Proceedings Paper:
Rugg-Gunn, D. and Aitken, J.M. orcid.org/0000-0003-4204-4020 (2022) Investigating scene visibility estimation within ORB-SLAM3. In: Pacheco-Gutierrez, S., Cryer, A., Caliskanelli, I., Tugal, H. and Skilton, R., (eds.) Towards Autonomous Robotic Systems: 23rd Annual Conference, TAROS 2022, Culham, UK, September 7–9, 2022, Proceedings. 23rd Annual Conference (TAROS 2022), 07-09 Sep 2022, Culham, UK. Lecture Notes in Computer Science, LNAI 13546 . Springer International Publishing , pp. 155-165. ISBN 9783031159077

https://doi.org/10.1007/978-3-031-15908-4_13

Investigating Scene Visibility Estimation within ORB-SLAM3*

Dominic Rugg-Gunn, Jonathan M. Aitken

Department of Automatic Control & Systems Engineering, University of Sheffield
{drugg-gunn1, jonathan.aitken}@sheffield.ac.uk

Abstract. Scene Visibility Estimation offers a collection of metrics that give a good indication of the quality of images that are being supplied to a Visual or Visual-Inertial Simultaneous Location and Mapping algorithm. This paper will investigate the application of these metrics during switching between camera and IMU-based localisation within the popular visual-inertial ORB-SLAM3 algorithm. Application of the metrics provides more flexibility compared to a static threshold and incorporating the metrics within the switch provides a reduction in the error in positioning.

Keywords: Simultaneous Localisation and Mapping · Visibility

1 Introduction

1.1 Background

Simultaneous Localisation And Mapping (SLAM) is a method used by mobile robots to construct a map of the surrounding environment and to estimate it's position within that map. It is now used in an increasing number of practical fields due to improvements with computation and sensing.

Many environments pose challenges to existing pose estimation strategies SLAM in conjunction with appropriate sensors, provides a strong alternative [1]. Cameras commonly used as sensors as they provide a large amount of information at a relatively low cost [14]. For this, algorithms which perform Visual-Inertial-SLAM (VI-SLAM), harness the localisation capabilities of Visual Odometry (VO) in conjunction with inertial measurements in order to estimate the robot's pose. VI-SLAM also performs a mapping process that tracks observed features relative to the agent, enabling a computational understanding of an unknown environment [4].

Cameras suffer from visual artefacts such as lens flares and occlusion, often reducing the reliability of the localisation and mapping. These artefacts are detected as features, and propagate into the algorithm as incorrect associations, significantly impairing performance [2].

* Supported by the Department of Automatic Control and Systems Engineering at the University of Sheffield. Also this work is supported by the UK's Engineering and Physical Sciences Research Council (EPSRC) Programme Grant EP/S016813/1

Previous work [7] developed Scene Visibility Estimation (SVE) metrics, that use a camera feed to assess the quality of the visual data and the visibility of the scene captured. This paper will implement SVE in a state-of-the-art VI-SLAM algorithm and use the metrics to evaluate performance with the aim of generating a more accurate and reliable pose estimation.

2 Related Work

2.1 Visual Simultaneous Localisation and Mapping

Visual Simultaneous Localisation and Mapping (V-SLAM) is a subsection of SLAM focused on using cameras as the sensor input to the algorithm. This incorporates functionality from Visual Odometry (VO) and combines it with the map generation capabilities inherent in SLAM algorithms [12]. VO works by isolating features in each frame and tracking the movement of these through multiple frames to infer the motion the agent must have taken.

V-SLAM is an advancement upon VO, introducing capabilities such as loop-closure which reduces the drift suffered in pose estimation, by referencing features to past features in the map and adjusting accordingly [1].

ORB-SLAM2 is a prominent V-SLAM algorithm, but is indirect as it includes an additional feature identification step. Built upon ORB-SLAM, it uses an ORB feature detector to incorporate clusters of pixels into a feature description before selecting features to use in the algorithm [11]. This is often faster and more data efficient than direct (individual pixel) methods, though some solutions such as LDSO use feature extraction methods to aid pixel selection and can produce results comparable to indirect methods [9].

2.2 Visual-Inertial Simultaneous Localisation and Mapping

VI-SLAM incorporates an IMU sensor to the V-SLAM pipeline. This adds precise measurements of movements and rotations but is prone to drift over extended periods of time, it therefore complements the V-SLAM algorithm well to increase system precision and robustness [5]. Two of the most significant VI-SLAM algorithms are VINS-Mono and ORB-SLAM3 [6].

VINS-Mono (monocular Visual Inertial Navigation System) is a tightly coupled mono VI-SLAM algorithm. Using a single camera and IMU, it considers the coupling of all sensors before pose estimation is performed. To optimise performance VINS-Mono pre-integrates the IMU data between frames, which reduces computation of superfluous pose estimation nodes [10]. The algorithm was primarily designed for use on-board UAS for pose estimation, but has proven capable in many other fields such as the automotive and agricultural research areas [8][15].

ORB-SLAM3 is built upon ORB-SLAM2 with the integration of an IMU sensor. It supports a variety of camera types and configurations, as well as incorporating a multi-map strategy to increase robustness to poor quality visual

data (either from sensor errors or from a feature-sparse environment) [4]. ORB-SLAM3's superior performance is well noted [4, 13, 6] and motivates its use within this paper.

2.3 Scene Visibility Estimation

Scene Visibility Estimation [7] is a novel method to improve robustness by estimating the visibility of detected features. This was achieved on top of the ORB-SLAM2 algorithm by extracting and analysing tracked visual features. The results showed it responded reliably in difficult conditions such as fog, direct sunlight, and dirt on the lens.

To track the scene visibility, three metrics (S_{a-c}) were proposed, where each tracked a different aspect of the presented scene. The metric S_a represents a general ratio of the features N_F to the desired number of features $N_{F,\ max}$ such that: $S_a = \frac{N_F}{N_{F,\ max}}$. With a low S_a implying poor visibility, this would also be effective in low contrast operation or for handling feature sparse environments.

Component S_c broadly operates in the same way but using tracked features N_T. It does, however, incorporate an estimation step to determine the number of features that should be visible based on the local map N_{Lv}, and is defined as $S_c = \frac{N_T}{N_{Lv}}$. This should be effective in dynamic environments where tracked features a quickly and frequently lost between keyframes, however results for this didn't show a strong correlation .

The final metric (S_b) is the most complex. It aims to capture the distribution of features, the frame is divided into equal bins and a Chi-Square test is performed to quantify the homogeneity of the extracted features (χ^2). This result is then scaled to a 'worst case' scenario (χ^2_w) where all features appear in one eighth of the bins. When a part of the view is obscured any features in that section cannot be identified, however, estimating which previously tracked features should be identified can be used to identify feature absence. This presents as a skew in feature homogeneity, and propagates such that $S_b \to 0$ as $\chi^2 \to 0$ [7].

2.4 Adaptation of Scene Visibility Estimation in ORB-SLAM3

A modification was required to component A. The concept of this is that with decreased visibility the number of features extracted would decrease and lower the value of SVE_a. As ORB-SLAM3 almost always extracts the requested number of features, making this value uninformative. This metric was altered to express the ratio of tracked features (N_{F_T}) to the number required for high quality localisation expressed as a fraction of the desired number of features (N_{F_max}).

3 Disturbance Generation

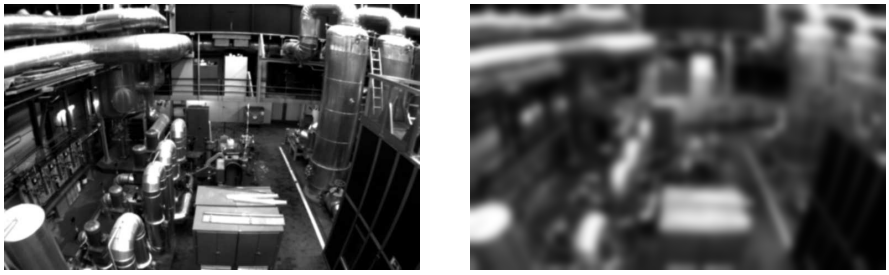
This section will discuss Blur, Downsampling and Occlusion disturbances, and how they can be applied to the data sequence to be used. These disturbances can then be used to corrupt elements of the EuRoC dataset [3] commonly used

with ORB-SLAM3 [4], allowing an insight to the response of ORB-SLAM3 to varying intensities of different visual disturbances.

3.1 Blur

Bluring of the image feed is an attempt to replicate fog and rain, which are both very common-place in the real world applications of SLAM for example in smoke, dust, and generally decreased visibility in underwater situations.

The blur was implemented via OpenCV's Gaussian Blur, which generates a Gaussian kernel with dimensions $n \times n$ and convolves the input image with this. To vary the strength of the blur, the kernel dimensions are varied leading to increased blurring of the image with higher n. In preliminary testing, all localisation attempts by ORB-SLAM3 failed before the blur kernel size reached 50 pixels. To best explore the range before this point, a step-size of 5 was selected, giving a testing set of 11 intensities with dimensions $n \times n$ for $n \in [0, 5...45, 50]$. Examples of these kernel sizes can be seen in Figure 1, with a weak blur shown in Figure 1a and a strong shown in Figure 1b.



(a) Kernel Size $n = 5$ (b) Kernel Size $n = 50$

Fig. 1: Example Frames with Different Blur Kernel Sizes

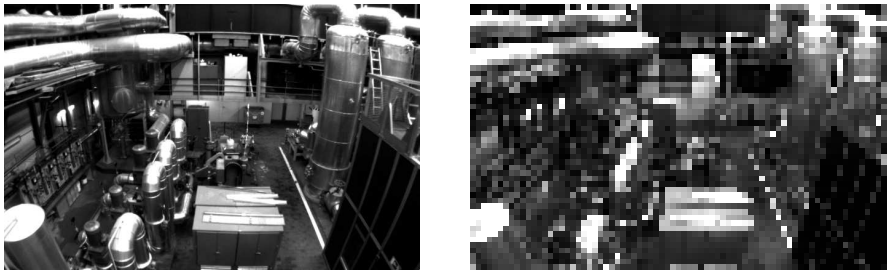
3.2 Downsampling

The second augmentation performed was resolution downsampling. This was chosen partially to present an alternative implementation to blur for low visibility scenarios. In addition, image resolution is important to consider in it's own right as it is closely linked to hardware costs, with higher resolution cameras costing significantly more. This augmentation helps to investigate the extent to which a lower resolution, low cost, cameras impact the performance of the localisation in ORB-SLAM3.

The resolution downsampling was implemented via OpenCV's Resize, which is used to resamples the image array according to a provided downsampling factor (applied to both x and y dimensions). Preliminary tests were conducted

that identified a downsampling factor of 1/0.1 was the point at which the initialisation failed, which is a resolution of just 75×48 px. Ten sample points were selected from the range [1,0.1] in order to cover it with a reasonable level of granularity. This resulted in the final downsampling factors to be tested of $1/[1.0, 0.9...0.2, 0.1]$.

Examples of the frames at these resolutions can be seen in Figure 2 where Figure 2a shows the smallest downsampling factor, taking the image to a resolution of 676×432 px. Figure 2b depicts the largest amount of downsampling (enlarged for visibility), and is the point at which the resolution is too low for the system to initialise.



(a) Downsampling Factor $DF = 1/0.9$ (b) Downsampling Factor $DF = 1/0.1$

Fig. 2: Example Frames with Different Downsampling Factors

3.3 Occlusion

The third disturbance selected was Occlusion. The ORB-SLAM3 system incorporates a place recognition module from ORBSLAM2, which is designed to resume tracking after temporal occlusions. In this test set the occlusion is static and persistent throughout the sequence, in order to investigate the systems robustness to scenarios with dirt or rain on the camera lens. To achieve this, a black square is generated in the center of the image as though an object was stuck to the lens of the camera. This appear in the centre of the image and different intensities are generated by adjusting the region's dimensions to produce a $n \times n$ absence of data. The smallest dimension of the image was 480 px, so the upper bound for n was chosen to be 450 px as this could be easily divided into ten steps to produce the final range of $n \in [50, 100...400, 450]$. Examples of the extremes of the occlusion disturbance can be seen in Figure 3 with the smallest and largest occlusion sizes are shown in Figure 3a and Figure 3b respectively.

(a) Occlusion Size $n = 50$ (b) Occlusion Size $n = 450$

Fig. 3: Example Frames with Different Occlusion Sizes

3.4 Data Aggregation

In summary, three simple disturbances are applied to a baseline sequence from the EuRoC dataset. These each have 9-10 different levels, on which ORB-SLAM3 was be tested 100 times in order to aggregate the results and reject anomalies in the system's performance.

4 Improving SVE and Adapting ORB-SLAM3

The original ORB-SLAM3 pipeline incorporates a switching-behaviour for using visual or inertial data which is based on the number of features extracted from the image, typically set to 15 with the IMU initialised, or 50 when not. This is used as a primitive equivalent of the SVE, used to dictate to the system whether the image localisation should be used or the IMU odometry.

To test the current state of behaviours these evaluations were replaced with comparisons to the SVE using thresholds of 0.1 and 0.3 respectively. With little to no disturbances applied, the majority of the SVE ($\mu \pm \sigma$) lies above 0.3 so this was chosen to be a reasonable upper threshold. The lower threshold of 0.1 was chosen from manually inspecting how the SVE responds to the visually challenging portions of the sequence.

Combining the insights learnt so far, the components of the SVE metric appear to work well to indicate the visibility. However the process of combining them does not focus on the most important aspects for ORB-SLAM3's switching-behaviour. The equation currently used to do this is outlined in Equation (1), where it can be seen that the number of extracted features (represented in SVE_a) only constitutes 20% of the overall metric. In order to improve the performance of the ORB-SLAM3 pipeline when the SVE metric is used to govern the switching-behaviour, two improvements are proposed to weight SVE_a more heavily.

$$\text{SVE} = 0.2 \times \text{SVE}_a + 0.4 \times \text{SVE}_b + 0.4 \times \text{SVE}_c \tag{1}$$

The first of improvement proposal is outlined in Equation (2), and alters the weightings such that component A bears twice the weighting of either com-

ponents B or C, which are equally weighted. This keeps to the same structure in [7], but seeks to redistribute the influence such that the switching-behaviour is more accurate and robust.

$$\text{SVE} = 0.5 \times \text{SVE}_a + 0.25 \times \text{SVE}_b + 0.25 \times \text{SVE}_c \tag{2}$$

The alternate improvement deviates from the original structure as shown in Equation (3), making the number of tracked features of paramount importance to dictate the scene's visibility. This structure uses the SVE_a metric as a scaling factor for the remaining metrics, and returns this result to a linear response by taking the square-root which results in a behaviour similar to a geometric mean.

The thresholds for this method were determined by scaling the original thresholds of 15 and 50, with best visibility (SVE = 1) set to occur with 250 tracked features, the original thresholds were divided by this to produce values of 0.04 for 15 features and 0.2 for 50.

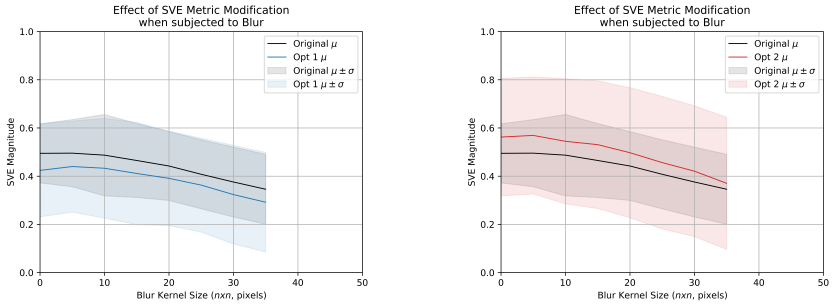
$$\text{SVE} = \sqrt{\text{SVE}_a \times (0.5 \times \text{SVE}_b + 0.5 \times \text{SVE}_c)} \tag{3}$$

5 Results

To analyse the performance of the two options for improving the SVE metrics, each was built into separate ORB-SLAM3 instances with the thresholds described in the previous section. These two ORB-SLAM3 builds were then run with each of the 31 sequences 100 times.

5.1 SVE Improvements

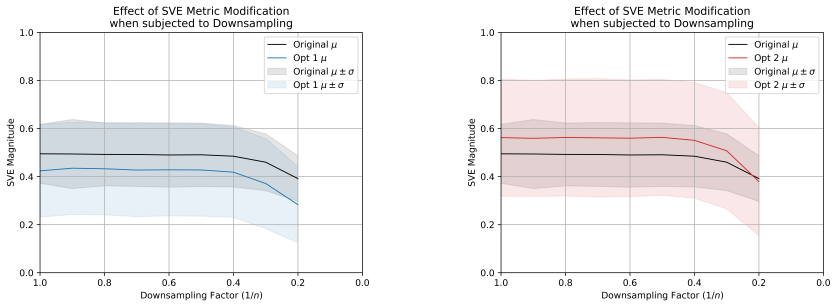
The plots of the SVE metrics are shown in Figure 4, Figure 5, and Figure 6, showing the responses to each of the different disturbances.



(a) Option 1 (b) Option 2

Fig. 4: Comparison of Modified SVE Metrics Subjected to Blur

The modified metrics showed the least difference from the original metric when subjected to the Blur disturbances, the results of which can be seen in Figure 4. These mean values for both metric options are almost completely parallel to the original for all intensities. Which indicates they generally agree on the relative visibilities of the different blur intensities. The larger standard deviation illustrates a greater variation of the metric throughout each test run, so it is likely that the modified metrics provide a more varied and accurate indication of the true visibility throughout the sequence.

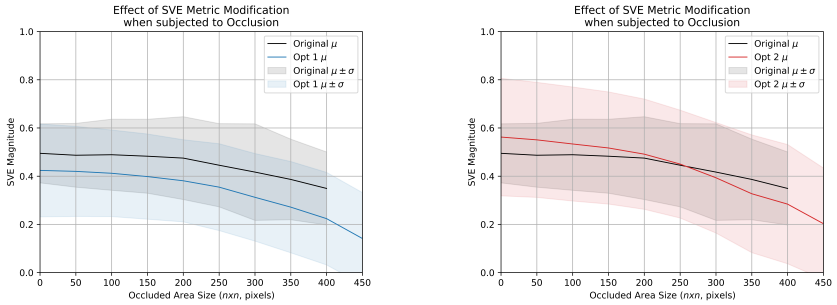


(a) Option 1 (b) Option 2

Fig. 5: Comparison of Modified SVE Metrics Subjected to Downsampling

The responses to downsampling shown in Figure 5 also exhibit the increased standard deviation of the modified metrics, with this being continually demonstrated throughout the sequence. As was seen in responses to the Blur, the mean value of both metrics stay parallel up to a downsampling factor of 1/0.5. A this point Option 2 drops first, but it shortly followed by Option 1 and the Original at 1/0.4, however both the modified versions do so with a steeper gradient. This shows an increased sensitivity to poor visibility conditions which is even more exaggerated in Option 2.

Figure 6 shows some of the most significant improvement with the modified metrics. Option 1 demonstrates an increased sensitivity to poor visibility by diverging from the unmodified equivalent from the lowest occlusion size and continuing this through to the highest. It also exhibits the increased standard deviation making it likely to be identifying high and low visibility areas more effectively throughout the individual sequences. This shows a very promising step towards representing the true scene visibility, as it is known that there is disturbance at this point yet neither of the other two metrics indicate this.



(a) Option 1 (b) Option 2

Fig. 6: Comparison of Modified SVE Metrics Subjected to Occlusion

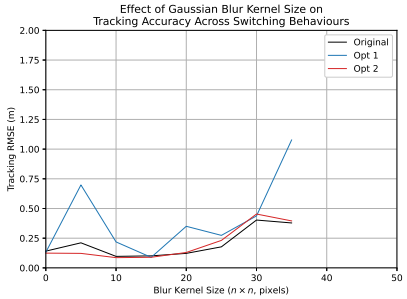
5.2 Adapted ORB-SLAM3 Results

With the original ORB-SLAM3 switching algorithm as the baseline, Figure 7a, Figure 7b, and Figure 7c show comparisons between this baseline and to two pipelines with the SVE-based switching-behaviour implemented.

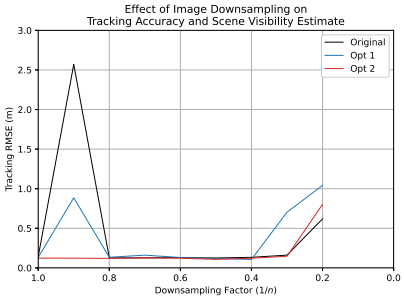
Figure 7a shows the relative performance of the localisation as the Blur kernel size increases. It can be seen that all algorithms completely fail beyond a kernel size of 35, indicating the first steps in successfully stopping visual localisation when scene visibility is severely affected. Throughout the lower ranges Option 1 displays poor performance, multiple points deviate significantly from the baseline and almost always have greater RMS error. By contrast, Option 2 exhibits very good performance with little deviation from the performance of ORB-SLAM3's Original switching-behaviour, and strongly reflects the expected general trend.

With Downsampling applied to the visual data, Figure 7b shows a generally similar performance between all three systems, this is most evident when DF = $1/0.8 \rightarrow 1/0.4$. Both the Original and Option 1 switching-behaviours show spikes at DF = $1/0.9$, this is due to a couple of smaller outliers not being rejected which appear to be the same reason for the peak in Option 1. Interestingly, Option 2 seems to be more robust to this with no unexpected spikes observed due to outliers in either the blur or downsampling tests.

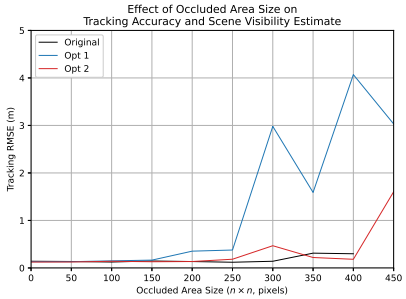
The final filter applied was occlusion, the results of which can be seen in Figure 7c. In this it can be seen that Option 1 is again showing relatively poor rejection of bad tracking data, with the SVE metric failing to trigger adequate switching in order to prevent failures within the localisation process. The performance of this remained in line with the other two systems until an occluded area of 150×150 px, at which point the error began to increase before sharply failing beyond 250×250 px. Option 2's performance remained much closer to the Original results, with a slight peak likely due to outliers at 300×300 px. Notably, neither Option 1 or Option 2 were successful in preventing tracking at 450×450. The Original failing to reach this point indicates that the number of tracked features was very low which would make localisation estimates poor.



(a) Subjected to Blur



(b) Subjected to Downsampling



(c) Subjected to Occlusion

Fig. 7: Comparison of Original ORB-SLAM3 Switching Behaviour and Proposed SVE-Based Switches

6 Conclusions and Future Work

This paper has investigated the application of Scene Visibility Estimation within ORB-SLAM3. The EuRoC dataset has been corrupted using a collection of techniques that produce effects analogous to blur, occlusion and downsampling. The SVE metrics have been implemented within ORB-SLAM3, and used to provide an adaptive threshold for switching between localisation using either the camera or IMU feeds. SVE provides a more complete set of measurements of the quality of the image feed, and this allows the switching process to be more efficient resulting in a more accurate localisation when compared to the standard ORB-SLAM3 baseline. Option 2 performed very well, being consistently similar to the original pipeline and in some cases out performing it. Future work will focus on the optimisation of the mix of the SVE metrics to minimise position error.

References

1. Aitken, J.M., Evans, M.H., Worley, R., Edwards, S., Zhang, R., Dodd, T., Mihaylova, L., Anderson, S.R.: Simultaneous localization and mapping for inspection

robots in water and sewer pipe networks: A review. IEEE access **9**, 140173–140198 (2021)
2. Bailey, T., Durrant-Whyte, H.: Simultaneous localization and mapping (slam): part ii. IEEE robotics & automation magazine **13**(3), 108–117 (2006)
3. Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The euroc micro aerial vehicle datasets. The International Journal of Robotics Research **35**(10), 1157–1163 (2016)
4. Campos, C., Elvira, R., Rodríguez, J.J.G., M. Montiel, J.M., D. Tardós, J.: Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. IEEE Transactions on Robotics pp. 1–17 (2021)
5. Chen, Y., Zhou, Y., Lv, Q., Deveerasetty, K.K.: A review of v-slam*. In: 2018 IEEE International Conference on Information and Automation (ICIA). pp. 603–608 (2018)
6. Cheng, J., Zhang, L., Chen, Q., Zhou, K., Long, R.: A fast and accurate binocular visual-inertial slam approach for micro unmanned system. In: 2021 IEEE 4th International Conference on Electronics Technology (ICET). pp. 971–976 (2021)
7. Haggart, R., Aitken, J.M.: Online scene visibility estimation as a complement to slam in uavs. In: Towards Autonomous Robotic Systems, pp. 365–369. Lecture Notes in Computer Science, Springer International Publishing, Cham (2021)
8. He, Y., Chai, Z., Liu, X., Li, Z., Luo, H., Zhao, F.: Tightly-coupled vision-gyro-wheel odometry for ground vehicle with online extrinsic calibration. In: 2020 3rd International Conference on Intelligent Autonomous Systems (ICoIAS). pp. 99–106 (2020)
9. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5), 1255–1262 (2017)
10. Qin, T., Li, P., Shen, S.: Vins-mono: A robust and versatile monocular visual-inertial state estimator. IEEE Transactions on Robotics **34**(4), 1004–1020 (2018)
11. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International Conference on Computer Vision. pp. 2564–2571 (2011)
12. Savaria, D.T., Balasubramanian, R.: V-slam: Vision-based simultaneous localization and map building for an autonomous mobile robot. In: 2010 IEEE Conference on Multisensor Fusion and Integration. pp. 1–6 (2010)
13. Sharafutdinov, D., Griguletskii, M., Kopanev, P., Kurenkov, M., Ferrer, G., Burkov, A., Gonnochenko, A., Tsetserukou, D.: Comparison of modern open-source visual slam approaches (2021)
14. Zaffar, M., Ehsan, S., Stolkin, R., Maier, K.M.: Sensors, slam and long-term autonomy: A review. In: 2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS). pp. 285–290 (2018)
15. Zhou, S., Zhao, H., Chen, W., Liu, Z., Wang, H., Liu, Y.H.: Dynamic state estimation and control of a heavy tractor–trailers vehicle. IEEE/ASME Transactions on Mechatronics **26**(3), 1467–1478 (2021)