# Performance analysis of Simultaneous Localization and Mapping to reconstruct aircraft engines in 3D

**Thomas Markhorst**[1] , **Jan van Gemert**[1] , **Burak Yildiz**[1]

[1]TU Delft

## Abstract

Proper maintenance and inspection of aircraft and their engines is important for society. These engine inspections are performed using borescopes of which the footage is manually analysed. Having the opportunity to reconstruct a 3D model of the rotors would ease the inspection and introduce the possibility to automate the process. Monocular SLAM systems are capable of reconstructing such models in real-time using a video of the rotors. However, SLAM is not tested in environments similar to the aircraft turbine. This study, therefore, assesses the performance of different SLAM approaches in this specific setting. The results show that 3D reconstruction of aircraft engines using direct SLAM has potential for damage assessment. Further research into damage assessment using SLAM is therefore viable.

## 1 Introduction

Inspections of aircraft turbines can both be sped up and simplified by automating part of the process, which is likely to positively affect the safety of aeroplanes. These inspections are performed using a borescope, a monocular camera on a semi-rigid body, that is used to reach and film narrow spaces. In a turbine, this would be the space between one of the approximately 40 blade sets, where the camera is manoeuvred to assess the blades on damage, see Figure 1. Using the footage of the camera to reconstruct a 3D model of such a blade set, opens the possibility to partially automate the measurement of damage on the blades. Video-based reconstruction could be done using Visual Simultaneous Localization and Mapping (VSLAM), which is an industry standard for modelling a scene and keeping track of the camera's position within that scene. It estimates the depth of the scene using relative motion between the object and the camera based on the parallax effect. The parallax effect states that objects close to the observer move more than distant ones when the perspective changes. These depth estimations are made for every frame and are processed to improve existing or add new parts to the reconstruction. Said process is hindered by the lack of texture and shininess of the metallic blades. Using SLAM instead of for example Structure from Motion, which can reconstruct denser models, allows reconstructing the blades in real-time. Therefore in addition to damage assessment, the real-time constructed model could be used by the inspector to firstly localize the borescope in the complex internals of the engine and secondly keep track of how many blades have been assessed. Aiir[1], a company specialising in automating processes in the aviation industry, has expressed interest in this study and therefore assists by providing borescope inspection videos.

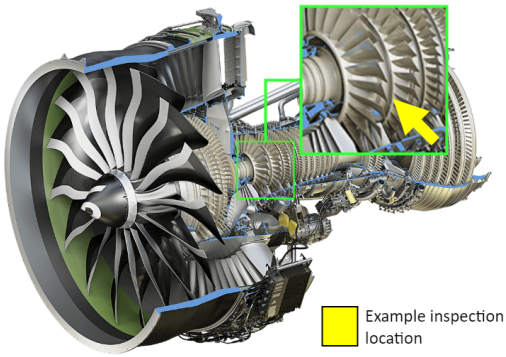



Figure 1: GE9x Commercial Aircraft engine (i.e. Boeing 777) Source: https://geaviation.com

Many different algorithms like [1, 2] introduce VSLAM methods and show the feasibility of running VSLAM on general scenes. These algorithms are evaluated using datasets like [3] where scenes are not comparable to borescope inspections as they are generally not shiny and textured. Therefore, the performance analysis of VSLAM on data with these properties, which characterises borescope inspection videos, is an unvisited topic. Hence this study will try to answer:

- *How well does monocular SLAM work on borescope inspection videos of aircraft turbines?*

The main contribution of this paper is a performance analysis of different SLAM methods on borescope inspections of aircraft turbines. In addition, the traditional feature matcher

[1]Aiir Innovations, Bringing Artificial Intelligence to Aviation: https://aiir.nl/

in a popular SLAM system is replaced by more recent neural-network-based approaches that are more fit to find matches in scenes with a lack of texture. This adapted system is tested and shown to outperform the original implementation. But in turn, is outperformed by methods that do not use feature matching.

The overall structure of this study takes the form of seven chapters, including this Introduction. Chapter 2 highlights relevant related work, followed by the method in Chapter 3 in which a general outline of the analysis approach is given. In Chapter 4 the specific experiments are described in detail and their results are displayed. Then, Chapter 5 reflects on the reproducibility of the study. The results are used in Chapter 6 for analysis and future recommendations. The paper is concluded in Chapter 7.

## 2 Related work

Within VSLAM there are several approach divisions. The first split is the approach for geometric estimation of the surroundings, the alternatives are indirect and direct SLAM. Indirect SLAM first extracts a set of features, interest points such as corners, from the image to then estimate the camera position and scene geometry based on said features. The downside of this is that only points in the image that conform to the feature type will be used [1]. In contrast, direct methods compare pixel intensities in successive frames to estimate geometry directly, using the brightness consistency constraint [4]. This results in higher accuracy and robustness in environments with few features [1]. It is however not illumination invariant, which could prove to be an issue for the shiny metallic surfaces.

The second split in VSLAM methods is introduced by the high computational complexity of using all landmarks from every frame for geometry mapping and camera localization. Unlike offline Structure from Motion, this is not possible for SLAM which operates in real-time. This can be solved by using either a filter- or keyframe-based approach. In filter-based methods localization and mapping takes place every frame using all detected landmarks. Consequently the number of landmarks processed per frame decreases. While in keyframe-based (KF) methods localization and mapping are separated. Localization is done every frame using a subset of the landmarks and mapping happens on a subset of the frames using all landmarks. This use of subsets enables more landmarks to be detected and maintained per frame. In [5] Strasdat et al. showed that using keyframes for visual SLAM generally yields higher accuracy than filtering. This claim is based on the observation that it is more profitable to increase the number of features than the number of processed frames. For this reason, combined with the trend of recent publications focusing on KF approaches, this study will consider only VSLAM approaches using keyframes. Strasdat's conclusion might however not be valid for environments with a lack of distinct landmarks, studying this could be worthwhile in another research.

### 2.1 Indirect VSLAM

ORB SLAM is considered to be the most promising indirect SLAM method with a public implementation for this study. It is based on and improves upon most components of PTAM by Klein and Murray [6], which is the first system to separate localization and mapping. One of the extensions is the use of the BRIEF visual descriptor with the FAST corner detector already used in PTAM. This combination of detector and descriptor was first introduced by Rublee et al. in [7] as Oriented FAST and Rotated BRIEF (ORB). Said change in descriptor makes the system more invariant to changes in viewpoint and illumination. Especially the illumination invariance is an important improvement for this study, considering the shiny material of the blades. Furthermore, when tracking is lost PTAM tries to match the current frame with low-resolution thumbnails of previous keyframes, using the sum-square-difference between the images. This introduces a problem for the borescope videos as two frames showing different blades are likely to be matched as they look similar, but are not the same. ORB slam solves this problem by relocalizing based on the BRIEF descriptors. Combined with the findings in [8, 2] which show that ORB slam is more robust and accurate in versatile environments, this argues to use ORB slam for borescope inspection videos.

### 2.2 Direct VSLAM

In contrast to indirect methods, there are multiple recent publications on direct VSLAM that introduce approaches differing on crucial parts in their implementation and therefore differing in their output. Firstly, Large Scale Direct SLAM (LSD) that produces semi-dense 3D reconstructions and is shown to be a robust and versatile industry standard. It does however only use pixels in high gradient areas like edges for direct matching. This is not beneficial for turbine reconstruction, as the blades are essentially a smooth surface with 3 main edges [1]. Secondly, Direct Sparse Odometry (DSO) which contrary to its name produces semi-dense maps comparable to LSD. However, unlike LSD the pixels used for matching are distributed over the frame. This is achieved by splitting the frame into blocks, from which the most interesting pixel patches are used. As stated in [9] it can sample pixels from all image sections including "smooth intensity variations on essentially featureless walls", which could prove useful for matching the blades. Lastly, Semi-Direct Visual Odometry (SVO) which outputs a sparse reconstruction and unlike the latter two does use feature detection. Said detection using FAST is used to determine interesting areas of the frame for direct matching and is therefore classified as a hybrid, taking a step towards indirect approaches. This design choice makes SVO run at framerates over 300 Hz resulting in good performance under fast and variable motion, especially the former applies to our use case [10]. LSD, DSO, and SVO are interesting approaches to assess in this study, as all three have interesting components.

Apart from the difference in video scenes used in this study, it also differs from all papers listed above in the evaluation of reconstruction and mapping. As most studies quantitatively evaluate their systems using a ground truth of the camera path. The path computed by the systems on datasets like [11, 3] is then scored using a metric like the Root Mean Square Error, compared to the corresponding ground truth provided

by the dataset. Since we will evaluate the performance of VSLAM systems on borescope inspection videos that are not part of such dataset, ground truth is missing. Besides, as the camera does not move in our use case but the scene does, constructing such ground truth is not trivial. Therefore a different method of evaluation needs to be formulated for this study.

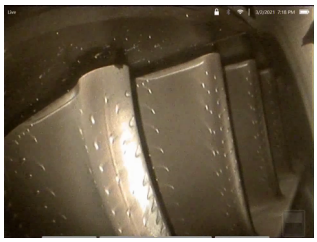
# 3 Methodology

The performance of each direct and indirect system will be evaluated on borescope inspection videos provided by Aiir. In the following sections, the evaluation method will be described.

## 3.1 Borescope inspection test data

Aiir provided us with a dataset of ten different borescope inspections of aeroplane turbines, out of these two were selected and trimmed to 15 seconds in duration. Since reconstructing models using SLAM on borescope videos is expected to be difficult, the videos from which the 3D structure was most clear for humans were selected, see Figure 2. To ensure coverage of more borescope videos, the two samples are selected to be different on key aspects. Such as the angle at which the videos were taken as well as the amount of texture and shape of the blades. Ideally, videos within the subset should only differ on one aspect (e.g. same blades and texture but different angle), this is however not feasible in this study due to time and data limitations.

(a) Example frame from video A, less texture and the camera perpendicular to the blades.



(b) Example frame from video B, more texture and the camera parallel to the blades.

Figure 2: Example frame from the test videos, showcasing the differences.

All discussed SLAM systems can process videos when structured like [3, 11] which are public data sets, we, therefore, convert the videos to the RGB-D TUM format [3]. Apart from formatting the video, the intrinsic parameters of the camera are also required to undistort the frames. Using a chessboard-based method to calculate these parameters is standard, this is however not possible as the cameras used for data collection are not available. Other proposed methods like [12, 13] need consecutive frames with a pure translation of the camera and use feature matchers, both of these properties will not work well in our use case. Even manual distortion removal in the frames will be speculative, as the geometry of the blades is unknown [14]. We therefore, have opted to set the distortion parameters for the cameras to zero in this study, treating the videos as undistorted.

## 3.2 Indirect VSLAM

ORB SLAM2[2] will be installed and used with the default settings to generate results for evaluation. In [15] it is shown that ORB performs relatively well on a car, a shiny and non-textured surface, and outperforms SIFT which is a standard for matching in many 3D reconstruction approaches. Still, the pitfall of indirect approaches remains a lack of features, hence it is interesting to analyse the performance when the matching algorithm is replaced with:

- recent approaches utilizing neural networks to match interest points between frames. Using SuperGlue [16] and LoFTR [17], which are found to perform significantly better in our use case compared to e.g. SIFT and ORB as discussed by Huizer in [18].
- a ground truth constructed using the method of Lieuw A Soe discussed in [19].

## 3.3 Direct VSLAM

All direct approaches, LSD[3], DSO[4] and SVO[5], can be installed from their respective repositories and used without modifications. As these methods are expected to handle the lack of features on the blades much better than indirect VSLAM, modifications to the matching are not implemented. Besides, changing the direct matching is less trivial than indirect matching.

## 3.4 Evaluation

As mentioned, the standard in evaluating the performance of SLAM systems is calculating the error in the reconstructed camera trajectory compared with ground truth. This suffices as the accuracy of localization is directly correlated with the quality of the model [20]. However, this approach is not possible for this study as such ground truth of camera positions does not exist and is non-trivial to generate. Even qualitative evaluation as proposed in [21] cannot be used as it is targeted for the evaluation of SLAM in rooms. Since all algorithms generate different results, e.g. dense or sparse, a metric has to be established that is not influenced by these differences. As a result, a new qualitative evaluation method is proposed. This evaluation is based on the observation that an optimal performing SLAM system will only initialise a model once and then track and extend this model for the entire duration of the video. However, when running SLAM on the borescope videos it becomes apparent that the model is frequently lost, after which the system resets the reconstruction or needs to be manually reset. Since it is likely that a model needs to be initialised multiple times in one video and will be tracked for a shorter duration than the video, the following measures are proposed:

1. Ability to initialise a model in which a blade can be recognized. Quantified by the number of times the system can do this after a reset.

[2]https://github.com/raulmur/ORB_SLAM2

[3]https://github.com/tum-vision/lsd_slam

[4]https://github.com/JakobEngel/dso

[5]https://github.com/uzh-rpg/rpg_svo

2. The maximum and average duration that a model can be extended with new features and the camera movement is tracked correctly. The camera movement is defined to be correctly tracked while the reconstructed path moves alongside the blades.
3. The constructed model can be used for: damage assessment, localization/counting of blades or nothing. Where the first is the best performance and the last the worst.

As mentioned, the ideal tracking duration is the entire length of the test video. However, when this is not the case, an increasing amount of initialisations would show that the system is capable to construct a new model after losing the previous one, which indicates robustness.

# 4 Experimental Setup and Results

Each of the experiments on both direct and indirect SLAM systems performed in this study will be described, and their results will be presented in the following sections.

## 4.1 Indirect VSLAM

The indirect VSLAM system used for evaluation is ORB SLAM2. The advised image resolution for this system is 640x360 pixels. Since the test data has a higher resolution each frame has been rescaled. Because there is no camera calibration available the frames are treated as undistorted. The calibration file can be found in Appendix A. In the next section, the replacement of ORB with SuperGlue, LoFTR and ground truth is discussed. The sections after that discuss the performed experiments and present their results.

**Replacement of ORB matcher**

To understand the replacement of the original ORB matcher in the ORB SLAM2 system, it is required to first be aware of what the matcher does. The matcher first detects the key points in the frame and then calculates a corresponding descriptor. Said descriptor is used to match the same key point in another frame and is represented as a 32x1 vector of 8-bit unsigned integers. Both this detection of features and conversion to descriptors needs to be replaced when removing the ORB matcher. Instead of implementing the three replacement matchers separately in the SLAM system, we have opted to introduce a method where matches can be read in from a text file. This text file can be generated by running either of the replacement matchers and storing the outcome. For each of the frames in the video, a list of present features is stored along with their position and match-ID. The ID is unique for a specific feature and is consistent throughout the frames. It can therefore be used as a descriptor to match features. Converting the ID, which is a non-negative integer, to a descriptor is done using base-10 to base-256 conversion, as the descriptor can be seen as a vector of 32x1 with 8-bit digits. Following this text file approach does not result in a SLAM system that can be run directly on a video, since the matcher has to compute its output on the full video before the SLAM system can run. This is not a problem for this study, as the goal is to analyse the performance and not to deliver a proper SLAM system.

Both the SuperGlue and LoFTR neural network (NN) matchers output matches between consecutive frames instead of descriptors of key points, so there is no descriptor available to use as match ID. However, when the same keypoint $X$ is present in frame 1, 2 and 3 with the NN matcher matching $X$ between both 1-2 and 2-3, we can link $X$ in all three frames by observing that the location of $X$ on frame 2 is the same in match 1-2 and 2-3. Using this technique we can track a feature in subsequent frames until it is not detected in the next frame anymore, the feature will then get a unique match ID and is stored in the text file.

Both SuperGlue[6] and LoFTR[7] were run using their default settings for indoor environments on resolution 640x360 and 640x480 pixels respectively. These settings were proposed by Huizer in [18]. The ground truth method generating the matches-file is described in [19], and is based on running SuperGlue and filtering matches.

**Experiment 1: ORB matcher on normal videos**

The original system is not able to initialise a reconstruction when run on the test videos. On video B, between 1500 and 1700 key points are detected in each frame. Out of these less than 80 points are matched of which less than 35 points are correct matches, while the system needs at least 100 matches by design. This low number of correct matches, a large amount of noisy matches and the short duration a correct match lasts makes the original system incapable of initialising any point cloud. Similar results were obtained when testing with video A. Since the indirect system using the ORB matcher does not perform well, further experiments with the indirect system will be performed using SuperGlue, LoFTR and ground truth as matchers.

**Experiment 2: Replaced matchers on normal videos**

In this experiment, the indirect SLAM approach is tested using SuperGlue, LoFTR and ground truth (GT) as a replacement for the original ORB feature matcher. As the new matchers significantly outperform ORB it is expected that these systems are capable of initialising and tracking a model of the blades. The results are shown in Table 1.

| Vid. | Matcher | # Inits | Max. duration (s) | Avg. duration (s) | Use for |
|---|---|---|---|---|---|
| A | SuperGlue | 0 | - | - | - |
| | LoFTR | 15 | 1.4 | 1.0 | Count/loc |
| | GT | 2 | 1.5 | 0.90 | Count/loc |
| B | SuperGlue | 2 | 1.3 | 1.3 | Count/loc |
| | LoFTR | 5 | 0.40 | 0.20 | - |
| | GT | 13 | 1.6 | 0.80 | Count/loc |

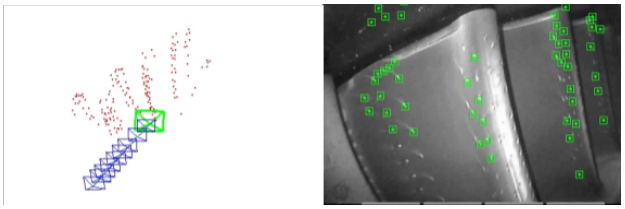
Table 1: Qualitative evaluation on the results of running indirect SLAM system ORB SLAM2 where ORB is replaced with another matcher on video A & B.

The results show that SuperGlue is not able to initialise a model on video A, on video B it manages to do so 2 times. Although only twice, both times the system managed to track

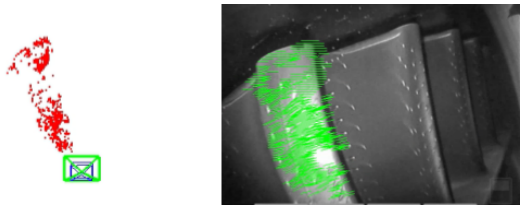
[6] https://github.com/tzvikif/SuperGlue

[7] https://github.com/zju3dv/LoFTR

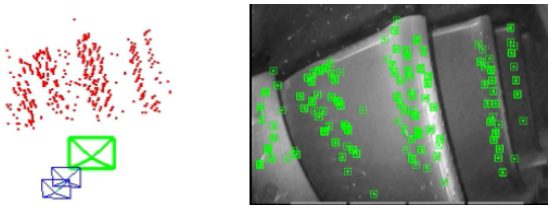
the blades correctly for 1.3 seconds yielding the highest average tracking time. The reconstruction and tracking are shown in Figure 3a. LoFTR on video A can initialise 15 times with an average duration of 1 second. The total tracking time is therefore 15 seconds, which is almost the duration of the video. So the system is always able to initialise a model, but can only track and extend it for approximately 1 second, which is the duration a blade is in view. When ran on video B, the LoFTR system can initialise a correct model occasionally but only for a short amount of time. Furthermore, the constructed models are not useful as only a small part of one blade is modelled, as seen in Figure 3b. Then the ground truth, which is not able to initialise frequently on video A. However when initialised, it does manage to track and extend well. On video B, with 13 initialisations and an average duration of 0.8 seconds, the total tracking time is approaching the duration of the video, the reconstruction is shown in Figure 3c. These results show that each of the matching methods can initialise a model. However, they are not able to track and extend that model after all blades that were in view initially left the frame, which is after 1 second. This observation introduces the follow-up experiment, which looks into how long the systems can track and extend a model when the same blades are kept in the frame.



(a) Indirect reconstruction using SuperGlue matcher; a sparse model with long tracking duration and non-robust initialisation.



(b) Indirect reconstruction using LoFTR matcher; useless model with short tracking duration and non-robust initialisation.



(c) Indirect reconstruction using groundtruth matcher; a semi-dense model with long tracking duration and robust initialisation.

Figure 3: Reconstructions from video B using different matchers with indirect SLAM. *Red dots - points in the model; blue squares - previous camera positions; green square - current camera position; green dots - features tracked in the frame.*

**Experiment 3: Replaced matchers on looped video**

In this experiment the videos were edited to be played and then reversed, which combined is repeated three times such that at least one blade is in the frame for the full edited video. The hypothesis in this experiment is that after a model is initialised the system under test cannot extend the model with more blades, because the original key points move out of frame. If the hypothesis is correct, the well-performing systems from the tests on the normal videos should be able to track the models longer. Therefore increasing the duration of tracking which in turn decreases the number of initialisations. The results are shown in Table 2.

| Vid. | Matcher | # Inits | Max. duration (s) | Avg. duration (s) | Use for |
|---|---|---|---|---|---|
| A loop | SuperGlue | 3 | 0 | 0 | - |
| | LoFTR | 2 | 7.5 | 5.7 | Count/loc |
| | GT | 3 | 1.1 | 0.7 | Count/loc |
| B loop | SuperGlue | 4 | 2.4 | 1.5 | Count/loc |
| | LoFTR | 6 | 0.10 | 0.10 | - |
| | GT | 4 | 9.2 | 2.9 | Count/loc |

Table 2: Qualitative evaluation on the results of running indirect SLAM system ORB SLAM2 on the looped videos where ORB is replaced with another matcher on video B.

The well-performing systems from Experiment 2, LoFTR on video A, and SuperGlue as well as ground truth on video B, do confirm the hypothesis. Since the other systems were not able to initialise a model in the previous experiment, it was not expected that their performance would change, which is also confirmed. Although less than previously, the systems still lose track of their models occasionally since all systems have more than one initialisation. It has to be mentioned that both LoFTR on video A and the ground truth on video B were cut off during their longest tracking attempt due to the video ending. Apart from these results, the quality of the models did not improve nor did the model become denser. Therefore all models that were correctly initialised are still only useful for counting blades and localization.

## 4.2 Direct VSLAM

The direct VSLAM systems used for evaluation are LSD, SVO and DSO. The advised image resolution for the first two systems is 640x360 and for the latter 640x480 pixels. The frames have been rescaled respectively and the calibration file shown in Appendix A is used. The following sections discuss the performed experiments and present their results.

**Experiment 4: Direct SLAM on normal videos**

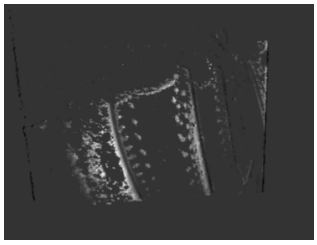
In this experiment both videos are tested on the direct SLAM systems. It is expected that the reconstruction of the blades is more consistent and denser. This hypothesis is based on the claim that direct methods are more robust on featureless surfaces compared to indirect methods [1]. If the hypothesis is correct, the systems will always be able to initialise a model. Although, said model might be lost when the modelled blade moves out of frame. There are no expectations with regards

| Vid. | Matcher | # Inits | Max. duration (s) | Avg. duration (s) | Use for |
|---|---|---|---|---|---|
| A | LSD | 9 | 2.0 | 1.5 | Count/loc |
| | DSO | 5 | 3.6 | 2.7 | Count/loc |
| | SVO | 0 | - | - | - |
| B | LSD | 7 | 1.9 | 1.8 | Damage |
| | DSO | 4 | 5.1 | 3.0 | Damage |
| | SVO | 0 | - | - | - |

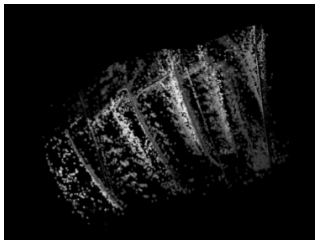
Table 3: Qualitative evaluation on the results of running direct SLAM systems: LSD, DSO & SVO on videos A and B without modification.

to the ability of the systems to connect separate blades. The results are shown in Table 3.

Like ORB SLAM in the indirect approaches, SVO was not able to initialise a reconstruction on either test video and is therefore not tested in further experiments. However, the hypothesis can be accepted since both LSD and DSO were always able to initialise a model, even though it could only be tracked and extended for a few seconds. Apart from this, the ability of DSO to extend the model and track the camera significantly longer than LSD on both videos stands out in the table. Using video A the models of both systems can be used to count blades and localize the camera, whereas on video B the models can be used for damage assessment. The quality of the models is similar as can be seen in Figure 4. LS, however, introduces less noise and the full shape of the blades is more visible. Compared to the indirect systems tested in Experiment 2, both direct systems reconstruct significantly denser models, can track said model longer and are more robust in their initialisation.



(a) Direct reconstruction using LSD, a dense reconstruction with short tracking duration and robust initialisation. Useful for damage assessment.



(b) Direct reconstruction using DSO, a noisy dense reconstruction with long tracking duration and robust initialisation. Useful for damage assessment.

Figure 4: Comparing reconstructed models by LSD and DSO SLAM on video B in Experiment 4.

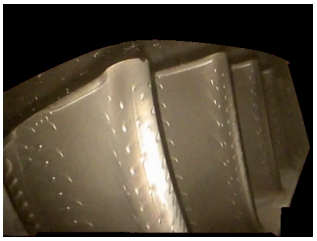
### Experiments 5-7: Direct SLAM follow-up

Based on the results of Experiment 4, several follow-up experiments were performed in which one aspect of the Experiment 4 was changed to test its influence on the performance. The results of Experiments 5-7 were either similar to Experiment 4 or decreased performance significantly, it, therefore, suffices to describe their results.

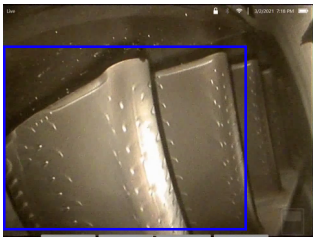
**5: Running SLAM near real-time** is an option for both LSD and DSO, which allows the systems to take longer in processing the videos. In near real-time, the system takes about 3 times longer than in normal real-time, with the normal variant processing 30 frames per second. The hypothesis for this experiment was that performance would improve since every frame could be fully processed without time restrictions as discussed in [1]. The results, however, were the same as for Experiment 4. Running the systems with fewer time restrictions, therefore, does not improve performance.

**6: Looping the blade**, like in Experiment 3, will test if the systems can track and improve the initialised model when the corresponding blade stays in the frame. Experiments showed that both LSD and DSO were able to initialise and track a model for the full duration of both looped videos. Therefore all tests had only one initialisation, that was tracked for 15 seconds. Apart from that, no increase in the quality of the model was registered.

**7: Removing static pixels** from the frame, to test how much these parts that are not part of the blades influence the performance of the systems. To hide most of the static pixels we use two different approaches, A: adding a mask like Figure 5a, and B: cropping the frames to the size of the blue square like in Figure 5b which eliminates most of the static parts. For both approaches LSD did not act differently when compared to Experiment 4, apart from some of the noise disappearing. DSO on the other hand consistently failed to either initialise a proper model or track the model after initialisation. When running DSO it was observed that the system was trying to match the masked parts of the image, instead of ignoring them which was the intent of the mask. Removing the static pixels using a mask or by cropping the images does therefore not have a positive effect on the performance of LSD and DSO.



(a) Frame from video B. *Black mask denotes how the frames are masked for experiment 7A.*



(b) Frame from video B. *Blue square denotes the cropped version for experiment 7B.*

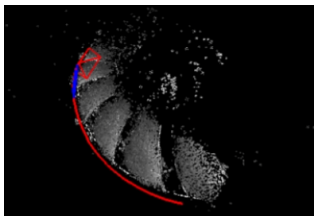
Figure 5: Frame from video B showing the changes made for Experiment 7 to remove static pixels by either masking or cropping. Both approaches do not have a positive effect on the performance of LSD nor DSO.

### Experiment 8: The influence of camera calibration

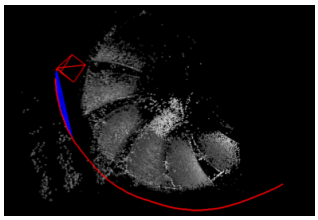
Unlike indirect SLAM the performance of direct methods suffers from geometric distortion in the frame [9]. This distortion can be removed by calibrating the camera. However, this calibration is not available for the borescope videos as discussed in Section 3.1. To overcome this, a borescope similar to the one used for video A was calibrated. Video C was

then shot using the borescope where the turbine was replaced by a computer fan, a frame from C can be seen in Figure 6c.

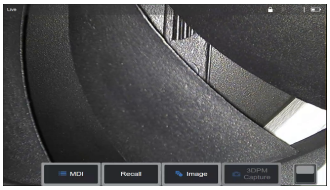
LSD and DSO were tested both calibrated and not calibrated on video C. For LSD there was no notable difference in performance when the camera was calibrated and in both tests, the system did not manage to extend the model after the modelled blades left the frame. On the other hand, DSO managed to initialise and extend the model for the full duration of the video and reveals the influence of calibrating the camera. This difference can be seen by comparing Figure 6b, which is calibrated, with Figure 6a, which is not. The calibrated version is superior in linking the blades together.



(a) Direct reconstruction using uncalibrated DSO, the blades **are not** linked together properly.



(b) Direct reconstruction using calibrated DSO, the blades **are** linked together properly.




(c) Example frame from video C.

Figure 6: Video C and its reconstruction showing the difference when calibrating DSO. *Red line - past camera position; red square - current camera position; grey dots - points in the model*

## 5 Responsible Research

The two main ethical aspects of this study are the possible replacement of a human inspector by an automated system and the responsible use of the system to assist the inspector in damage assessment. Starting with the replacement issue which is a well-known topic in the field of robotics and AI. Like many similar cases, this system would assist a human operator instead of replacing it. The system can therefore be seen as an extra tool to make the inspection easier and more thorough. Which introduces the second aspect concerning the responsible use of this tool. Since the system would be designed to be used as a tool rather than a stand-alone system, the inspector should also use it as such. To ensure safe use of the system a clear explanation of what the system can do but equally important what the system cannot do is required. This scenario is in line with the principle of meaningful human control, described in [22], stating that humans should remain in control and therefore be morally responsible for relevant decisions made by intelligent systems. No moral competencies are required for the system and the human operator is the moral decision-maker, as the system would not be designed to make decisions. Designing the system to be more autonomous requires the system to have an understanding of moral, the implications of this are discussed in [23].

Regarding the reproducibility of the study, three aspects require further discussion. Firstly, the test data used in this study is not publicly available as it is owned by Aiir. However, there are several available example videos[8] that are similar to the data used in this study. Secondly, the SLAM systems used for evaluation are all publicly available and are run with the settings described in Chapter 4. The adapted ORB SLAM implementation can be shared upon request. Thirdly, the qualitative evaluation is somewhat subjective and therefore makes reproducing the exact results difficult. However, when following the described method similar results will be produced. It can therefore be concluded that the method and results are sufficiently reproducible.

## 6 Discussion

The results of this study show that both direct and indirect VSLAM approaches are capable of reconstructing an initial model of the turbine blades, although they tend to lose track after the modelled blades leave the frame. When compared, it is apparent that the direct methods outperform the indirect ones in every aspect; the density of the model, robustness of initialisation and the duration of tracking. The next section will discuss the results of the performed experiments and how they help in answering the main question of this study, after which the limitations and future work will be discussed.

### 6.1 Results analysis and explanation

For the indirect methods in Experiment 1 and 2, it was shown that the original ORB SLAM system was not capable of reconstructing the blades due to a lack of features. Replacing the ORB matcher with neural networks LoFTR and SuperGlue demonstrated that indirect approaches are capable of initialising and tracking the blades. In exceptionally low feature environments such as video A, LoFTR seems more robust in initialisation than SuperGlue. Due to the higher amount of matches LoFTR registers, it also results in denser models. However, when SuperGlue does manage to initialise a model it displays the ability to track and extend the model longer than LoFTR. One unanticipated finding was the inferior performance of LoFTR on video B. This can be explained by the nature of LoFTR, which needs a feature to move at least 8 pixels between frames to detect a useful match [18]. In video B the movement of the blades in the back of the frame is lower than this threshold, resulting in a lack of matches on the more distant blades. Summarizing, it depends on the scenario if LoFTR or SuperGlue performs better. LoFTR is better suited for low feature environments but needs the camera to be perpendicular to the movement of the blades, whereas SuperGlue is more invariant to the positioning of the camera but works better if some texture is present.

The direct methods in Experiment 4 show that direct approaches are capable of reconstructing the blades properly. As stated, this study confirms that direct approaches are more robust in featureless environments than indirect systems as claimed previously in [1, 9]. Comparing the direct methods

[8] https://www.rvi-ltd.com/borescope.html

between themselves shows that LSD reconstructs a model with less noise than DSO, but is inferior to DSO when it comes to tracking the blades after they move out of the frame. However, even DSO is not capable of tracking its models indefinitely. Which in turn is likely to improve significantly when a calibrated camera is used, as shown in Experiment 8 and stated previously in [9]. This tracking of the model and linking blades together is however not required for the damage assessment mentioned as one of the aims, since one could model all blades separately and perform damage on the individual blades. In addition, it can be argued that the quality of the models reconstructed by these direct approaches is good enough to identify cracks and holes on the blade. Especially when taking into account that damages create features that will be picked up by the systems and will therefore show up as anomalies in the model of the blade. LSD and DSO can be used for this modelling of one blade, although LSD would reconstruct models with less noise. However, if it is required to model all blades together, DSO will outperform all other options.

Then testing the systems on their ability to track the model when the blade stays in the frame in Experiment 3 & 6, showing that the blade moving out of the frame also hinders performance. This finding supports the theory that the lack of features is not the only limitation to be researched to increase the performance of VSLAM systems on aircraft turbines.

Overall it can be observed that the SLAM systems perform better on video B than video A. Which can partly be attributed to the higher amount of features on the blades of video B. However, as discussed in [9], the bigger the difference in depth within the frame, the better the depth estimation. Since the blades in video B move towards the camera instead of past the camera like in video A, the difference in depth in video B is larger compared to A. This larger difference in depth is likely to contribute to the better performance on video B.

### 6.2 Limitations and future work

The results of this study should be interpreted with caution as each of the main experiments in this study is limited to some extent. For the experiments in which the ORB matcher is replaced with neural networks, it has to be noted that these matchers were not run in real-time. The frequency at which they can process frames and return matches is around 5 Hz compared to the standard ORB matcher which is implemented to reach over 30 Hz. However, the current implementation of the neural networks in the indirect SLAM system is far from optimal and meant to test the quality of the models, not the computational time of the system. When properly implemented and optimized, the indirect system will likely be able to run in real-time. Besides, the real-time direct matchers outperform the indirect system even when the indirect approach runs near real-time.

The calibration of the camera is missing in the experiments where the direct systems were tested. With previous studies as well as this study showing the importance of this calibration, testing the systems without calibration might be labelled stubborn. However, this study can be interpreted as a preliminary study trying to get results with as few resources as possible. Now that the concept is shown to work to some extent, a more in-depth study with more resources can be considered to be worthwhile. Apart from this, it would increase the usability of the VSLAM method if no calibration of the camera is required to get proper results. The study into uncalibrated direct VSLAM is therefore not perfect but still interesting.

The last main limitation of this study is the material of the blades used in Experiment 8 to test the influence of calibration. In this experiment, a plastic computer fan with more texture than the metal blades of the aircraft turbine was used. Although ideally the texture of the test object should have had less texture, it is clear that the other aspects of the test video were similar to the actual borescope inspection videos. Besides, the calibrated and uncalibrated results of Experiment 8 were only compared between themselves and not with experiments on other videos.

These last two limitations introduce the need for future research into direct VSLAM similar to how it is performed in Experiment 4, but with a calibrated camera. To further improve the performance of the SLAM systems, we advise looking into how the noise that is likely introduced by the static parts of the scene can be removed. This study tried to solve this issue in Experiment 7 but did not succeed. The final suggested improvement is looking into the possibilities of keeping the blades in the frame for a longer period of time and introducing more depth into the scene. This could be done by taking videos, similar to video B, in which the blades move towards the camera.

## 7 Conclusions

3D modelling the blades visible in borescope videos of aircraft turbines opens up the possibility to automate the mandatory rotor inspection. The need to perform these inspections quickly urges for a 3D reconstruction method that runs in real-time such as SLAM. Since SLAM is designed for general scenes such as offices, it has not been thoroughly tested in environments similar to the inspection videos. Therefore, this study has tried to determine how well different SLAM approaches perform on this niche task. To achieve this, both direct and indirect systems have been tested. The results show that existing indirect approaches fail, due to the lack of features on the blades. Yet, when the traditional feature matcher is replaced by more recent neural networks, the system does succeed at reconstructing a few blades. Still, the direct approaches perform significantly better as they are more robust, create denser models, and can track the model longer. Results can be improved by making use of a calibrated camera. Therefore this study shows that it is possible to reconstruct the blades from an aircraft turbine using SLAM, although modelling more than a few consecutive blades is not shown yet. Even modelling the blades separately allows to automate damage assessment and therefore potentially simplifies and speeds up inspections of aircraft turbines.

## References

[1] Jakob Engel, Thomas Schöps, and Daniel Cremers. "LSD-SLAM: Large-Scale Direct Monocular SLAM". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet

et al. Cham: Springer International Publishing, 2014, pp. 834–849. ISBN: 978-3-319-10605-2.

[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. "ORB-SLAM: A Versatile and Accurate Monocular SLAM System". In: *IEEE Transactions on Robotics* 31.5 (2015), pp. 1147–1163. DOI: 10.1109/TRO.2015.2463671.

[3] Jürgen Sturm et al. "A benchmark for the evaluation of RGB-D SLAM systems". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012, pp. 573–580. DOI: 10.1109/IROS.2012.6385773.

[4] Georges Younes et al. "Keyframe-based monocular SLAM: design, survey, and future directions". In: *Robotics and Autonomous Systems* 98 (Dec. 2017), pp. 67–88. ISSN: 0921-8890. DOI: 10.1016/j.robot.2017.09.010. URL: http://dx.doi.org/10.1016/j.robot.2017.09.010.

[5] Hauke Strasdat, J.M.M. Montiel, and Andrew J. Davison. "Editors Choice Article". English. In: *Image and Vision Computing* 30.2 (2012), pp. 65–77. DOI: 10.1016/j.imavis.2012.02.009.

[6] Georg Klein and David Murray. "Parallel Tracking and Mapping for Small AR Workspaces". In: *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*. 2007, pp. 225–234. DOI: 10.1109/ISMAR.2007.4538852.

[7] E. Rublee et al. "ORB: An efficient alternative to SIFT or SURF". In: *2011 International Conference on Computer Vision*. 2011, pp. 2564–2571. DOI: 10.1109/ICCV.2011.6126544.

[8] Maksim Filipenko and Ilya Afanasyev. "Comparison of Various SLAM Systems for Mobile Robot in an Indoor Environment". In: *2018 International Conference on Intelligent Systems (IS)*. 2018, pp. 400–407. DOI: 10.1109/IS.2018.8710464.

[9] Jakob Engel, Vladlen Koltun, and Daniel Cremers. "Direct Sparse Odometry". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.3 (2018), pp. 611–625. DOI: 10.1109/TPAMI.2017.2658577.

[10] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. "SVO: Fast semi-direct monocular visual odometry". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014, pp. 15–22. DOI: 10.1109/ICRA.2014.6906584.

[11] A Geiger et al. "Vision meets robotics: The KITTI dataset". In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1231–1237. DOI: 10.1177/0278364913491297. eprint: https://doi.org/10.1177/0278364913491297. URL: https://doi.org/10.1177/0278364913491297.

[12] Zuzana Kukelova and Tomas Pajdla. "A minimal solution to the autocalibration of radial distortion". In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 2007, pp. 1–7. DOI: 10.1109/CVPR.2007.383063.

[13] Carsten Steger. "Estimating the fundamental matrix under pure translation and radial distortion". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 74 (2012), pp. 202–217. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2012.09.012. URL: https://www.sciencedirect.com/science/article/pii/S0924271612001815.

[14] G. Stein. "Internal Camera Calibration using Rotation and Geometric Shapes". Unpublished MSc Thesis MIT. 1993.

[15] H. Sperker and A. Henrich. "Feature-based object recognition — A case study for car model detection". In: *2013 11th International Workshop on Content-Based Multimedia Indexing (CBMI)*. 2013, pp. 127–130. DOI: 10.1109/CBMI.2013.6576568.

[16] Paul-Edouard Sarlin et al. "SuperGlue: Learning Feature Matching with Graph Neural Networks". In: *CVPR*. 2020. URL: https://arxiv.org/abs/1911.11763.

[17] Jiaming Sun et al. "LoFTR: Detector-Free Local Feature Matching with Transformers". In: *CVPR* (2021).

[18] Rick Huizer. "A performance analysis of interest point detection/matching on shiny and non-textured surfaces - A case study on aircraft engine borescope inspection videos". Unpublished BSc thesis TU Delft. 2021.

[19] Devin Lieuw A Soe. "Ground Truth for Evaluating 3D Reconstruction of Jet Engines". Unpublished BSc thesis TU Delft. 2021.

[20] Felix Caesar. "A Novel SLAM Quality Evaluation Method". Unpublished MSc thesis KTH Royal Institute of Technology. 2019.

[21] Anton Filatov et al. "2D SLAM Quality Evaluation Methods". In: *CoRR* abs/1708.02354 (2017). arXiv: 1708.02354. URL: http://arxiv.org/abs/1708.02354.

[22] Filippo Santoni de Sio and Jeroen van den hoven. "Meaningful Human Control over Autonomous Systems: A Philosophical Account". In: *Frontiers in Robotics and AI* 5 (Feb. 2018), p. 15. DOI: 10.3389/frobt.2018.00015.

[23] Jasper van der Waa et al. "Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach". In: *Engineering Psychology and Cognitive Ergonomics. Cognition and Design*. Ed. by Don Harris and Wen-Chin Li. Cham: Springer International Publishing, 2020, pp. 203–220. ISBN: 978-3-030-49183-3.

# A Appendix: Calibration files

The following lines show the text file following the format described in [3] used to undistort the images for all systems but LSD.

```
517 516 320 180 0 0 0 0
640 360
crop
640 360
```

The following lines show the text file following the format described in [3] used to undistort the images for LSD.

517 516 320 180 0 0 0 0
640 360
crop
640 480