Occlusion handling in spatio-temporal object-based image sequence matching

Simon Nietiedt [1], Petra Helmholz [2], Thomas Luhmann [1]

[1] Institute for Applied Photogrammetry and Geoinformatics, Jade University of Applied Sciences, Oldenburg, Germany
[2] Spatial Sciences, School for Earth and Planetary Sciences, Curtin University, Australia

KEY WORDS: Close-range photogrammetry, dynamic, occlusion, image matching, robust optimisation.

ABSTRACT:
Dynamic photogrammetry is an established method for acquiring 3D information of deforming objects or dynamic scenes in various close-range applications. A crucial impact has occlusions caused by object deformations, obstacles or camera movements. Temporal occlusions are highly application-specific and sometimes difficult to predict, resulting in a significant reduction of reconstruction quality or the aborting of image sequence processing. Previous approaches usually model such occlusions as semantic information and consider them using image masks. However, generating these image masks requires complex methods and extensive training data. Due to the unpredictability of the complexity and movements of dynamic scenes, generating training data is challenging in many applications. Therefore, this paper proposes an alternative modelling approach, which can be part of a spatio-temporal matching process. Based on the characteristic high redundancy, occlusions can be detected using robust estimation methods and considered in the optimisation. Therefore, no information about the occlusions and further processing steps are necessary. We evaluate our approach with synthetic and real data of an industrial application regarding the accuracy and ability to detect occlusion simultaneously. The evaluation of the proposed approach shows that the impact of occlusion can be eliminated, and the quality of the results is comparable to conventional methods.

1. INTRODUCTION

Dynamic photogrammetry is an established method for the precise reconstruction of dynamic scenes in many applications. Especially in high-speed processes where area-based information is required, photogrammetry is used more and more (Luhmann et al., 2023). In general, the photogrammetric processing of image sequences to create 3D trajectories in close-range applications consists of several steps. Once the data has been acquired, image pre-processing and system calibration are required. Then, spatial and temporal matching procedures are carried out. The matching steps are usually the most computationally expensive and can be performed simultaneously and independently of each other. However, the spatial and temporal matching are based on the extracted image features or image intensities, resulting in high spatio-temporal redundancy. This redundancy can be used as a motion model to support the spatial and temporary matching process and reduce ambiguities. In general, the dynamic characteristics lead to additional requirements for the measurement technology and algorithms that surpass the requirements for reconstructing static scenes. For example, the synchronisation of the cameras, handling possible motion blur, and storing large amounts of data must be considered. In specific applications where repeatability cannot be guaranteed, temporary occlusion can pose a significant challenge. For example, in materials testing, the probe is destroyed during the test (Hampel and Maas, 2009; Guccione et al., 2020). The presence of obstacles can lead to a loss of image data and affect the quality of the results. A similar situation applies to car safety tests, where the dummy's movement and the footwell's deformation are interesting (Raguse et al., 2004). However, temporary occlusions caused by movements of the object or the camera itself and occlusions caused by damaged components can reduce the quality of the image sequences or even lead to aborted processing. Consequently, this can result in higher costs or reduce the usage of photogrammetric methods.

Therefore, reliable detection is an essential task, and several approaches have been developed. In general, solution strategies can consist of purely image-based approaches or a combination of image and object-based methods. A purely image-based approach is to detect occlusions through semantic segmentation, which can be generated by various Machine Learning approaches (Wu and Nevatla, 2009; Saleh et al., 2021). Afterwards, corrupted pixels can be excluded from further processing steps. However, application-specific training data is required to generate accurate results. Alternative approaches use the optical flow that produces information about movements. If the optical flow is performed back and forth (in both sequence directions), occlusions can be classified using an energy function (Alvarez et al., 2007). Rúa et al. (2016) use parametric motion models to check the plausibility of the flow and detect occlusions based on it. The occlusion detection can be improved if a CAD model of the obstacle (Bethmann et al., 2009) or the deforming object (Malti et al., 2011) is available. For this purpose, a Kalman filter is used to predict the CAD model. Then, the prediction will be transformed into the image domain, where corresponding positions are excluded from the processing.

Despite the wide range of applications, these approaches have one thing in common: the state of a pixel (occluded or not) is represented by an image mask. The image mask can be used in the respective matching process to determine whether the corresponding pixel will be considered, which can increase the complexity and computational time. If the reconstruction of the object surface is combined with the tracking, a spatio-temporal matching method is formed, enabling better utilisation of the spatio-temporal redundancy. In the spatio-temporal approach published by Ngo et al. (2015), all observations are used in the numerical optimisation. A relevancy score, calculated by normalised cross-correlation, weights the observations to control the impact of corrupted observations. However, the method assumes that the static object surface is known. Instead, Lin et al. (2022) use an RGBD sensor to derive spatio-temporal information from the sensor data. Here, the object motion is modelled using a graph whose spatio-temporal consistency is optimised using a Long Short-Term Memory (LSTM) model. The network is trained in a supervised manner. Therefore,

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-163-2024 | © Author(s) 2024. CC BY 4.0 License.

training data is required for this approach, which is limited or unavailable in many dynamic applications such as wind energy science or car safety. As a result, the generation of generic training data is also difficult. Therefore, previous Machine Learning solutions can probably only be used successfully in new applications through complex adaptation.

This paper introduces an alternative approach that significantly minimises the impact of temporal occlusions in dynamic close-range applications without requiring prior knowledge about the behaviour of occlusions. The method is based on a novel spatio-temporal matching procedure that takes advantage of the high spatio-temporal redundancy and eliminates the impact of occlusions using classic statistical methods. The paper begins by describing the spatio-temporal matching procedure and robust estimation methods. Subsequently, the synthetic and real datasets are explained, followed by the evaluation process. Finally, the performance of the proposed approach is demonstrated through a comparison with a Machine Learning method based on image masks.

2. OBJECT-BASED IMAGE SEQUENCE MATCHING

While classical approaches of dynamic photogrammetry in close-range applications perform reconstruction and tracking separately, various methods can exploit the full potential of spatio-temporal redundancy. Some methods, called non-rigid structure from motion (NRSfM), aim to determine a dynamic scene and the unknown camera positions (Parashar et al., 2018; Kong and Lucey, 2019). Other methods use a known rigid object surface instead (template-based). Here, a spatio-temporal model is used, describing the template's deformation. Based on this, a cost function can be formulated to describe the photo consistency of homologous image intensities, matched by the template and spatio-temporal model. This cost function is minimised, and the unknown spatio-temporal parameters are determined by numerical optimisation, where smoothness constraints are considered (Yu et al., 2015).

This concept can be adapted to object-based image matching to reconstruct dynamic surfaces with an established photogrammetric method. In object-based image matching, the colour intensities of all images are matched, and unknown model parameters of radiometry, object surface, and cameras are estimated simultaneously (Wrobel, 1987). The model can be extended for dynamic surfaces by a kinematic model, which describes the movement/deformation of the rigid object surface. This description results in a closed mathematical formula, shown in Figure 1.

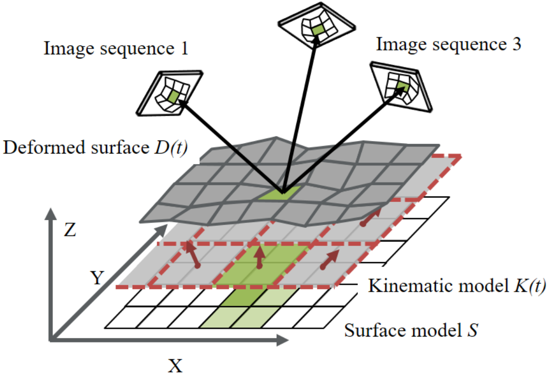

Figure 1. Relationship of image sequence domain and object-spaced model. The model consists of several patches.

The mathematical formulation of spatio-temporal object-based image sequence matching describes the relationship between colour intensity g_{ij} in an image sequence j and the corresponding surface texture $G_i(t)$ of the dynamic surface point $D_i(t)$, which consists of a geometric model S and a kinematic model K.

$$G_i(t) = g_{ij}(S, K, O, R, t) \tag{1}$$

As with the original approach, a geometric model forms the basis of the method, which determines any 3D object point $\vec{p}$. In the past, numerous models have been developed that differ in dimensionality (2.5D or 3D) and computational time (Heipke, 1991; Schneider, 1991; Schlüter, 1999). As a geometric model, we use independent 2.5D planes to approximate the rigid surface. Each plane is described by the height Z of the centre point $\vec{p}_P$ and the angles α and β. The required coordinates X and Y are assumed to be known of $\vec{p}_i$ and centre point $\vec{p}_P$.

$$S :\Leftrightarrow\ Z_i(X,Y) = \underline{Z_P} + (X_i - X_P)\cdot tan(\underline{\alpha}) + (Y_i - Y_P)\cdot tan(\underline{\beta}) \tag{2}$$

Each object point $\vec{p}_i$ can be explicitly associated with a static plane P, whose motion is described by the kinematic model K. Depending on the application, several planes can use the same motion parameters (see the highlighted area in Figure 1). In this study, the kinematic model is modelled by a 6-parameter similarity transformation, consisting of translation $\overrightarrow{x(t)}$ and rotation $R(t)$. However, each set of parameters is only valid for one discrete time step t and a discrete area of the surface model.

$$K :\Leftrightarrow\ \vec{p}_{di}(t) = \underline{\overrightarrow{x(t)}} + \underline{R(t)}\cdot \vec{p}_i \tag{3}$$

After applying the transformation, the transformed points $\vec{p}_{di}(t)$ defines the deformed surface $D(t)$. These discrete points can be transformed into the image domain of the respective image sequence using the collinearity equations and the orientation parameters O of the camera j. The colour intensities are generated by bilinear interpolation in the image domain. In addition, radiometric correction terms r_{0j} and r_{1j} can be applied through the radiometric model R. The complexity of these correction terms varies depending on the application. Different radiometric models can be found in Weisensee (1992) and Gehrke (2008).

$$R :\Leftrightarrow\ g_{ij} = \underline{r_{0j}} + \underline{r_{1j}}\cdot g'_{ij} \tag{4}$$

The underlined parameters in Equations 1 to 4 are unknown and determined using a least squares adjustment. The observation equation can be derived from Equation 1. Here, the colour intensities of the image sequences are defined as observations so that the primary observations of the cameras can be used directly. Furthermore, this description allows the use of the entire spatio-temporal redundancy. Another advantage of the method is the dynamic scene's modelling, which allows comprehensive error propagation. In addition, all image sequences are matched equally, so no reference image is needed. Hence, the method is characterised by great flexibility and highly achievable accuracy. However, the method needs initial values and long computational times.

3. OCCLUSION DETECTION

In the context of this work, we assume that occlusions can be reduced to obstacles with an unknown motion model. In addition, the obstacle's texture is assumed to differ from the dynamic object's surface texture. The following describes two approaches to handling occlusion in the spatio-temporal matching method.

3.1 Semantic approach

The semantic approach refers to the use of binary image masks. For this purpose, we use the network architecture of the SegFormer network, which is characterised by a small number of trainable parameters (Xie et al., 2021). The neuronal network processes each image of a sequence and generates a sequence of binary image masks. Occluded pixels, which are supposed to be processed by the spatio-temporal matching method, are thus excluded from the optimisation process.

3.2 Statistical approach

We define another approach as the statistical method based on robust estimation procedures. Such methods are characterised by the ability to separate observations from outliers and robustness regarding deviations from the expected distribution of observations (Huber and Ronchetti, 2009). For example, RANSAC algorithms are widely applied methods that allow good parameter estimation regardless of many outliers (Fischer and Bolles, 1981). However, RANSAC is characterised by high computational times and additional memory requirements, which would further increase the computational times of the proposed spatio-temporal matching method. An alternative are M-estimators that are based on the maximum likelihood principle. Here, a loss function is used for parameter estimation that maximises the likelihood function, which allows a parameter estimation even though the underlying observation distribution is disturbed. Thus, M-estimators have already been successfully investigated and used in stereo matching (Arya et al., 2007; Li and Wang, 2014).

Nevertheless, to our knowledge, no publications are related to spatio-temporal object-based image sequence matching in which robust estimation methods are used to handle corrupted observations. M-estimators are particularly suitable for spatio-temporal object-based image sequence matching because they can be easily integrated into the least-squares method. Equation 5 shows the generic estimation of unknown parameters using the least-squares method.

$$\vec{x} = (J^T \cdot J)^{-1} \cdot (J^T \cdot \vec{b}) \tag{5}$$

Here $\vec{x}$ describes the parameter vector, J the Jacobian matrix, and $\vec{b}$ the observations. In robust estimation, this equation is extended by a weight matrix W, whereby the influence of each observation can be individually controlled. The parameter vector is determined following the principle of iteratively reweighted least squares (IRLS, Holland and Welsch, 1977), where the determination of the weight matrix is repeated simultaneously in each iteration.

$$\vec{x} = (J^T \cdot W \cdot J)^{-1} \cdot (J^T \cdot W \cdot \vec{b}) \tag{6}$$

In the first iteration, the weight matrix corresponds to an identity matrix whose diagonal is redefined by a weight function $w(\cdot)$ in the next iteration.

$$W = \begin{pmatrix} w_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{nn} \end{pmatrix} \; ; \; w_{ii} = w(r_i) \tag{7}$$

The weight function is the second derivation of a loss function, for which numerous variants have been proposed (Holland and Welsch, 1977). The weights are usually based on the respective residual r_i and a threshold c known as the tuning factor. We refer to Huber and Ronchetti (2009) for more information about the tuning factor. Two of the most widely used functions are the Huber (Equation 8) and the Tukey weight function (Equation 9). While the Huber weight function reduces the weights of outliers, the Tukey weight function modifies the weights of valid observations. At the same time, the outliers are given a weight of zero. However, this can lead to unstable equation systems. Thus, we use a weight of 10^{-6} instead.

$$w(r) = \begin{cases} 1 & |r| \leq c \\ \frac{c}{|r|} & |r| > c \end{cases} \qquad ; c = 1.345 \tag{8}$$

$$w(r) = \begin{cases} \left[1 - \left(\frac{r}{c}\right)^2\right]^2 & |r| \leq c \\ 0 & |r| > c \end{cases} ; c = 4.685 \tag{9}$$

In the context of the requirements for temporal occlusions described in Chapter 3, the occlusions can be defined as outliers in the described matching procedure. The occlusions can be detected simultaneously with the parameter estimation through the statistical approach, and the influence of occlusions can be minimised. That means that no additional processing of the image data or prior information about the characteristics of the temporal occlusions is required. It should be noted that it is impossible to separate the occluded pixels from other outliers. As a result, a higher false negative rate can be assumed for the classification of occlusions compared to the semantic approach.

4. EXPERIMENTS

In the following experiments, we investigated the performance of the proposed method using synthetic and real data. The datasets simulate a modal analysis, a widely used experiment in wind energy science to verify the quality of operating and new rotor blades (Larsen et al., 2002). A modal analysis aims to evaluate the behaviour of a rotor blade by occurring vibrations. Minor vibrations can be expected at the root and the highest at the tip position of the blade. Different measurement systems can acquire the vibrations that occur. Photogrammetric methods are especially suitable for the area-based acquisition of displacements. However, repeatability cannot guaranteed due to changing environmental conditions. Therefore, precise calibration and high-quality processing strategies are necessary to achieve the required accuracy (Sabato et al., 2018).

4.1 Synthetic dataset

The modelling and creation of the synthetic dataset and the reference masks were carried out with Blender (Blender Development Team, 2022). A CAD model of a NREL 5MW turbine was used to model the rotor blade. For this experiment, the model was scaled to 900 mm. To add the deformation to the blade, we used the modifier *SimpleDeform Bend*. The largest deviations were 31.84 mm at the blade's tip and showed a sinusoidal pattern over time. A speckle pattern signalises the blade, interpreted as a diffuse Lambertian emitter.

The photogrammetry system consisted of four cameras corresponding to the high-speed camera PCO Dimax HD+ specifications, summarised in Table 1. To simplify the processing, no distortion model was applied to the image sequences.

Sensor size [mm]	21.12 x 15.84
Sensor size [pixel]	1920 x 1440
Pixel size [mm]	0.011
Focal length [mm]	24
Principal point [mm]	-0.0055 (X) 0.0055 (Y)

Table 1. Specification of the camera Dimax HD+ and CS3.

Using four perspective views, which ensured that the whole rotor blade was captured in each view and time step. The average distance to the object was 1500 mm, resulting in a GSD of 0.72 mm. To optimise the runtime of the procedure, only the red area marked in Figure 2 was reconstructed.

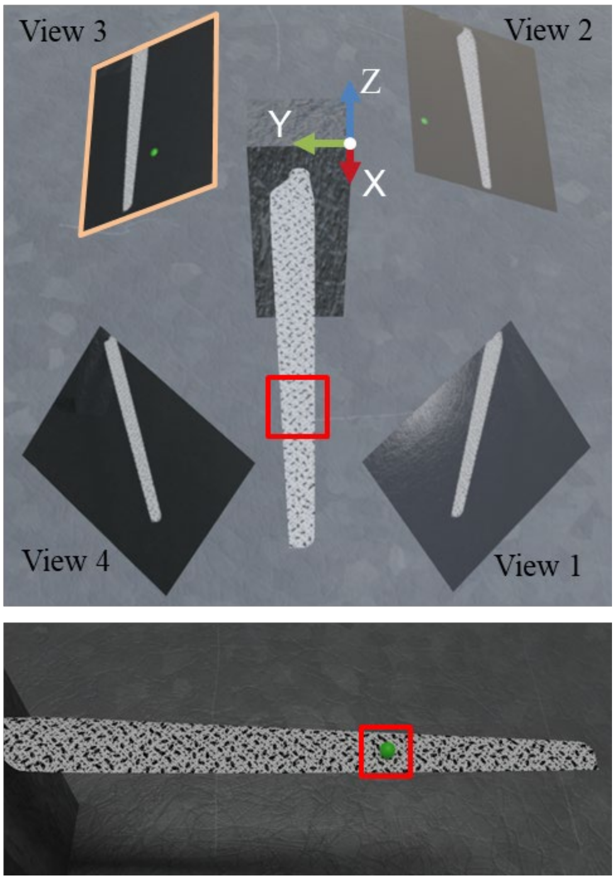

Figure 2. Image configuration of the synthetic dataset (top). View 3 (orange) with the reconstruction area and the occlusion (bottom).

The size of a geometric patch was 15 mm x 15 mm, so the geometric model consisted of 254 elements. The kinematic model was structured in strips along the Y-axis, resulting in a varying number of geometric meshes per kinematic patch. Each kinematic patch consisted of at least 13 geometric patches, which led to a different number of observations. The general specifications of the adjustment are summarised in Table 2.

A green obstacle was modelled as an occlusion. This obstacle dropped alongside the Z-axis, resulting in occlusions in epoch 4 and obstructing 25% of the observation per view. The dataset was used in three experiments, where the number of occluded observations varied. In experiment A, occlusions were only included in view 3. In experiment B, additional occlusions were added in view 4. Finally, in experiment C, occlusions were present in views 2, 3, and 4. This means that for experiment C, only view 1 provided observation for the reconstruction, and consequently this meant that the reconstruction was based on view 1 and the spatio-temporal redundancy.

4.2 Real dataset

The real dataset simulated a modal analysis of a metal bar in the laboratory, shown in Figure 3. For data acquisition, two PCO Dimax HD+ and two PCO CS3 cameras were used, and they were applied in a similar image configuration to the synthetic dataset. The camera system was calibrated through bundle adjustment and synchronised with an external device. A hammer hit the mounted metal bar and deformed the object. Therefore, the metal bar oscillated only in the Z-direction (blue arrow) during the experiment, where the most significant deviation occurred at the tip position (horizontal metal bar).

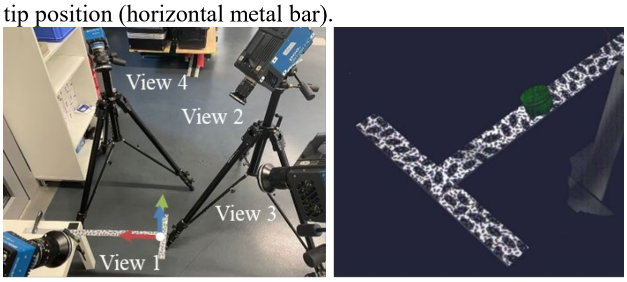

Figure 3. Experiment setup of real dataset (left). View 2 with the green obstacle and shaded area (right).

The kinematic patches were arranged along the X-axis (red arrow), similar to the synthetic dataset. The GSD was 0.25 mm, and each image sequence consisted of 8 images. This specification led to more observations and parameters, summarised in Table 2.

	Synthetic dataset	Real dataset
Sequence length	6	8
Colour depth	3 (RGB)	3 (RGB)
Number of geometric patches	254	219
Number of kinematic patches	18	21
Total number of observations	1,407,168	5,797,824
Total number of parameters	59,664	182,280

Table 2. Specification of the datasets.

The occlusion was based on a green obstacle that dropped alongside the Z-axis. The occlusions occurred in views 2 and 4 at time 4, 5, and 6. Furthermore, the obstacle created shadows at time 2 to 6, which could additionally reduced the reconstruction quality.

4.3 Evaluation procedure

The procedure's evaluation can be divided into three groups. First, the quality of the reconstruction was analysed. For this purpose, the point cloud of each epoch was compared with the simulation's reference data and evaluated using the RMS. The comparison of all epochs considers the influence of all models (geometric and kinematic). In addition, the number of iterations was considered but was of minor importance.

The second part of the evaluation included the quality of the segmentation. The evaluation aims to analyse the ability of the procedure to detect the occlusions. Precision, recall, and the IoU were used in the assessment. For this purpose, the observations were transformed into the image domain and compared with the reference masks. The classification with the robust methods

(Huber and Tukey) used the calculated weight. However, the weights are not binary, thus a threshold was required. Therefore, we used an adaptive threshold determined through Otsu's method (Otsu, 1979).
Image masks were required to compare with the semantic-based method. Therefore, the SegFormer network (Section 3.1) was used for the pixelwise detection of the objects. The training used 600 synthetic images, which show different numbers of rotor blades, perspective and radiometry. After training, the network was characterised by a precision of 99.6 %, a recall of 95.4% and an Intersection over Union (IoU) of 0.95.

The third part of the evaluation includes the real dataset. However, due to the dynamic and uniqueness of each experiment, generating highly accurate reference data for the real dataset was not possible. Therefore, only the residuals of each method were used to validate the procedure.

5. RESULTS

5.1 Synthetic dataset

The results of the three synthetic experiments were analysed and summarised in Table 3. The presented metrics were determined for all methods and experiments. Here, the *Standard* method shows the results of the spatio-temporal matching method described in Chapter 2. *Ground Truth* differs from *Standard* by using occlusion-free image sequences. Therefore, the achievable accuracy level can be seen as the best result based on the configuration. The RMS was 0.02 mm, which resulted in a reprojection error of 0.03 pixels. The accuracy level decreased significantly with increasing occlusions. The RMS decreased by factors of 2.5, 4, and 5 due to the different experiments. In addition, considerably more iterations were required until the method converged. However, the three methods can significantly reduce this effect, allowing high-quality experiment results.

Differences can be seen between the methods regarding the deviations from the CAD model, which are visualised in Figure 4. Figure 4 shows the reconstructed surface for all time steps of experiment C, coloured according to the deviations. Although the occlusions only occurred at time 4, high deviations are visible at all time steps with the *Standard* method. The largest deviations occurred in the area of the occlusions. Therefore, the occlusions seem to influence the affected positions independent of time due to the spatio-temporal relationship. If the occlusions were considered through image masks, the areas with high deviations were significantly smaller. Individual deviations occurred in the occlusion area, presumably attributed to incomplete segmentation. The use of robust estimation methods resulted in even fewer deviations and shows a high level of accuracy comparable to *Ground Truth* throughout all epochs. In addition to the RMS values, Table 3 summarises the segmentation metrics. High segmentation ratios can be achieved using the *Mask* method, which has the same quality as the training dataset. The robust estimation methods *Huber* and *Tukey* show a slightly different performance but on a same level. Figure 5 visualises the weights of the respective detected occlusions. Compared to *Ground Truth* (GT), *Mask* did not detect the occlusion correctly, although the metrics in Table 3 show a very high level. It should be noted that the results of all image sequences were cumulated to determine the metrics, which means that local discrepancies have a minor impact. If the metrics were determined only for the affected images at time 4 the IoU for experiment A would be 0.89 (precision: 94.60% - recall: 94.39%), experiment B 0.95 (precision: 97.64% - recall: 97.55%) and experiment C 0.87 (precision: 93.11% - recall: 92.56%). From our point of view, this is not suitable for the analysis because the spatio-temporal method considered all images of the sequences equally. Therefore, spatio-temporal redundancy allowed for the achievement of high-quality results even with less reliable detection. Nevertheless, it showed the importance of high-quality training data for the segmentation task.

By comparing the reference with the calculated weights by *Huber*, it can be seen that the whole occlusion can be detected. However, noise could also be recognised in all image views. This noise cannot be seen in the *Tukey* weighting, which the higher tuning factor can explain. In addition, the occlusion can be detected comprehensively, although the occlusion edge was weighted slightly higher.

Experiment	Method	RMS [mm]	Precision [%]	Recall [%]	IoU	Iteration
Ground Truth		0.02	-	-	-	10
A	Standard	0.05	-	-	-	27
	Mask	**0.02**	**99.57**	99.72	**0.99**	**10**
	Huber	**0.02**	90.83	**99.99**	0.91	21
	Tukey	**0.02**	95.90	**99.99**	0.96	14
B	Standard	0.08	-	-	-	19
	Mask	0.03	**99.37**	99.53	**0.98**	14
	Huber	**0.02**	94.83	**99.99**	0.95	17
	Tukey	**0.02**	95.89	99.96	0.95	**13**
C	Standard	0.10	-	-	-	36
	Mask	**0.03**	**99.10**	99.27	**0.98**	**14**
	Huber	**0.03**	96.31	**99.98**	0.96	23
	Tukey	**0.03**	95.96	99.93	0.96	35

Table 3. Results of spatio-temporal object-based image sequence matching. The best results are marked in bold.

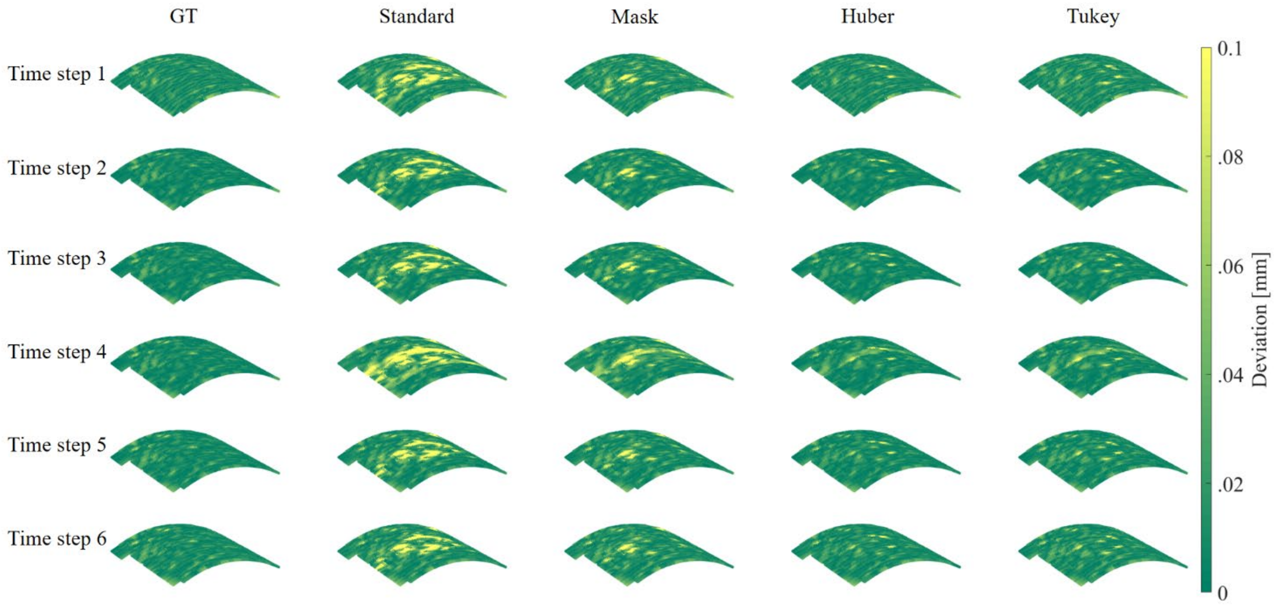

Figure 4. Deviations from the reference CAD model for experiment C. Modelled occlusions were in time step 4.

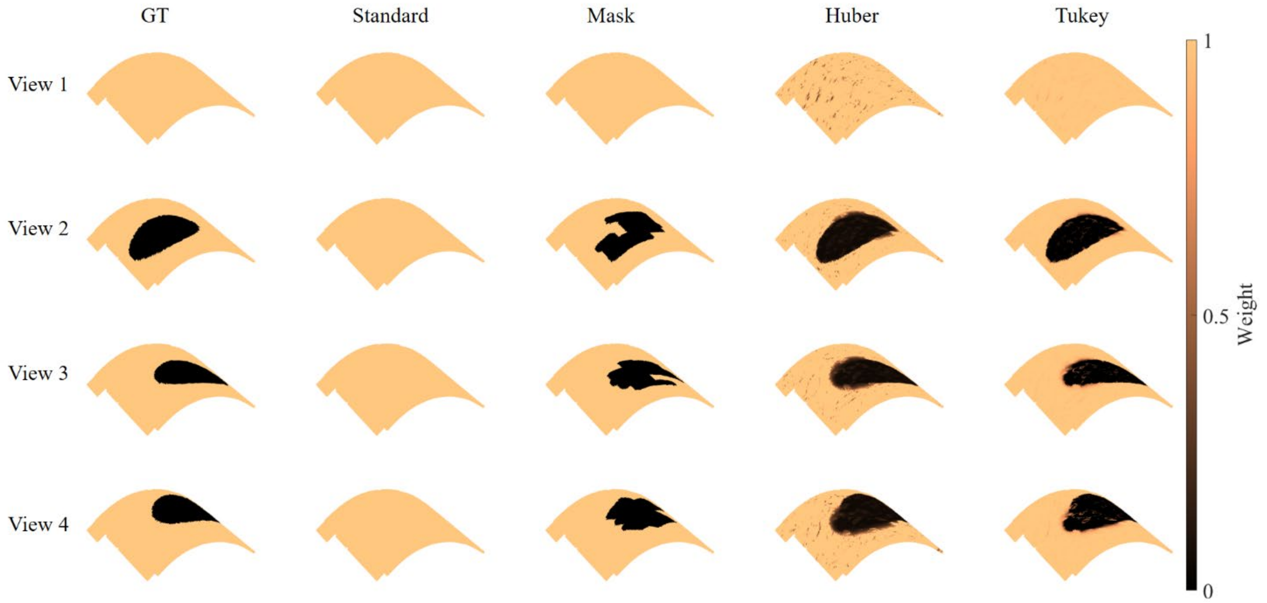

Figure 5. Colourised detected occlusion (black) through weights. Correct observations are bright, and corrupted observations are dark.

5.2 Real dataset

Figure 6 shows the reconstructed surface of the real dataset. The surfaces are colourised based on the residuals of the view's 2 observations for time 2 to 7. The method *Standard* (first row) showed high residuals in the occlusion and shadow area. In addition, all residuals were higher than the other methods, leading to a σ_0 of 21.36 intensity. Furthermore, the method converged after 37 iterations. The reconstructed object with the *Mask* method obtained some gaps due to eliminating invalid pixels. In addition, high residuals could be seen in some areas caused by the inefficient detection of shaded areas. However, only 24 iterations were needed to obtain a σ_0 of 17.65 intensity. *Huber* and *Tukey* achieved a similar performance. The methods converged after 35 and 28 iterations and reached a σ_0 of 15.26 and 15.73 intensity. Contrary to *Standard*, the residuals were homogenous for the whole surface. Although minor artefacts in the shaded area could be seen in *Tukey*, the reconstructed surface was of high quality.

Similar behaviour could be observed compared to the synthetic dataset, and the results from the synthetic experiments could be confirmed. In both datasets, occlusion's impact was significantly reduced with robust and masked-based methods. Although the masked-based approach required fewer iterations, robust methods detected all corrupted observations. However, both approaches increased the quality and allowed high-accuracy reconstruction of partly-occluded dynamic surfaces.

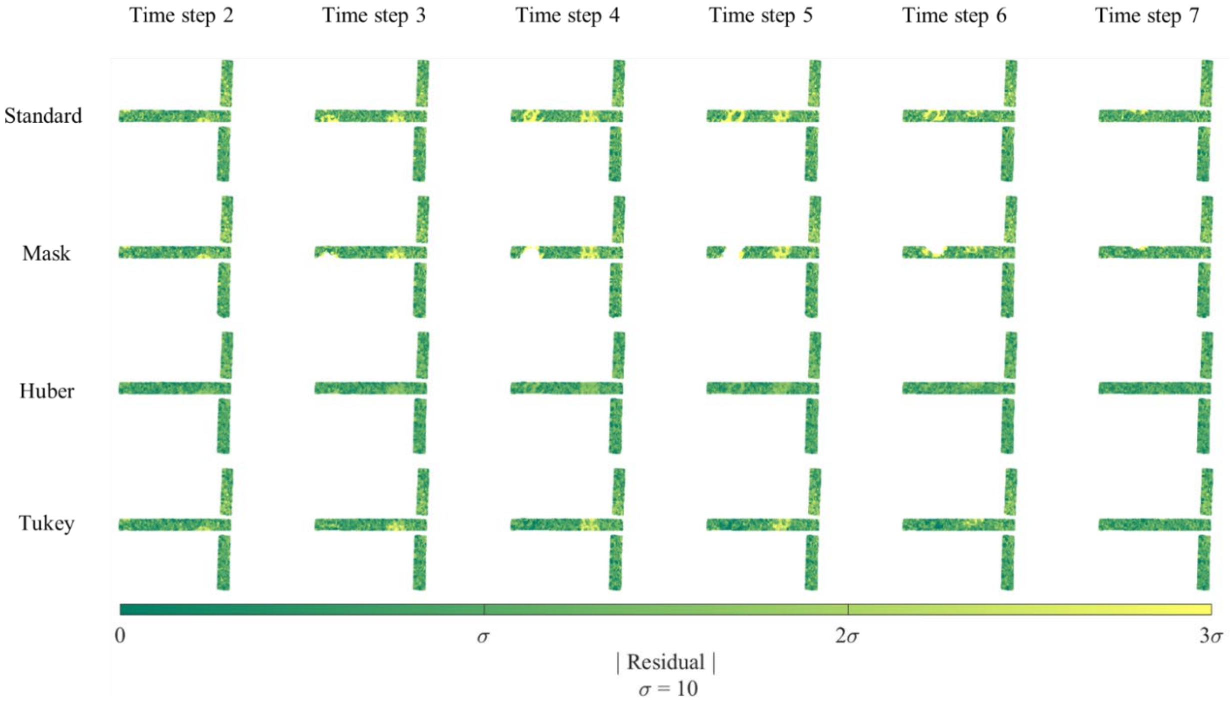

Figure 6. Reconstructed surface of the real dataset, colourised by the residuals of view's 2 observations.

6. CONCLUSION

Temporal occlusions can significantly affect the reconstruction process of dynamic scenes. In the past, image masks were often used to detect occluded pixels and exclude them from the reconstruction process. However, this requires prior knowledge of the occlusion and additional processing steps.

This paper proposed an alternative framework that exploits high spatio-temporal redundancy and uses robust estimation methods. The framework extends an established object-based image matching approach with a kinematic model, which allowed the processing of dynamic scenes. We combined the framework with different methods to handle occlusions. Moreover, we evaluated those approaches using synthetic and tested them on real data. It was shown that the spatio-temporal matching method reconstructs the dynamic surface with high quality but was sensitive to occlusions. Robust and mask-based methods significantly increased the accuracy in all experiments so that the influence of temporal occlusions was reduced entirely. Even with occlusions in 75% of all images of an epoch, the reconstruction could be achieved without significant loss of accuracy, which can be explained by high spatio-temporal redundancy.

Distinctions between the methods could be seen in the reconstruction results. On the one hand, the robust methods required a few more iterations than the mask-based approach. On the other hand, inaccurate image masks could corrupt the result or led to gaps in the reconstructed surface, especially in the case of poorly identifiable artefacts such as shadows. Therefore, an accurate segmentation method should be used if high-quality semantic information is required. In summary, both approaches were equivalent, and their use should depend on the application.

The investigated setup is based on the modal analysis of a rotor blade, where an artificial object generates the occlusions. Therefore, it is still interesting whether similar results can be achieved if the occlusion has radiometric properties similar to those of the reconstruction object. Due to the image configuration and the local radiometric uniqueness, similar results can probably still be achieved. However, this requires further experiments.

Upcoming steps will focus on expanding the proposed framework. For example, measured velocities or rotations of the object can be integrated into the optimisation process to improve the reconstruction quality.

ACKNOWLEDGEMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project number 497532406) and partly funded by the Jade University of Applied Sciences. We also thank the colleagues of the photogrammetry group for their support during the experiments and the helpful discussions.

REFERENCES

Alvares, L., Deriche, R., Papadopoulo, T., Sánchez, J., 2007: Symetrical Dense Optical Flow Estimation with Occlusions Detection. *International Journal of Computer Vision*, 75(3), 371-385. DOI: 10.1007/s11263-007-0041-4.

Arya, K. V., Gupta, P., Kalra, P. K., Mitra, P., 2007: Image registration using robust M-estimators. *Pattern Recognition Letters*, 28, 1957–1968. doi:10.1016/j.patrec.2007.05.006.

Bethmann, F., Herd, B., Luhmann, T., Ohm, J., 2009: Free-Form Surface Measurement with Image Sequences under Consideration of disturbing Objects. *Proceedings 9th Conference on Optical 3-D Measurement Techniques*, 51-6.

Blender Development Team, 2022: Blender Software, Version 3.3.1. Blender Foundation, blender.org (28 January 2023).

Fischler, M. A.; Bolles, R. C., 1981: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, *24*(6), 381-395.

Gehrke, S., 2008: Geometric and radiometric modeling of the martian surface based on object space matching and photoclinometry. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XXXVII, 1031-1036.

Guccione, D. E., Thoeni, K., Giacomini, A., Buzzi, O., and Fityus, S., 2020: EFFICIENT MULTI-VIEW 3D TRACKING OF ARBITRARY ROCK FRAGMENTS UPON IMPACT. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 589–596. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-589-2020.

Hampel, U., Maas, H.-G., 2009: Cascaded image analysis for dynamic crack detection in material testing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64, 345-350. doi:10.1016/j.isprsjprs.2008.12.006.

Heipke, C., 1991: Integration von Bildzuordnung, Punktbestimmung, Oberflächenrekonstruktion und Orthoprojektion innerhalb der digitalen Photogrammetrie. PhD-Thesis, DGK-Reihe C, 366.

Holland, P. W., Welsch, E., 1977: Robust regression using iteratively reweighted least-squares. Communications in Statistics – Theory and Methods, 6(9), 813-827. https://doi.org/10.1080/03610927708827533.

Huber, P. J., Ronchetti, E. M., 2009: *Robust Statistics – Second Edition*. John Wiley & Sons, Hoboken, NJ, USA.

Kong, C. and Lucey, S., 2019: Deep Non-Rigid Structure From Motion. *IEEE/CVF International Conference on Computer Vision (ICCV),* 1558-1567.

Larsen, G. C., Hansen, M. H., Bumgart, A., Carlén, I., 2003: Modal analysis of wind turbine blades. Forskningscenter Risoe, Risoe-R No. 1181(EN).

Li, Z., Wang, J., 2014: Least squares image matching: A comparison of the performance of robust estimators, *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, II-1, 37–44. https://doi.org/10.5194/isprsannals-II-1-37-2014.

Lin, W., Zheng, C., Yong, J.H., Feng, X., 2022: OcclusionFusion: Occlusion-aware Motion Estimation for Real-time Dynamic 3D Reconstruction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1736-1745.

Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2023: *Close-Range Photogrammetry and 3D Imaging*. 4th ed., de Gruyter, Berlin, Boston. https://doi.org/10.1515/9783111029672.

Malti, A., Bartoli, A., Collins, T., 2011: A pixel-based approach to template-based monocular 3D reconstruction of deformable surfaces. *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 1650-1657. doi: 10.1109/ICCVW.2011.6130447.

Ngo, D. T., Park, S., Jorstad, A., Crivellaro, A., Yoo, C. D., Fua, P., 2015: Dense image registration and deformable surface reconstruction in presence of occlusions and minimal texture. *Proceedings of the IEEE International Conference on Computer Vision,* 2273-2281.

Otsu, N., 1979: A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62-66. doi: 10.1109/TSMC.1979.4310076.

Parashar, S., Pizarro, D., Bartoli A., 2018: Isometric Non-Rigid Shape-from-Motion with Riemannian Geometry Solved in Linear Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40, 2442-2454. doi: 10.1109/TPAMI.2017.2760301.

Pérez-Rúa, J.M., Crivelli, T., Bouthemy, P., Pérez, P., 2016: Determining occlusions from space and time image reconstructions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1382-1391. doi: 10.1109/CVPR.2016.154.

Raguse, K., Derpmann-Hagenström, P., Köller P., 2004: Verifizierung von Simulationsmodellen für Fahrzeugsicherheitsversuche. *Publikationen der Deutschen Gesellschaft für Photogrammetrie, Fernerkundung und Bildverarbeitung*, Band 13, 367-374.

Sabato, A., Poozesh, P., Avitabile P., Niezrecki, C., 2018: Experimental modal analysis of a utility-scale wind turbine blade using multi-camera approach. *Journal of Physics: Conference Series*, 1149. DOI 10.1088/1742-6596/1149/1/012005.

Saleh, K., Szénási, S., Vámossy, Z., 2021: Occlusion Handling in Generic Object Detection: A Review. *EEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 477-484. doi: 10.1109/SAMI50585.2021.9378657.

Schlüter, M., 1999: Von der $2\frac{1}{2}$D- zur 3D-Flächenmodellierung für photogrammetrische Rekonstruktion im Objektraum. PhD-Thesis, DGK-Reihe C, 506.

Schneider, C.T., 1991: Objektgestützte Mehrbildzuordnung. PhD-Thesis, DGK-Reihe C, 375.

Weisensee, M., 1992: Modelle und Algorithmen für das Facetten-Stereosehen. PhD-Thesis, DGK-Reihe C, 374.

Wrobel, B., 1987: facets stereo vision (FAST vision) – A New Approach to computer Stereo Vision and to Digital Photogrammetry. *Proceedings of a Conference on Fast Processing of Photogrammetric Data*, 231-258.

Wu, B., Nevatia, R., 2009: Detection and Segmentation of Multiple, Partially Occluded Objects by Grouping, Merging, Assigning Part Detection Responses. *International Journal of Computer Vision,* 82, 185–204. https://doi.org/10.1007/s11263-008-0194-9.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021: SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34, 12077-12090.

Yu, R., Russel, C., Campbell, N.D.F., Agapito, L., 2015: Direct, Dense, and Deformable: Template-Based Non-Rigid 3D Reconstruction from RGB Video. *2015 IEEE International Conference on Computer Vision (ICCV),* 918-926.