2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Multimodal tracking framework for visual odometry in challenging illumination conditions

Axel Beauvisage*, Kenan Ahiska*, Nabil Aouf†
*Centre for Electronic Warfare, Information and Cyber, Cranfield University, DA of the UK, Shrivenham, SN6 8LA.
†Department of Electrical and Electronic Engineering, City University of London, London.
Email: *a.beauvisage@cranfield.ac.uk, *k.ahiska@cranfield.ac.uk, †Nabil.Aouf@city.ac.uk

***Abstract*—Research on visual odometry and localisation is largely dominated by solutions developed in the visible spectrum, where illumination is a critical factor. Other parts of the electromagnetic spectrum are currently being investigated to generate solutions dealing with extreme illumination conditions. Multispectral setups are particularly interesting as they provide information from different parts of the spectrum at once. However, the main challenge of such camera setups is the lack of similarity between the images produced, which makes conventional stereo matching techniques obsolete.**

This work investigates a new way of concurrently processing images from different spectra for application to visual odometry. It particularly focuses on the visible and Long Wave InfraRed (LWIR) spectral bands where dissimilarity between pixel intensities is maximal. A new Multimodal Monocular Visual Odometry solution (MMS-VO) is presented. With this novel approach, features are tracked simultaneously, but only the camera providing the best tracking quality is used to estimate motion. Visual odometry is performed within a windowed bundle adjustment framework, by alternating between the cameras as the nature of the scene changes. Furthermore, the motion estimation process is robustified by selecting adequate keyframes based on parallax.

The algorithm was tested on a series of visible-thermal datasets, acquired from a car with real driving conditions. It is shown that feature tracking could be performed in both modalities with the same set of parameters. Additionally, the MMS-VO provides a superior visual odometry trajectory as one camera can compensate when the other is not working.

I. Introduction

As visual based localisation techniques are becoming more advanced and reliable, the use of cameras has widely spread in research fields such as robotics, autonomous driving or mapping. They have now become the primary choice for scene perception when building autonomous systems.

Monocular Visual Odometry (VO) is especially interesting for small and light-weight platform such as ground robots or unmanned air vehicles (UAVs). It consists in estimating the 6 Degrees Of Freedom (DoF) ego-motion of a single moving camera. 2D features are detected in the different images acquired and matched to produce correspondences. The relative transformations between camera poses can then be recovered with 2D-to-2D motion estimation, solely based on the correspondences in the images [1], or alternatively with 3D-to-2D motion estimation [2], [3].

While 2D-to-2D estimation only describes the geometric relationship between two views, 3D-to-2D estimation can be used to jointly optimise the 3D points of the scene and the camera poses (position and orientation) in bundle adjustment scheme. These parameters are estimated by minimising an error metric between the observed and 3D features re-projected in the different images [4]. Two general approaches can be distinguished: *feature-based methods* which minimise the geometric error between an observed feature (obtained after matching) and its re-projected position in the image [1], [2]; and *direct methods* which minimise the photometric error between images, or in other words the difference in pixel intensities between local regions [5], [6].

Visual Odometry differs from Visual Simultaneous Localisation and Mapping (V-SLAM) in the way it makes use of the 3D points. While V-SLAM approaches build a consistent map of the environment and estimate the global pose of the camera in it, VO only estimates local motion which is accumulated over time to build the full camera trajectory [7]. Thus, V-SLAM can benefit from optimisation techniques such as loop closure, when the vehicle returns to a previously visited location [8]. The estimated trajectory can be further corrected and a more precise localisation is obtained. This, on the other hand, makes V-SLAM computationally heavier and leads to some specific issues such as map/feature data association [9]. Since the aim of the paper is to evaluate the use of multispectral images for self localisation, it focuses on VO and relative motion estimation rather than developing a full SLAM framework.

Due to the nature of projective geometry, monocular vision systems suffer from a scale ambiguity. At least two views are necessary to reconstruct the 3D position of a 2D feature, but because depth information is lost during the projection, an overall ambiguity remains. Even if the global scale cannot be recovered, several methods exist to estimate relative scale changes. For instance, in [10] a closed-form method with three views is proposed. It has the advantage of being much faster than iterative optimisation, but it is also less precise. To estimate the pose of a new frame relative to previous poses, a wide range of *Perspective N Points* (PnP) algorithms exist, such as P3P [11], [12] or EPnP [13]. These algorithms are based on 3D-to-2D correspondences. Compared to them, bundle adjustment is the most complete and accurate technique for computing relative poses over several views.

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

Even though multispectral imaging has been studied for a long time for medical applications [14], it only gained a certain interest in recent years for navigation purposes. Indeed, vision-based navigation has been largely dominated by the use of the different types of visible cameras. Yet, other spectral bands like the thermal band appear to be useful in many problems related to autonomous navigation such as space rendez-vous [15], navigation under varying illumination conditions [16] or pedestrian detection [17]. Although it remains overlooked, a few works on multispectral visual odometry can nevertheless be found. In [18], a new descriptor was developed to match multispectral stereo images and perform VO on car datasets. Then, Poujol et al. fused visible and thermal images to create enhanced images that they used for monocular VO [19]. Finally, in the concept of multispectral stereo odometry was extended to unmanned air vehicle (UAVs) in [20].

In order to extend the available literature, this work proposes the following innovations:

- it proves that, unlike feature matching, correspondences between consecutive frames can easily be computed with feature tracking in both visible and LWIR modalities without the need of tuning detection/matching parameters for each one of them. The features obtained can be used to perform accurate visual odometry.
- it tackles the challenge of extreme illumination conditions by presenting a multispectral monocular VO solution (MMS-VO) able to switch between modalities to always select the images that present the best tracking capabilities.

Section II introduces MMS-VO and presents the two main aspects of the proposed solution, namely feature tracking and bundle adjustment. In Section III, a series of experiments show the robustness and suitability of the solution for multispectral navigation in three different scenarios, including single modality failures.

II. Proposed approach

A new monocular VO framework for both infrared and visible images is presented. It carefully selects appropriate keyframes and processes them in a windowed bundle adjustment optimisation scheme (see Fig. 1). Features are registered and tracked over a sliding window using a pyramidal *Lucas-Kanade* algorithm (p-LK) [21]. The choice of feature tracking over feature matching was driven by the speed of the algorithm and consistency between images of each modality. Because it is fast, p-LK can deal with a greater number of features from which robust motion estimation can be performed. Furthermore, since it relies on the pixel consistency over consecutive frames, it can be used on various modalities with a common set of independent parameters (window size, optimisation stop condition, etc...). Hence, it does not require to tune specific thresholds and parameters intrinsic to each spectrum, which is usually the case when matching features, because the quality of the descriptors is directly affected by the nature of the images [22]. Based on optical flow, this technique is

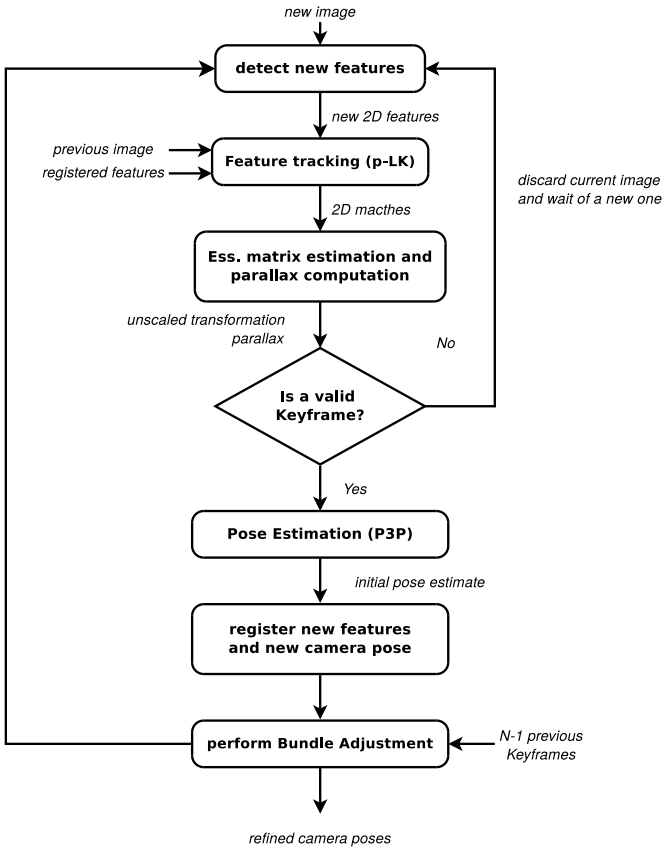


Fig. 1: Flowchart of the monocular VO algorithm for a single modality.

nevertheless affected by the texture and contrast present in the images.

For each new image, a set of additional features is detected in the previous image and added to the existing set of registered features. This guarantees that a sufficient number of points are tracked as the sliding window progresses, despite some of them being discarded or leaving the image scope. To ensure a robust tracking, the features that do not satisfy the epipolar constraint are removed. Additionally, to avoid degenerate cases where motion is limited between consecutive images, a keyframe approach is employed. As seen in Fig. 1, the validity of the current image is assessed by computing its parallax with the last keyframe. The parallax corresponds to the 2D shift produced between two images, due to pure translation between their camera centres. If the parallax is sufficient, a new keyframe is generated, otherwise, the image and tracked features are discarded. The observed points are then triangulated from the $N-1$ first views of the window. From the generated point cloud, the new camera pose is estimated with the P3P method [12], employed in a RANSAC framework to robustify the conducted estimation. Finally, bundle adjustment is applied to the whole window in order to refine the structure of the scene and the camera pose estimation. The points that have not been triangulated properly (do not appear in front of the camera) or which have been labelled outliers during

the RANSAC scheme are not used in the bundle adjustment optimisation.

A. Feature tracking (p-LK)

As mentioned previously, the use of optical flow based tracking [23] was preferred in order to develop a feature association technique which do not depend on the nature of the images. Thus, features are extracted based on the eigenvalues produced by the covariance of the gradient image. The threshold imposed to these eigenvalues is set low intentionally to detect enough features regardless of the contrast in the images. A certain subset (around 500 features) is then selected so all its elements are homogeneously distributed in the image. This guarantees that a constant number of features is being tracked, even if the environment/contrast varies over time, but it also ensures a generic extraction that could be adopted for all modalities.

Once extracted, features are tracked between consecutive frames with the pyramidal *Lucas-Kanade* method [21] and robustly rejected by running the 5-point algorithm in a RANSAC scheme [24]. Features are tracked until they fail to be tracked properly or leave the scope of the image. As the *Lucas-Kanade* method relies on the pixel intensity invariance between consecutive images, this guarantee that it will produce similar results in every modality as long as this principle is respected.

Finally, the processing speed of this method allows both modalities to be processed simultaneously, in real-time, which creates an opportunity to select the best modality for motion estimation. Therefore, only one set of 2D/3D points, from a single modality, is used to perform bundle adjustment. To determine which set to use, the camera providing the highest number of inliers after triangulation and P3P estimation is selected.

B. Bundle adjustment

Based on the pinhole camera model, a 3D feature $\mathbf{X}_i$ is re-projected into a specific 2D location $\mathbf{x}_{ij}$ in image I_j by the following re-projection function:

$$\begin{aligned} &\mathbf{x}_{ij} : \mathbb{R}^{6+3} \mapsto \mathbb{R}^2 \qquad i \in [1, N], j \in [1, M] \\ &\mathbf{x}_{ij} = \pi(\mathbf{S}) = P_j \mathbf{X}_i \end{aligned} \tag{1}$$

where P_j is the projection matrix for image I_j. It encapsulates the camera pose and calibration parameters. $\mathbf{S}$ corresponds to the vector of all parameters to be optimised. It contains the 3D location and 3D rotation vector each camera (6 DoF) as well as the 3D position of each feature (3 DoF). Rotation vectors are used as local parameterisation for the current iteration to improve the stability of the algorithm and are accumulated in global quaternion representations.

The least-squares cost function used for this bundle adjustment is defined as the sum of squared errors (SSE) for all N points in each camera:

$$f(\mathbf{S}) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} \Delta \mathbf{x}_{ij}^T W_{ij} \Delta \mathbf{x}_{ij}, \qquad \Delta \mathbf{x}_{ij} = \overline{\mathbf{x}}_{ij} - \mathbf{x}_{ij} \tag{2}$$

where W_{ij} is a symmetric positive definite matrix. For this least-squares formulation to have a meaningful statistical interpretation, W_{ij} should represent the inverse covariance of the observed feature $\overline{\mathbf{x}}_{ij}$ [4]. It then coincides with the log-likelihood of $\overline{\mathbf{x}}_{ij}$.

To avoid poorly tracked features to have a negative impact on the optimisation, a Maximum-Likelihood estimator function $\rho_i(\mathbf{x})$ is added to the formulation to reduce the influence of such features. In the implementation of this algorithm, the Hubert loss function was chosen [25]. The bundle adjustment formulation becomes:

$$\arg\min_{\mathbf{S}} \ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} \rho_i(\Delta \mathbf{x}_{ij}^T W_{ij} \Delta \mathbf{x}_{ij}) \tag{3}$$

Obviously, the main drawback of bundle adjustment is the number of parameters which grows every time a new keyframe is added or when new features are detected. This makes it impossible to use for real-time navigation when the framerate is high (more than 10 Hz) and the amount of points is significant (several hundreds). Hopefully, it does not have to be applied to all frames and points. Instead, it could be used locally with a limited number of parameters to refine the previously estimated structure and camera path [26], [27]. This method is called *windowed bundle adjustment* (WBA) or *local bundle adjustment* and consists in optimising only the latest camera poses, in a certain window, and the corresponding features that have been observed in that interval.

III. Experimental validation

The performance of the proposed monocular VO framework is evaluated with visible and infrared thermal images on a series of image sequences representing real driving conditions. The images were acquired from a car where a BlueFOX (Matrix Vision) camera and a Tau2 (FLIR) camera were installed on the car roof. The car was driven in a semi-urban environment. The third sequence presents particularly challenging illumination conditions. The sun being low in the sky, it produced image glare in the visible images when facing it, and low-textured images were generated when driving in the shade of the trees. Ground truth was recorded by an MTi-G GNSS sensor (Xsens) which combines inertial data and GPS coordinates to produce an accurate pose estimate. Finally, camera calibration parameters were obtained using the multispectral calibration algorithm described in [28].

A. Feature tracking between modalities

To show the suitability of the proposed approach, the number of features retained after outlier rejection, at each iteration, is analysed. Table I shows the average and standard deviation of the number of inliers for each modality. New features are added at every iteration to keep the total number of features constant at 500 features. A first rejection is applied by computing the essential matrix in a RANSAC scheme. The inlier threshold applied is 1px. Then, the remaining features are triangulated and P3P is performed as explained

TABLE I: Number of inliers remaining after outlier rejection for visible, infrared and the modality selected at each iteration.

	Visible	Thermal	MMS-VO
	Sequence 1		
mean inliers per frame	296	303	322
standard deviation	67	57	52
	Sequence 2		
mean inliers per frame	338	356	369
standard deviation	63	54	47
	Sequence 3		
mean inliers per frame	206	221	254
standard deviation	85	79	76

TABLE II: Number of images selected by MMS-VO for each sequence.

	Sequence 1	Sequence 2	Sequence 3
Visible	60	427	124
Thermal	96	668	179

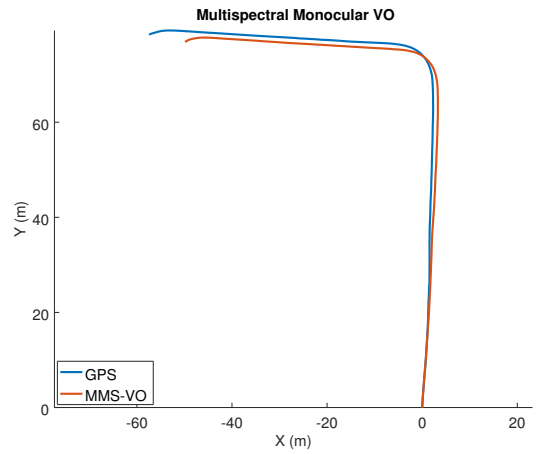
in Section II. The selected re-projection error threshold was 2px. This means that the inlier features at the end of the process satisfy the epipolar constraint between the last and current keyframe, but they are also re-projected correctly in the current image with an accuracy of 2px. The third column of Table I corresponds to MMS-VO, where both multispectral images are processed concurrently and the modality providing the highest number of inliers is selected to perform bundle adjustment.

As shown in Table I, more than half the total number of points is retained for the first two scenarios, offering a high number of strong features for subsequent motion estimation. As for the third sequence, where illumination conditions are challenging, the number of inliers remain important with more than 40% of the 500 original features retained. Additionally, Table II highlights the complementarity of visible-thermal images in the proposed multispectral solution as it can be seen that there is no camera that always produces the best matches. Instead, each modality is selected alternately as the vehicle moves and the environment changes. Nevertheless, it can be noticed that infrared images are selected more often.

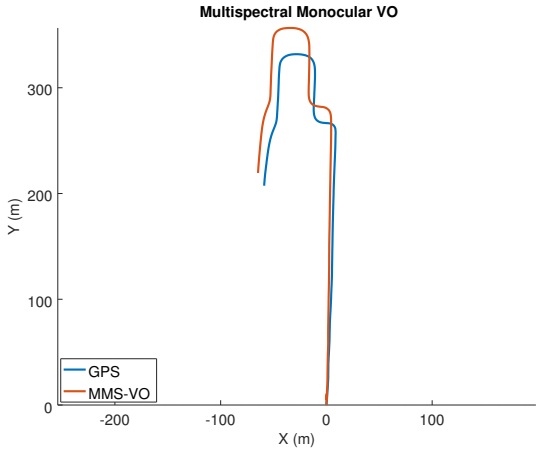
B. Multispectral Monocular VO

Errors are evaluated by computing the Euclidean distance between each estimated camera position and the closest GPS measurement in time. Because of the error accumulation, two types of errors are considered. The first one is the final position of the trajectory, which shows how far the estimation ends up compared to the ground truth, and the second one is the mean error which shows the quality of the estimation over the whole trajectory. These results are summarised in Table III.

The scale ambiguity is not addressed here because it exists several ways to solve it. Instead, the focus of this paper is to show the suitability of p-LK and bundle adjustment for both visible and infrared images. Therefore, a global scale is estimated over the first 10 frames, by computing the ratio between the 2D translation vector obtained from GPS measurements and the 2D translation vector obtained from BA.



a: Sequence 1.



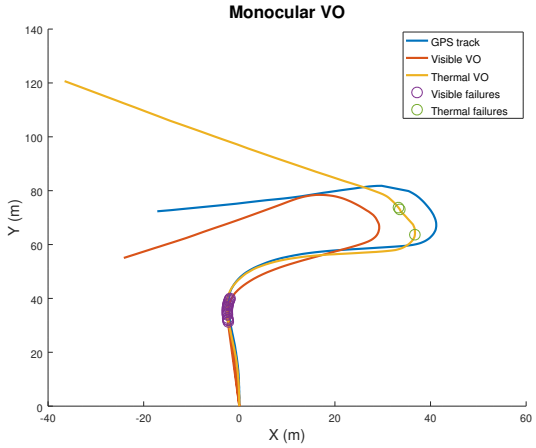
b: Sequence 2.

Fig. 2: Trajectories generated by MMS-VO on Sequence 1 and 2.

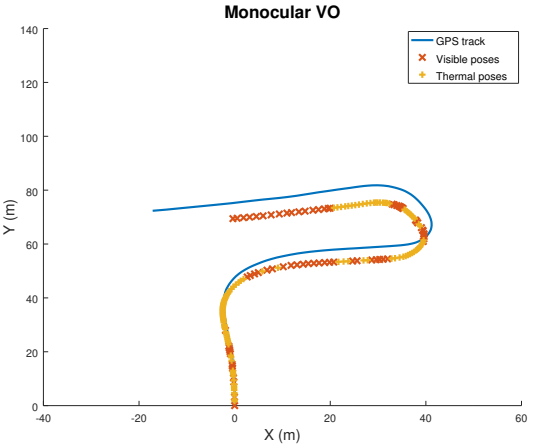
Even though some drift can be noticed in Fig. 2, the overall trajectories, for sequence 1 and sequence 2, are estimated properly. A final positioning error not exceeding **14m** after almost **500m** traveled and a mean error of **2%** the distance traveled on average are achieved. These results demonstrate the suitability of the proposed multispectral approach for visual odometry. It is hard to compare such multispectral technique with others since most state-of-the-art works were designed for stereo visible images or fused with inertial data in a SLAM framework. Nonetheless, the results presented are not so far from the one obtained by state-of-the-art algorithms on the Kitti dataset [29]. Because it utilises a simple BA framework to compute motion, it could still be enhanced in order to improve the odometry accuracy. However, what is demonstrated here is the suitability of both visible and thermal feature tracking for motion estimation.

TABLE III: Errors comparison between MMS-VO and the ground truth (GNSS).

	Sequence 1	Squence2
distance traveled	133m	494m
Error on final point (m)	7.70	13.37
Mean error (m)	1.98	12.45
Error on final point (%)	5.79	2.71
Mean error (%)	1.48	2.52



a: Independent trajectory estimation on Sequence 3.



b: Alternating trajectory estimation.

Fig. 3: Trajectories generated by thermal VO, visible VO and MMS-VO in Sequence 3.

C. Failure recovery

In this section, the advantage of multispectral monocular VO is further demonstrated through the results obtained by the proposed algorithm on the third dataset (sequence 3) with challenging illumination conditions. The trajectories obtained by visible VO and thermal VO are shown in Fig. 3a. It can be seen that both modalities fail to estimate motion for a few frames because the quality of the images deteriorates.

An example of images that caused these failures can be seen in Fig. 4. The visible camera is not able to detect and track features properly as it is facing the sun and the top-right corner of the image becomes excessively bright while the rest of image is completely dark. The thermal camera on the other hand, produces images of particularly low contrast when facing trees. Some features are detected in the right side of the images but they are too few and too close to each other to have a meaningful representation of the movement of the camera, making motion estimation impossible. Moreover, due to the accumulative nature of VO, these failures are affecting the rest of the trajectory which is then drifting away from the GPS track (see Fig. 3a).

TABLE IV: Errors comparison between each modality and the alternating technique on Sequence 3.

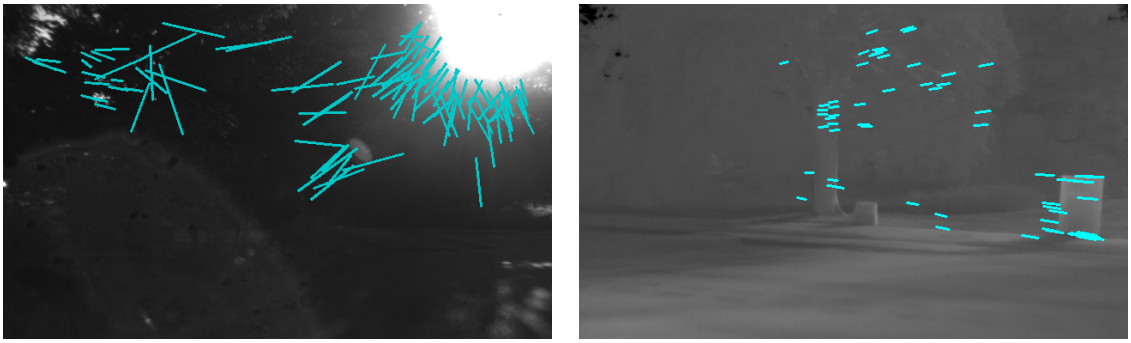
	Visible	Thermal	Multispectral
distance traveled	165m		
Error on final point (m)	18.68	55.12	**11.33**
Mean error (m)	9.67	6.81	**4.16**
Error on final point (%)	11.32	33.4	**6.87**
Mean error (%)	5.86	4.13	**2.52**

However, with a calibrated multispectral setup, MMS-VO makes it possible to continue performing VO when one of the cameras fails to produce decent images, as long as the other one captures enough details to run the p-LK method. This would be impossible with a single camera or with stereo cameras of the same modality as both sensors would be affected in the same way. To fully take advantage of the multispectral setup, a single VO trajectory is computed using each camera alternatively, instead of producing two trajectories corresponding to each modality. Fig. 3b shows the trajectory obtained with MMS-VO. Although a notable drift can be seen, the trajectory is fully recovered, even when one of the cameras fails to produce adequate images. This is confirmed by the analysis of the errors obtained (see Table IV), as each modality alone produces larger errors.

It must be kept in mind that this solution is only valid when p-LK is unsuccessful with one of the cameras. This would not be possible if both cameras fail to produce satisfying images at the same time. Nevertheless, the experiment shown demonstrates the superiority of MMS-VO over single band monocular VO solutions for challenging illumination conditions or low contrast environments.

D. Processing time

To show the suitability of the solution for real-time applications, the time spent to process each multispectral pair was measured and the results are gathered in Table V. The program was tested on a computer with an Intel i5-2410M CPU running at 2.30GHz on a Linux desktop environment. Because the task is identical for each modality, the program developed is multi-threaded and each image is processed in a separate thread. As it can be seen in Table V, the average processing time is around **60ms**. Thus, the program can handle images incoming at 15 fps. It can be noticed that the standard deviation is much more important for the last dataset. This is because it takes more time to reject outliers when tracking becomes less accurate in



a: Visible fails. b: Infrared fails.

Fig. 4: Example of images where p-LK fails and the corresponding p-LK tracking results. Each blue line represents the location of a feature in the current and the previous frame.

TABLE V: Time spent to process a multispectral image pair.

	Sequence 1	Sequence 2	Sequence 3
mean (ms)	65.7	57.3	60.8
std. dev. (ms)	11.4	9.2	30.7

a certain modality. Nevertheless, images could be queued if this happens, as the average processing time on that particular dataset is still below the 66ms time interval required to run at 15 fps. Additionally, the program does not always need to run at full speed because some frames are skipped if the parallax constraint is not reached. In this experiment 500 features where processed at each iteration but this number could be increased if using a more recent and more powerful processor. On the contrary, 500 points is substantial, so it could be reduced on embedded systems to reach the level of real-time required.

IV. Conclusion

A multispectral monocular VO system for visible and thermal images was proposed. Based on a keyframe approach, it automatically tracks features in every incoming image but only selects those presenting the best geometric properties. It then triangulates features over a certain window and refines both the structure and camera poses in a local BA process. It was demonstrated that this method can perform accurate monocular visual odometry in both visible and infrared domains with the same set of parameters.

Additionally, it was shown that, by evaluating the quality of the VO solutions in each modality and adaptively selecting the best performing result, MMS-VO is able to deal with failure cases, where it is impossible to track or estimate motion in one of the modalities. The proposed algorithm can be executed in real-time, allowing multispectral images to be processed simultaneously and the best modality to be selected for accurate visual odometry. Even though the number of datasets presented in this paper is limited, additional work could be carried out to highlight the benefits of MMS-VO.

Acknowledgment

The authors would like to acknowledge Dr Antonio Scannapieco for his help to acquire the multispectral datasets presented in this paper.

References

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge: Cambridge University Press, 2004.

[2] M. I. A. Lourakis and A. A. Argyros, "SBA: A software package for generic sparse bundle adjustment," *ACM Transactions on Mathematical Software*, vol. 36, no. 1, pp. 1–30, 2009.

[3] F. Fraundorfer and D. Scaramuzza, "Visual Odometry : Part II," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 78–90, 2011.

[4] B. Triggs, P. F. McLauchlan, R. I. Hartley *et al.*, "Bundle Adjustment A Modern Synthesis," *Vision Algorithms: Theory and Practice*, vol. 1883, pp. 298–372, 2000.

[5] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[6] R. Wang, M. Schworer, and D. Cremers, "Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras," in *IEEE International Conference on Computer Vision*, 2017, pp. 3923–3931.

[7] D. Scaramuzza and F. Fraundorfer, "Visual Odometry [Tutorial]," *IEEE Robotics & Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[8] K. L. Ho and P. Newman, "Loop closure detection in SLAM by combining visual and spatial appearance," *Robotics and Autonomous Systems*, vol. 54, no. 9, pp. 740–749, 2006.

[9] V. Ila, L. Polok, M. Solony *et al.*, "SLAM++ -A highly efficient and temporally scalable incremental SLAM framework," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 210–230, feb 2017.

[10] I. Esteban, L. Dorst, and J. Dijk, "Closed Form Solution for the Scale Ambiguity Problem in Monocular Visual Odometry," *International Conference on Intelligent Robotics and Applications*, vol. 6424, pp. 665–679, 2010.

[11] L. Kneip, D. Scaramuzza, and R. Siegwart, "A Novel Parametrization of the Perspective-Three-Point Problem for a Direct Computation of Absolute Camera Position and Orientation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 2969–2976.

[12] X.-S. Gao, X.-R. Hou, J. Tang *et al.*, "Complete solution classification for the perspective-three-point problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 8, pp. 930–943, 2003.

[13] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.

[14] P. Viola and W. M. Wells, "Alignment by Maximization of Mutual Information," *International Journal of Computer Vision*, vol. 24, no. 2, pp. 137–154, 1997.

[15] O. Yilmaz, N. Aouf, L. Majewski *et al.*, "Using infrared based relative navigation for active debris removal," in *10th International ESA Conference on Guidance, Navigation & Control Systems*, Salzburg, 2017.

[16] T. Mouats, N. Aouf, L. Chermak *et al.*, "Thermal Stereo Odometry for UAVs," *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6335–6347, 2015.
[17] S. J. Krotosky and M. M. Trivedi, "Mutual information based registration of multimodal stereo videos for person tracking," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 270–287, 2007.
[18] T. Mouats, N. Aouf, A. D. Sappa *et al.*, "Multi-Spectral Stereo Odometry," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1–15, 2015.
[19] J. Poujol, C. A. Aguilera, E. Danos *et al.*, "A Visible-Thermal Fusion Based Monocular Visual Odometry," in *Advances in Intelligent Systems and Computing*, 2016, vol. 417, pp. 517–528.
[20] A. Beauvisage, N. Aouf, and H. Courtois, "Multi-spectral visual odometry for unmanned air vehicles," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Budapest, Hungary: IEEE, oct 2016, pp. 1994–1999.
[21] J. Y. Bouguet, "Camera Calibration Toolbox for Matlab," 2000.
[22] T. Mouats, N. Aouf, and M. A. Richardson, "A Novel Image Representation via Local Frequency Analysis for Illumination Invariant Stereo Matching," *IEEE Transactions on Image Processing*, vol. 24, no. 9, pp. 2685–2700, 2015.
[23] J. Shi and C. Tomasi, "Good Features to Track," in *IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.
[24] D. Nister, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, jun 2004.
[25] C. V. Stewart, "Robust Parameter Estimation in Computer Vision," *SIAM Review*, vol. 41, no. 3, pp. 513–537, jan 1999.
[26] A. Eudes and M. Lhuillier, "Error Propagations for Local Bundle Adjustment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2411–2418.
[27] A. Eudes, S. Naudet-Collette, M. Lhuillier *et al.*, "Weighted Local Bundle Adjustment and Application to Odometry and Visual SLAM Fusion," in *British Machine Vision Conference*, 2010, pp. 1–10.
[28] A. Beauvisage and N. Aouf, "Low cost and low power multispectral thermal-visible calibration," in *2017 IEEE SENSORS*. Glasgow, UK: IEEE, oct 2017, pp. 1–3.
[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.