipated application for the system is the augmentation of existing laser and vision-based SLAM systems for more robust performance under adverse conditions.

Acknowledgment

This project has been funded in part by a Queensland Government Smart Futures PhD Scholarship and a CSIRO (Commonwealth Scientific and Industrial Research Organisation) OCE (Office of the Chief Executive) Scholarship. Equipment as well as manufacturing and computing facilities were provided by both CSIRO and QUT (Queensland University of Technology).

References

[1] S. El-Tawab, M. Abuelela, and Y. Gongjun, "Real-time weather notification system using intelligent vehicles and smart sensors," in *Mobile Adhoc and Sensor Systems, 2009. MASS'09. IEEE 6th International Conference on*, 2009, p. 627632. 1
[2] C. Brunner, T. Peynot, and T. Vidal-Calleja, "Automatic selection of sensor modality for resilient localisation in low visibility conditions," 2012. 1
[3] S. Lin, "Technical report MS-CIS-01-04." 1
[4] K. Hajebi and J. S. Zelek, "Structure from infrared stereo images." IEEE, May 2008, pp. 105–112. 1
[5] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, 2007, p. 225234. 1
[6] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "A Mask-Based approach for the geometric calibration of Thermal-Infrared cameras," *IEEE Transactions on Instrumentation and Measurement*, 2012. 2
[7] J. Shi and C. Tomasi, "Good features to track," in *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, 1994, p. 593600. 2, 3
[8] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," *Computer VisionECCV 2006*, p. 430443, 2006. 2, 3
[9] S. Vidas, R. Lakemond, S. Denman, C. Fookes, S. Sridharan, and T. Wark, "An exploration of feature detector performance in the thermal-infrared modality," in *Digital Image Computing Techniques and Applications (DICTA), 2011 International Conference on*, 2011, p. 217224. 2
[10] J. Bouget, "Pyramidal implementation of the lucas kanade feature tracker." 3
[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, p. 346359, 2008. 3
[12] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2000. 4
[13] P. Torr, "Geometric motion segmentation and model selection," 1998. 4
[14] T. Kato, S. Omachi, and H. Aso, "Asymmetric gaussian and its application to pattern recognition," *Structural, Syntactic, and Statistical Pattern Recognition*, p. 227242, 2002. 4