Evaluating Structure-from-Motion on shiny and non-textured surfaces in borescope videos

Alec Nonnemaker[1] , **Dr. Jan van Gemert**[1] , **Burak Yildiz**[1]

[1]TU Delft

Abstract

To aid in damage assessment, creating 3D reconstruction from borescope videos of jet engines could be very beneficial. However, jet engines often have shiny and non-textured surfaces, and the performance of 3D reconstruction methods is unknown in this case. This paper aims to qualitatively and quantitatively evaluate Structure from Motion (SfM) on these borescope videos. SfM is a technique for 3D reconstruction that uses collections of images to create 3D models. An evaluation was done on borescope videos with differing characteristics using SIFT, SuperGlue, and ground truth for feature detection. Even though small experiments with the global SfM approach produced insufficient results, more extensive experiments using incremental SfM show promising performance on borescope videos and potential for accurate damage assessment, especially when combined with multi-view stereo.

1 Introduction

Damage assessment through industrial videos is crucial when inspecting and dealing with certain products and should be done as precisely as possible. Especially when human lives depend on these products or when substantial amounts of money are involved. Both of these are the case when dealing with airplanes. Aiir Innovations works with aviation companies and creates software that automates borescope inspections of jet engines [6]. Borescope inspections are done when difficult-to-reach places need to be visually inspected and use a rigid or flexible tube with a camera and occasionally a light on the end of it. The ideal way to do this is to create a 3D model from an input video which can then be used to automatically make measurements.

This research paper focuses specifically on the Structure from Motion (SfM) technique for 3D reconstruction. SfM is a technique for the 3D reconstruction of scenes or objects utilizing a sequence of 2D images taken from different viewpoints. It is based on the same idea as finding structure from stereo vision (vision from two views). Finding structure from stereo vision is done by using triangulation to calculate the relative position of objects in 3D space. This is similar to what humans use to perceive depth. Instead of a view from two cameras (or eyes), SfM takes many overlapping images of an object taken from different angles. It uses a technique called feature matching, where features are tracked between images. These feature matches are then used to reconstruct their 3D position and the camera positions resulting in a point-cloud-based 3D model. There are several strategies for performing SfM which have been extensively researched including incremental SfM [14] and global SfM [22]. SfM can be combined with Multi-View Stereo (MVS) to obtain dense 3D models compared to the relatively spare 3D model resulting from just SfM. MVS using results from SfM has also been researched [15]. However, these researches focus on 3D reconstruction of objects or scenes with a lot of texture and where a lot of images are made from good angles at a good distance from the object or scene. Consequently, it is not clear how these methods would perform when scenes lack texture or are shiny and when the camera capturing the images is not in the optimal position. This and several other factors like the resolution of the images, the focal length of the camera, quality of the lens, and lighting conditions can affect the accuracy greatly [12]. All these factors are fairly unique in the case of borescope videos so this research aims to evaluate SfM in the context of borescope videos.

To formalize, the main research question that will be answered in this paper is ”How well does SfM work on borescope videos with shiny and non-textured surfaces?”. Since there are two main strategies for SfM with each their advantages and disadvantages this main research question can be split up into the two sub-questions:

- How well does incremental SfM work on borescope videos with shiny and non-textured surfaces?
- How well does global SfM work on borescope videos with shiny and non-textured surfaces?

This research will evaluate the most prominent implementations of incremental and global SfM on efficiency and accuracy when working with borescope videos. Accuracy will be evaluated qualitatively and quantitatively.

2 Structure-from-Motion approaches

Although the different approaches to SfM differ in how the final 3D model is constructed, they all start in identically. That

is the extraction of features and matching of images. Following this, the two approaches both go in different directions with each their advantages and disadvantages. In the following sections, all stages of these SfM approaches will be discussed and related work will be introduced. Starting off the task of feature extraction and matching is outlined, followed by the different processes of reconstruction for every approach.

2.1 Feature extraction and Matching

The first part of the SfM pipeline is to find overlapping images. With these overlapping images and their corresponding feature pairs, 3D reconstructions can be created.

Feature extraction is done on every image individually and results in a set of features that are characterised by an x and y coordinate and a feature descriptor. Important is that these features are invariant to rotation and scale so that identical points can be matched in images from different angles. The most widely used feature is SIFT [7] and its derivative SURF. However, these two features do not perform well on shiny and non-textured surfaces. Another newer class of features uses neural networks to detect and describe interest points [13; 18]. These seem to have superior performance compared to SIFT, SURF, and other traditional features. Especially when dealing with shiny and non-textured surfaces.

Next, images are matched using the features obtained from the previous step. For every feature in an image, it searches for the most similar feature in a potentially overlapping image. If these features are sufficiently similar, their corresponding point can be considered to be the same in the context of the scene. Finding the most similar feature between images is often done using a nearest neighbour approach [16; 7; 23]. If two images then have a set of points in common it means they see the same scene part. Exhaustively matching all these images takes $O(n^2)$ time which means it is one of the bottlenecks in the SfM pipeline. However, since the borescope videos have the images collected sequentially only consecutive frames can be matches which removes the need to exhaustively match all images. This significantly reduces the time complexity.

The final stage of feature extraction and image matching is geometric verification. This is a critical step in the SfM pipeline which removes outliers in image matches. The previous matching stage uses feature descriptors which only tells you that two points in an image have a sufficiently similar appearance. It does not guarantee that these feature matches also correspond to the same points in the scene. Geometric verification verifies these matches by estimating a geometric transformation that maps a sufficient amount of points between two matching images. To filter out as many of the outliers as possible a robust estimation technique such as RANSAC [3] is used. This geometric verification is especially important for global SfM since incremental SfM also uses RANSAC when adding images to the model. The result of this pipeline stage, after which the three SfM approaches diverge in methodology, is a graph with images as nodes and the edges connecting verified pairs of images.

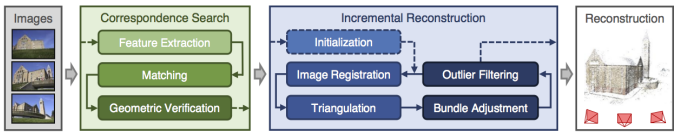

Figure 1: Incremental SfM pipeline [14].

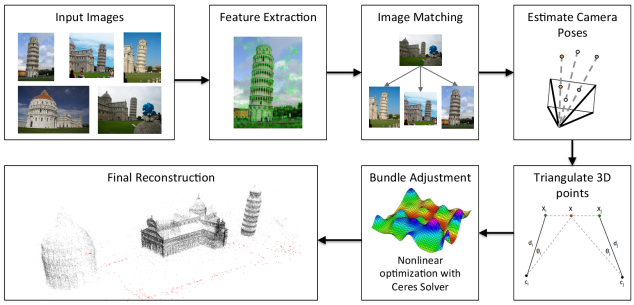

Figure 2: Global SfM pipeline [19].

2.2 Incremental SfM

The core idea of incremental SfM is to initialize a model with a two-view reconstruction and then incrementally registering new images to the model until you get a full reconstruction. This initialization is very important since a bad initialization will lead to an undesirable result. If a low run-time is important initializing from a sparse location in the image graph decreases the run-time since the Bundle Adjustment will have to deal with less information. However, this does come at the cost of robustness and accuracy which you will have more of if you start from a denser location in the image graph.

After the initialization of the model, the new images are added incrementally in three steps: Image Registration, Triangulation, and Bundle Adjustment(BA). Starting with image registration, where a new image is registered to the model while calculating the pose (position and rotation) of the camera where the image is from. This pose can be calculated by solving the Perspective-n-Point [3] problem which uses correspondence between the 2D features in the newly registered image and the 3D features already present in the model. Also, here RANSAC is used to get a robust estimation of the camera pose for the newly registered image.

The new image having been registered to the model, it now can add new points to the model using triangulation. A new point from the image can be added when at least one of the previously registered images also observes this point. The new 3D point is estimated by triangulation using the pose of both image cameras and the overlapping point in both images.

Since in both the image registration step and the triangulation step the results are estimated, these results can have inaccuracies. An inaccuracy in one of these steps affects the other step in both ways. Inaccurate camera poses affect triangulation and in turn, small errors in triangulation can affect the estimation of the next camera pose in image registration. These small inaccuracies can then quickly become large inaccuracies which will make the model unusable. Bundle adjustment (BA) [21] aims to refine camera poses and points in the

model to optimize the 3D reconstruction. It does this by minimizing the so-called reprojection error which is the difference between the 3D point reprojected back into the camera image and the 2D point of the camera image. This reprojection error is minimized over all the combinations of images and points, turning it into a large least-squares problem. The BA algorithm most often used is Levenberg-Marquardt [21]. BA is run locally on groups of highly connected images and globally when the model has grown by a certain percentage since the last global BA.

2.3 Global SfM

Compared to incremental SfM which registers cameras to the model one by one, global SfM initializes all cameras at once. Most of the earlier prominent SfM methods are incremental since they are simpler and fairly robust due to the extensive refining that is done using BA. They are however typically fairly slow because of this same extensive refining and recent studies seem to suggest that global SfM has the potential to be more accurate and efficient [2; 24]. The general pipeline for global SfM starts by estimating the set of camera poses, then triangulating all points, and finally doing bundle adjustment to optimize the cameras and 3D points together.

The first and most important part of the pipeline, estimating the camera poses is done using motion averaging. This motion averaging is split into rotation averaging and translation averaging where relative rotation R_i and position c_i of the cameras are estimated. These two are constrained by the following equations:

$$R_{ij} = R_j R_i^{\intercal} \tag{1}$$

$$\lambda_{ij} t_{ij} = R_j(c_i - c_j), \tag{2}$$

where R_{ij} is a 3 X 3 relative rotation matrix, t_{ij} is a unit vector representing the relative translation between i and j, and λ_{ij} a scale factor. Camera rotation and position can then be obtained by solving the minimization problem for:

$$arg \min_{R'} \sum_{R_{ij} \in R'_{rel}} d^R(R_{ij}, R_j R_i^{\intercal}) \tag{3}$$

$$arg \min_{T} \sum_{t_{ij} \in T_{rel}} d^T(t_{ij}, R_j(c_i - c_j)), \tag{4}$$

where d^R is a distance measure for 3 X 3 rotations, d^T is a dissimilarity measure, R'_{rel} and R' are the sets of relative and normal camera rotations, T_{rel} is the set of relative translation between cameras, and T is the set of global camera positions. Several methods have been proposed to perform this motion averaging as robust as possible [22; 1; 8; 11].

After obtaining the camera poses, 3D points can now be triangulated to create a reconstruction. Since all camera poses are known at once, this triangulation step can be done in parallel. Finally, to refine camera poses and 3D points BA is done, using the same algorithm as the global BA discussed for incremental SfM.

3 Evaluation of SfM on borescope videos

To find out how well SfM works on borescope videos with shiny and non-textured surfaces a few aspects need to be considered. Namely the feature matching performance issues on shiny and non-textured surfaces and how the SfM result will be evaluated qualitatively and quantitatively. The evaluation was done on borescope videos provided by Aiir Innovations using the state-of-the-art open-source systems *COLMAP*[1] for incremental SfM [14] and *OpenMVG* [2] for global SfM [9].

3.1 Feature matching

One of the main issues being dealt with in the borescope videos is the shininess and lack of texture on the engine rotors. This issue lies mostly with the feature matching stage of the pipeline. For the incremental SfM experiments, two different types of feature matching were used and evaluated. A peer research was done by R. Huizert specifically on the performance of different feature detectors and matchers on shiny and non-textured surfaces [5]. In addition, for both qualitative and quantitative evaluation, ground truth for feature matching was used.

SIFT. As mentioned previously, SIFT is the most common feature used for the SfM algorithm and is considered an industry-standard in the field of computer vision. However, its performance on shiny and non-textured surfaces drops significantly [20]. SIFT detects significant changes in pixel intensity in all directions to find corners in an image. It could potentially still detect the outlines of the jet engine rotors and any visible damage therefore it was still used for the experiments. The open-source implementations utilized use SIFT by default so no pre-processing was necessary to run SfM with SIFT.

Neural Networks. Using neural networks is a newer way of finding and matching features that could potentially work better on shiny and non-textured surfaces. The peer research on feature detectors and matching mentioned earlier recommended using SuperGlue (SG) and LoFTR as the neural networks to perform the feature detection and matching step. As such, they provided us with the feature and matching data to do so. Some pre-processing had to be done to get this data in a format that could be used by *COLMAP*. This format can be found in the documentation of these implementations.

Ground Truth. To accurately evaluate SIFT and the neural networks for performance on SfM it is beneficial to have a ground truth (GT) to compare these methods against. This ground truth will have a very high accuracy but takes a lot of time to create so will not be useful for the production use of SfM. The feature and matching data for the ground truth have been provided by peer research from D. Liew A Soe [17].

3.2 Multi-View Stereo

After running the SfM pipeline the result is a scene reconstruction in the form of a sparse point cloud. Following this, a dense point cloud can be computed using multi-view stereo(MVS) [4]. MVS takes as input a set of camera poses and the sparse point cloud obtained from SfM to compute

[1] https://github.com/colmap/colmap

[2] https://github.com/openMVG/openMVG

depth and normal information for each pixel in an image. Fusing this information for multiple images results in a dense point cloud. *COLMAP* has its own built-in implementation of MVS [15]. *OpenMVG* does not so only sparse reconstruction will be looked at.

3.3 Qualitative evaluation

The goal of creating these models from borescope videos is to accurately assess potential damage to the jet engine. To evaluate this you want to look at the absolute best and final result that would be used to achieve this goal. Evaluating the sparse models resulting from only SfM does not give sufficient insight into whether damage can be assessed. Therefore the qualitative evaluation besides looking at the sparse models also focuses on how well SfM paired with MVS achieves the goal of damage assessment. For this, a comparison was done of the different models and the videos, seeing how similar the 3D models look to the actual jet engine in the videos and if damage visible in the borescope videos is also visible in the 3D models.

Since OpenMVG does not allow for the ability to compute these dense point clouds using MVS only the sparse reconstruction between the incremental and global SfM implementations will be compared. Although this is not a full comparison, comparing the sparse point clouds could give sufficient to predict their relative performance when combined with MVS.

3.4 Quantitative evaluation

Besides comparing the 3D models to the video qualitatively some quantitative analysis of the models was also done. Looking at the model data a few evaluation metrics can be looked at and analysed. These metrics apply to the sparse model since the dense model created with MVS adds an extra layer to the algorithm which is out of scope for this project to analyse the data of.

The first metric is the number of registered images. All experiments used the same amount of images as input so a difference in the number of registered images can indicate a disconnect between images through a lack of feature matches or an issue with the 2D-3D correspondence. For global SfM all images are always considered so this metric only applies to incremental SfM. The second metric is the number of points in the model. This will most likely say something about the number of feature matches found in the matching step of the pipeline but is also affected by the number of images registered in incremental SfM. The next metric is efficiency which is measured in the amount of time it takes to construct the model. This time excludes the feature matching step since these feature matches for neural network and ground truth were provided by peer researches [5; 17]. Finally, the average reprojection error is looked at. These are important measurements since they will essentially show how far BA was able to refine the model which gives an idea of the accuracy of the points in the model. BA tries to minimize the reprojection error over the whole model as much as possible while registering new images to it in incremental SfM.

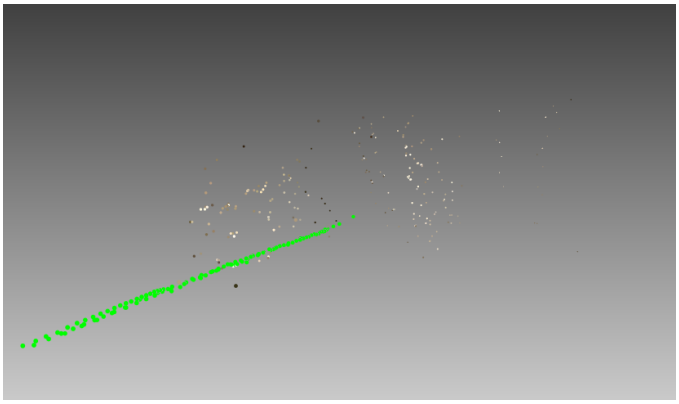

Figure 3: Video 3 global SfM sparse model with camera poses using SIFT. A rough outline of an engine blade can be recognised.

4 Experimental Setup and Results

The experiments are run on datasets provided by Aiir. These consist of videos of different jet engines filmed from different angles. Covering multiple types of engines and angles helps in determining which factors affect the reconstruction. The datasets are collected using a borescope camera of which the intrinsic parameters are not known. The videos are cut into 150 frames each with a resolution of 1280 X 720 px. A low amount of frames was chosen since more frames do not aid in the accuracy of the model, it only makes the model larger which is not needed for evaluation. Then the previously mentioned COLMAP and OpenMVG were run on the datasets using SIFT, SuperGlue, LoFTR, and a ground truth for feature matching. This was done on a 2.2GHz machine with 16GB RAM. To get the best possible reconstruction a few settings were changed from the COLMAP default settings. These setting are: *min_num_matches = 8*, *init_min_num_inliers = 50*, *init_min_tri_angle [deg] = 8,00*, and *abs_pose_min_num_inliers = 12*.

4.1 Model comparison

The first part of the qualitative evaluation is a visual comparison between the different models and the dataset videos. Three different borescope videos were evaluated where two were similar in video angle but different material of blades which meant a higher degree of shininess in video two compared to video one. The third video was a completely different angle from the other two videos. This gives insight into the optimal angle for the borescope to be filmed in to obtain a quality model. For all three videos, LoFTR did not produce any model to be visually evaluated.

Global SfM

Since the global SfM implementation used did not allow for importing features only the already available SIFT features could be used for reconstruction. Experiments were run on all three videos using SIFT for feature detection. Reconstruction on the first video produced a few points but nothing resembling the jet engine blades displayed in the video. Video two did not produce any points. The third video performed better resulting in a sparse point cloud where a rough outline of the blades can be recognised (Fig. 3).

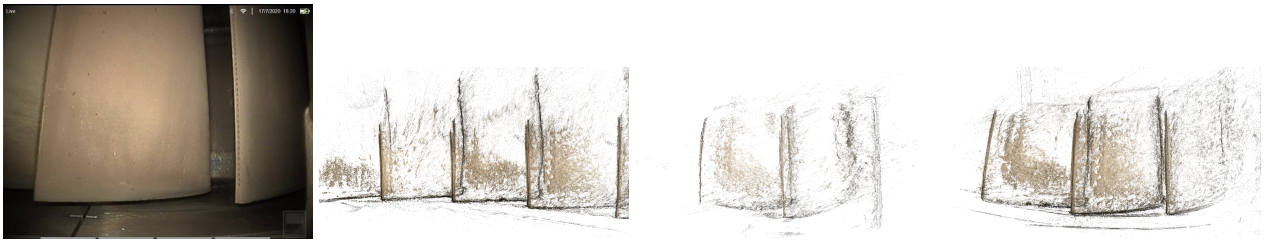

Figure 4: Video 1 comparison of reference video and dense models for GT, SIFT and SG (fLTR). The SuperGlue model is closer to the ground than SIFT model.

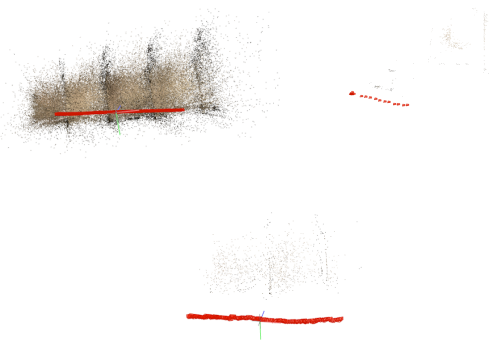

Figure 5: Video 1 sparse models of GT, SIFT and SG models with camera poses(fLTR). GT is substantially more dense compared to SIFT and SG.

Incremental SfM video 1

The first video that incremental SfM was evaluated on was in theory the video that was going to give the best results. The blades do not have a lot of texture. but are not excessively shiny and move horizontally across the screen. Running incremental SfM on this video answers the question of how well incremental SfM performs on a borescope video with these characteristics. A visual comparison between the video itself and the result of running incremental SfM + MVS with SIFT, SuperGlue, and the ground truth can be seen in Fig. 4. True to what was theorised, video 1 has the most visually accurate ground truth and superglue generated dense models.

As for the comparison between the dense models of SuperGlue, SIFT, and the ground truth on video 1, it can be seen that superglue models the blades well but was not able to reconstruct as many blades. SIFT reconstructs only two blades and is quite noisy. Fewer blades being reconstructed means incremental SfM was not able to register all the images in the dataset. The SIFT and to a lesser degree SuperGlue models get saved by MVS however since the sparse reconstructions do not perform well compared to the ground truth as seen in Fig. 5 which partly explains the noisiness of the model. Also visible in the sparse model comparison is the camera poses. These appear to be accurate for all three models which would mean accurate camera poses translate to accurate dense models when using MVS.

Figure 6: Video 2 comparison of reference video and GT dense model. Blades can be clearly recognised.

Incremental SfM video 2

The second video is geometrically similar to the first one as the blades also move horizontally across the screen. However, the blades are shinier and have less texture. Due to the high degree of shininess and lack of texture SIFT and SuperGlue do not produce any model to be visually evaluated. This also hinders the ground truth as seen in the models. They are quite noisy and only the edges of the blades are sufficiently visible. A visual comparison between the video itself and the result of running incremental SfM + MVS with the ground truth can be seen in Fig. 6.

Incremental SfM video 3

This video is filmed from a side angle with the blades moving diagonally and towards the camera. The blades themselves have shiny surfaces but they have grooves on them and the edges are thick and prominent which are well picked up by feature detectors. These grooves can be seen in the dense reconstruction but are also picked up in the sparse reconstruction when using SIFT(Fig. 8). A visual comparison between the video itself and the result of running incremental SfM + MVS with SIFT, SuperGlue, and the ground truth can be seen in Fig. 7.

The dense models all look greatly alike with the ground truth and SuperGlue resulting in more blades than SIFT similar to video 1. Looking at SuperGlue and the ground truth. The ground truth has more noise compared to the SuperGlue, but the blades themselves are slightly sharper.

For the sparse models a visual comparison between the ground truth, SIFT and SuperGlue can be seen in Fig. 8. The SuperGlue model seems to perform the worst here even though a rough outline of the blades can be seen. Comparing SIFT and the ground truth an observation that can be made is

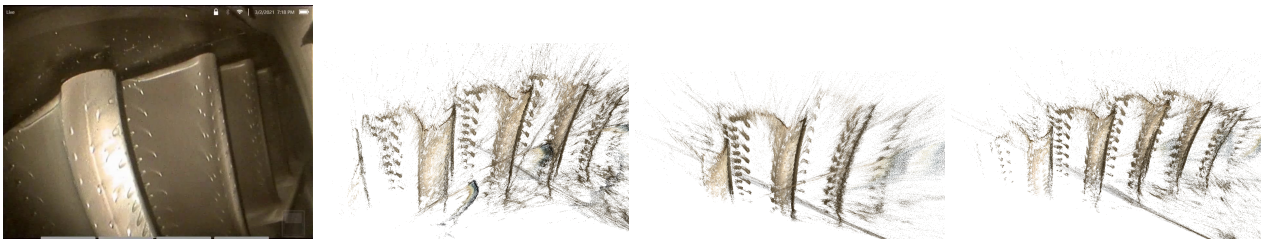

Figure 7: Video 3 comparison of reference video and dense models for GT, SIFT and SG (fLTR). The three models look greatly alike.

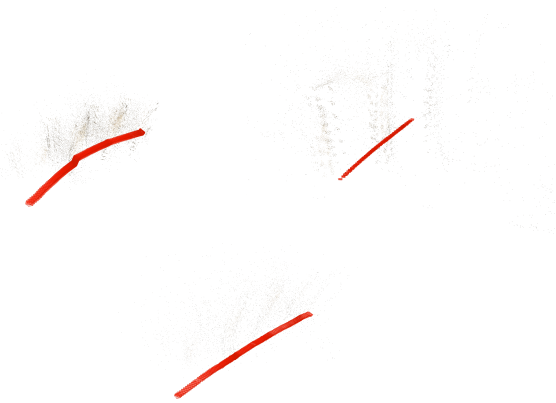

Figure 8: Video 3 sparse models of GT, SIFT and SG models with camera poses(fLTR).

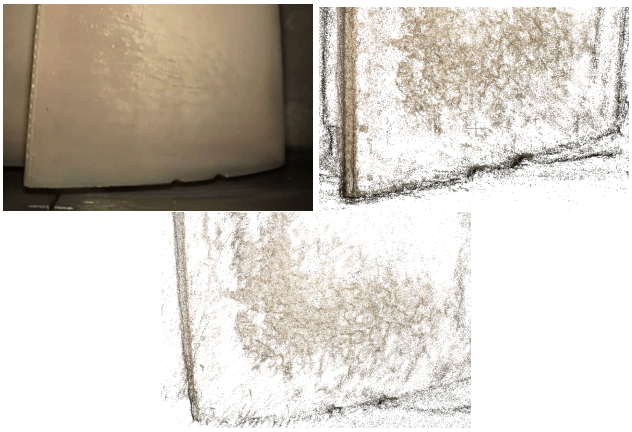

Figure 9: Damage in video 1 seen in the dense GT (right) and SG (bottom) models.

that the SIFT model is still a lot less dense compared to the ground truth model. However, the SIFT model captures the details of the blades better as it appears to have less noise than the ground truth model.

4.2 Damage visualisation

To qualitatively evaluate if damage can be assessed, some datasets were chosen where damage was clearly visible in the images. A side-by-side comparison between that image and the dense models can visualise how well incremental SfM can be used for damage assessment. Dents in the blade edges are visualised well in the ground truth as seen in the model of video 1 (Fig. 9). For the SuperGlue model, there is a visible change in density of the edge where the damage in the blade is, but it is substantially less clear that those density changes are meant to represent dents. However, the models are not dense enough to visualize any scratches or dents that don't alter the shape of the blades visibly.

4.3 Data analysis

Data results of the sparse reconstructions and thereby an evaluation of the incremental SfM system can be seen in Table 1. For each video, the best reconstruction resulting from the qualitative analysis is reported. A few standouts and takeaways will be highlighted here.

Global SfM experiments produced only one sufficient model for data analysis so no table is needed to lay out these results. The result of running global SfM with SIFT on video 3 contained 360 points with a mean reprojection error of 1.08 pixels.

On all three videos, LoFTR did not produce any results. For the first video, SIFT and SuperGlue do not register all images in the dataset resulting in fewer points but also a lower runtime and mean reprojection error compared to the ground truth. Video two only produced a reconstruction with the ground truth so not much comparison could be done between the different features. The third video had SIFT and Superglue performing substantially better compared to the first two with all images being registered and a significantly higher amount of points in the sparse point-cloud. SIFT here performing better than SuperGlue and also breaking the trend in the data of more points being registered resulting in a higher mean reprojection error.

5 Responsible Research

To make sure the results of this study are published responsibly some ethical aspects were considered while researching as well as the reproducibility of the research and experiments done. These considerations will be outlined in the following section.

Probably the most important ethical aspect to be considered is that the damage assessment of jet engines has to be done very thoroughly since a lot of human lives can be in danger when this damage is missed. Relying solely on the programs used in this study or any program that is not proved to be perfect would be extremely irresponsible. Therefore

	# Images	# Registered				#Points				Time (min)				Mean Reproj. Error (px)			
		SIFT	SG	LoFTR	GT	SIFT	SG	LoFTR	GT	SIFT	SG	LoFTR	GT	SIFT	SG	LoFTR	GT
Video 1	150	18	83	-	150	716	1497	-	52409	2.0	1.7	-	14.78	0.58	1.28	-	1.35
Video 2	150	-	-	-	80	-	-	-	10283	-	-	-	20.8	-	-	-	1.35
Video 3	150	150	150	-	150	6924	4741	-	20782	14.1	22.0	-	45.2	1.09	1.41	-	1.28

Table 1: Results for sparse reconstruction on video 1, 2 and 3 using SIFT, SuperGlue, LoFTR and a ground truth

any use of SfM techniques should be overseen by a human and any conclusions that a damage assessment program using SfM makes should be verified. Tying into this is the fact that two open-source implementations were used for this study, which raises the question of who is legally responsible when this software is used in production and fails. *COLMAP* and *OpenMVG* are licensed under the BSD and the MPL2 licenses respectively, which both indemnify every contributor of the software for any liability when these implementations are used in commercial products.

To ensure that the experiments done in this study are reproducible a lot of documentation was done while performing them. The open-source implementations used both have extensive documentation and any setting changes have been mentioned in section 4 of this paper. Software used to convert feature matching files received from peer researches to files that could be used by *COLMAP* can be shared. The videos used for experiments are not publicly available, but similar borescope videos can be found online.

6 Discussion

In this research, knowledge on the performance of SfM is extended with experiments that showcase the result of running SfM on borescopes videos with shiny and non-textured surfaces. Discussed in this section are an interpretation of the experiment results, limitations of these experiments, and future work that can be done on this topic.

The performance of SfM on borescopes videos with shiny and non-textured surfaces is considerably lower than reported SfM results run on more traditional datasets using the *COLMAP* and *OpenMVG* open-source SfM implementations as used in this research [14] and other studies done on the performance of SfM. Comparison of the amount of points in the model and run-time is difficult since these studies used much larger datasets and stronger machines to run the SfM algorithms on.

Looking at the difference in results between global and incremental SfM is difficult since few experiments were able to be done using global SfM. For both approaches SIFT performance was poor for videos 1 and 2, SIFT on video one performing slightly better for incremental SfM. On video 3 SIFT performed the best for both approaches. Here it is very clear that incremental SfM outperforms global SfM. Having only one experiment where this can be clearly seen makes it so no definite conclusions can be made about the relative performance of the two approaches, but the results from this single overlapping experiment do agree with existing literature which most often picks incremental SfM as the more robust approach.

Going more in-depth for incremental SfM, a comparison between the videos and features used shows SuperGlue and the ground truth performing the most consistent over all videos. SIFT does outperform SuperGlue and also the ground truth on accuracy through the lowest mean reprojection error on video 3. This suggests that the combination of edges and grooves in the blades itself increases the performance of SIFT because of the extra interest points it can detect.

The main goal of this research was to investigate if damage can be assessed using incremental SfM reconstructions. It was shown that large dents in the edges of the blades can clearly be seen which would suggest that this goal of damage assessment can be successful achieved. However, the clear visualisation of these dents only shows that a human can detect those types of damage in the model. Since the models all had quite some noise, future algorithms that look for dents in these models might be unsuccessful.

For the qualitative analysis, a major aspect that might hinder the accurate evaluation of incremental SfM is the use of MVS to create dense models. An explanation of why MVS was used is written in 3.3, but the result of using it is that bad sparse reconstructions can get 'saved' by running MVS on it. They can only be saved however, if the camera poses are correct and in most cases, the best sparse result corresponded to the best dense results.

In the evaluation of incremental SfM, it is assumed that for these videos the experiment results are the best possible results that could be achieved using incremental SfM on the borescope videos. However, settings had to be changed from the default to improve the reconstructions or in a few cases even obtain a reconstruction. These settings were picked out and changed by experimentation which means they are not necessarily the setting changes that achieve the best results. Also, intrinsic parameters of the camera were not known so could not be used as input during experiments for both SfM approaches. Literature shows that knowing these intrinsic parameters improves the performance of SfM. Further investigation can be done into these points.

Another assumption that was made is that the 'ground truth' used for evaluation provided by peer research is really the ground truth for the feature matching. Creating a feature detection ground truth that is perfect is challenging and it being a 10 week research project, a perfect ground truth is not realistic to expect. That the provided ground truth is not perfect can be seen in the results of video 3 where using SIFT resulted in a lower mean reprojection error compared

to the ground truth. There is also no guarantee that a perfect feature matching ground truth would even translate into a perfect ground truth model. In the future better ground truth feature matching or other ground truth methods, like utilizing synthetic data as researched by E. Klein Onstenk in a peer research [10], could be used.

7 Conclusions

In this paper Structure from Motion (SfM) is evaluated in the context of jet engine borescope videos with shiny and non-textured surfaces. Small experiments were performed using a global approach to SfM and more extensive research was conducted on incremental SfM. The goal of this research was to determine if SfM works well enough to aid in damage assessment of jet engines. Different feature matching methods and different borescope videos were compared qualitatively and quantitatively.

Results indicate that global SfM using SIFT underperforms compared to incremental SfM due to a lack of robustness against outliers in feature matches. In general, incremental SfM shows promising results when using SuperGlue for feature matching. SIFT only shows these promising results if blades contain enough large discernible features like grooves. Experiments performed show that incremental SfM definitely has potential as a tool for aiding damage assessment or as a method of assessing damage by itself. To be able to asses damage itself, work needs to be done on reducing noise in the models. Further work can be done to improvethe evaluation of SfM such as using improved ground truth and utilizing point cloud comparison techniques. The performance of SfM can also be improved in future researches by using known intrinsic parameters of the borescope cameras instead of having SfM estimate them.

References

[1] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[2] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 864–872, 2015.

[3] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[4] Yasutaka Furukawa and Carlos Hernández. Multi-view stereo: A tutorial. *Found. Trends. Comput. Graph. Vis.*, 9(1–2):1–148, June 2015.

[5] Rick Huizer. How well does interest point detection/matching work on shiny and non-textured surfaces?, 2021.

[6] Aiir innovations. Bringing artificial intelligence to aviation. https://aiir.nl/.

[7] David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[8] Daniel Martinec and Tomas Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[9] Pierre Moulon, Pascal Monasse, and Renaud Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3248–3255, 2013.

[10] Eduard Klein Onstenk. Synthetic data for damage assessment in aircraft turbines, 2021.

[11] Onur Ozyesil and Amit Singer. Robust camera location estimation by convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[12] P. Sapirstein. Accurate measurement with photogrammetry at large sites. *Journal of Archaeological Science*, 66:137–145, 2016.

[13] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020.

[14] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016.

[15] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[16] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.

[17] Devin Lieuw A Soe. Ground truth for evaluating 3d reconstruction of jet engines, 2021.

[18] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *CoRR*, abs/2104.00680, 2021.

[19] Chris Sweeney, Tobias Höllerer, and M. Turk. Theia: A fast and scalable structure-from-motion library. *Proceedings of the 23rd ACM international conference on Multimedia*, 2015.

[20] Federico Tombari, Alessandro Franchi, and Luigi Di. Bold features to detect texture-less objects. In *2013 IEEE International Conference on Computer Vision*, pages 1265–1272, 2013.

[21] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In Bill Triggs, Andrew Zisserman,

and Richard Szeliski, editors, *Vision Algorithms: Theory and Practice*, pages 298–372, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[22] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 61–75, Cham, 2014. Springer International Publishing.

[23] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision - 3DV 2013*, pages 127–134, 2013.

[24] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2018.