










# A STRUCTURE FROM MOTION PIPELINE FOR ORTHOGRAPHIC MULTI-VIEW IMAGES

Kai A. Neumann<sup>\*,†</sup>  Philipp P. Hoffmann<sup>\*</sup>  Max von Buelow<sup>\*</sup>   
Volker Knauth<sup>\*</sup>  Tristan Wirth<sup>\*</sup>  Christian Kontermann<sup>‡</sup>   
Arjan Kuijper<sup>†</sup>  Stefan Guthe<sup>\*,†</sup>  Dieter W. Fellner<sup>\*,†,§</sup> 

<sup>\*</sup>Technical University of Darmstadt, Interactive Graphics Systems Group, Germany

<sup>†</sup>Fraunhofer IGD, Germany

<sup>‡</sup>Technical University of Darmstadt, Institute for Materials Technology, Germany

<sup>§</sup>Graz University of Technology, Institute of Computer Graphics and Knowledge Visualization, Austria

## ABSTRACT

Structure from Motion (SfM) plays a crucial role in unstructured capturing. While images are usually taken by perspective cameras, orthographic camera projections do not suffer from the foreshortening effect, that leads to varying capturing quality in image regions. Most contributions to orthographic image SfM assume a perspective setup with nearly infinite focal length. These assumptions lead to potentially sub-optimal camera pose estimation. Therefore, we propose a SfM pipeline that is optimized for orthographically projected images. For this, we estimate initial camera poses using the factorization method by Tomasi and Kanade. These poses are further refined by a specialized bundle adjustment implementation for orthographic projections. The proposed pipeline surpasses the precision of state-of-the-art work by an order of magnitude, while consuming considerably less computational resources.

**Index Terms**— Camera Pose Estimation, Structure from Motion, Orthographic Geometry, Multi-View Geometry

## 1. INTRODUCTION

Modern photogrammetry pipelines use Structure from Motion (SfM) as a main component to facilitate camera pose estimation from multi-view image datasets. These pipelines have a multitude of real-world applications from the generation of digital twins over high precision specimen to usage in virtual reality application for educational or recreational purposes.

Classical SfM algorithms [2, 3] use a perspective image projection because this is the way the real world is represented by most lenses and perceived by human beings. While the pixel density in the image plane is constant, the foreshortening effect of perspective projections leads to a pixel density imbalance for object surfaces that are further away from the camera, i.e. fewer pixels per real-world surface area. This may have serious quality implications for 3D reconstruction pipelines, especially in texture fitting stages. The applica-

tion of orthographic geometry by utilizing telecentric lenses is suited to mitigate these drawbacks due to the absence of the foreshortening effect. Therefore, we propose a SfM algorithm for orthographic geometry that enables photogrammetric image processing pipelines for image data acquired with telecentric lenses.

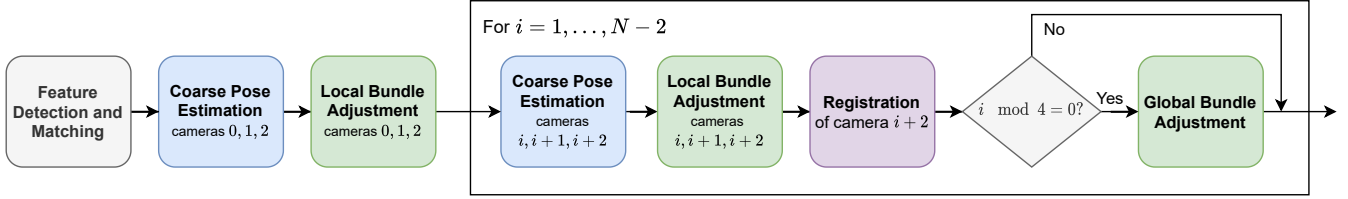
We perform an extensive comparison to the work of Julià et al. [4], which is considered state-of-the-art. This comparison shows that our system provides (a) a better scalability due to its iterative design, (b) a higher overall accuracy and (c) a superior run-time performance.

## 2. RELATED WORK

Structure from Motion (SfM) algorithms estimate camera poses based on multi-view image data, as initially defined by Ullman [5]. First implementations using orthographic projections are presented by Bennett et al. [6], Harris [7] and Koenderink and Van Doorn [8]. Tomasi and Kanade [1] propose an SfM factorization method that use more than two images to combat ambiguity of camera poses.

As simplification, all four aforementioned solutions use orthographic camera models. For this reason, Sturm and Triggs [9] modified the Tomasi and Kanade [1] factorization for the application on perspective images. The first large-scale incremental SfM pipeline was presented by Snavely et al. [10] and applied to web image collections. Following this implementation, improved SfM pipelines like *VisualSfM* [11], *MVE* [12] and *COLMAP* [13] have been introduced. An evaluation of state-of-the-art SfM pipelines is presented by Bianco et al. [14].

A commonly used technique in iterative SfM pipelines called *bundle adjustment* is used to prevent the accumulation of errors. An overview of this method is presented by Triggs et al. [15]. Blonquist and Pack [16] propose a bundle adjustment variant specially optimized for orthographic images. In addition, Oskarsson [17], Larsson et al. [18] and Julià et al. [4] introduce specialized orthographic SfM algorithms.



**Fig. 1:** Proposed SfM Pipeline Architecture. After an initial pose estimation with three cameras, single cameras are added iteratively by estimating their position with respect to two already estimated camera poses using the algorithm by Tomasi and Kanade [1] and a local BA. Every few iterations a global BA is performed to accumulating errors.

### 3. ORTHOGRAPHIC STRUCTURE FROM MOTION

Our approach initially extracts image feature descriptors as described in Section 3.1. On the base of these features we iteratively perform a coarse pose estimation and bundle adjustment. The coarse estimation of camera poses is based on the approach of Tomasi and Kanade [1] and is performed on groups of three images (see Section 3.3). The group selection is described in detail in Section 3.2. These estimates are further refined using local and global BA (see Section 3.4). Fig. 1 illustrates an overview of our proposed pipeline architecture.

#### 3.1. Feature Detection and Matching

We jointly extract SIFT [19] and SURF [20] feature positions  $\mathbf{x}_i$  and their descriptors from each image in the dataset, similar to MVE [12]. Subsequently, we apply a nearest neighbor approach to match image feature correspondences before filtering them using a RANSAC-based approach. The resulting feature positions and matches form the foundation for our coarse pose estimation.

#### 3.2. Group Building

The coarse pose estimation is always carried out on groups of three cameras with each group containing exactly two previously aligned cameras. The groups are built by maximizing the number of features shared between all three cameras of each group.

#### 3.3. Coarse Pose Estimation

For the coarse pose estimation, we use the algorithm proposed by Tomasi and Kanade [1] that calculates the relative rotations between a set of cameras for a set of feature point matches. In addition to that, the algorithm estimates a *shape matrix*  $\mathbf{S}$  consisting of the 3D positions of all feature points. The goal is to solve an equation system resulting from an initial matrix called the *measurement matrix*  $\mathbf{D} = [\mathbf{X}, \mathbf{Y}]^T$ . It contains the image coordinates of the feature matches and has the dimensions  $2f \times p$  with  $f$  frames and  $p$  feature points.

Together with the unknown *rotational matrix*  $\mathbf{R}$  that represents the camera pose transformations, we construct an equation system. In order to retrieve those matrices, we perform a singular value decomposition (SVD) on the registered measurement matrix. Due to noise, we only consider the top three eigenvalues and its eigenvectors which form the basis for the upcoming non-linear optimization. Because this factorization method is sensitive to outliers, it is run inside of a RANSAC loop.

#### 3.4. Bundle Adjustment

Bundle adjustment simultaneously optimizes the extrinsic and intrinsic camera parameters as well as the triangulated points, thereby preventing error accumulation during the iterative registration of cameras [21]. In our pipeline, we perform BA on the estimates given by the Tomasi and Kanade Factorization in a local manner. Due to this reoccurring local BA, we perform the computationally more expensive task of global BA only every four steps without loss of algorithm stability [14].

**Problem Statement** BA minimizes an error metric on the camera parameters and triangulated points. The most commonly used metric is the *reprojection error* [22]:

$$e(\mathbf{p}) = \sum_j \rho(\mathbf{r}_j) = \sum_j \rho(\mathbf{x}_j - \mathbf{f}(\mathbf{y}_j, \mathbf{p})) \quad (1)$$

where  $\mathbf{p}$  denotes the camera parameters,  $\mathbf{x}_j$  the image coordinates of observed features,  $\mathbf{y}_j$  the triangulated 3D point that corresponds to  $\mathbf{x}_j$ .  $\mathbf{f}$  denotes a function that projects a 3D point  $\mathbf{y}_j$  onto the same image plane as  $\mathbf{x}_j$ , while  $\rho$  denotes an arbitrary robust error function.

The optimization of the reprojection error is usually formulated as a non-linear least squares problem [15] and commonly consists of a large number of residuals. Because each residual is only relevant for a small subset of camera parameters, the problem is sparse in nature. We exploit this by using a sparse linear solver as part of the non-linear least squares solver *Ceres*.

**Camera Parameters** We define the camera parameters  $\mathbf{p}_i$  to be optimized as follows. Because view rays of a true orthographic projection are parallel, the projection of an object onto the image plane is invariant to translation in its view direction. For that reason, the position of a camera is less important than its orientation. Therefore, we represent the cameras by placing them on a sphere around the object with a fixed radius, resulting in the extrinsic parameters of a camera  $\mathbf{p}_i := [q_{x,i}, q_{y,i}, q_{z,i}, q_{w,i}, u_i, v_i, s_i]^T$ .

These are defined as the rotation quaternion  $\mathbf{q}_i = (q_{x,i}, q_{y,i}, q_{z,i}, q_{w,i})$ , the offset  $(u_i, v_i)$  relative to the camera’s image plane and a scale factor  $s_i$ . While the quaternion sufficiently defines the camera’s orientation, the offset is necessary to account for the fact that not all cameras look at the same point. Using these seven parameters, a point  $\mathbf{x}_{\text{cam}}$  in camera space is transformed into world coordinates using the transformation

$$\mathbf{T}_i(\mathbf{x}_{\text{cam}}) := \mathbf{q}_i (s_i \mathbf{x}_{\text{cam}} + \mathbf{o}_i) \mathbf{q}_i^{-1} \quad (2)$$

with the camera’s offset vector  $\mathbf{o}_i = [u_i, v_i, -r]^T$  and  $r$  the radius of the sphere.

**Residuals** After triangulating all tracks by using the method proposed by Traa [23], the reprojection error residuals  $\mathbf{r}_j$  can be estimated using the newly calculated point cloud. For a feature  $\mathbf{x}_j$ , the corresponding triangulated point  $\mathbf{y}_j$  in projected back onto the image plane by applying

$$\mathbf{x}_j^{\text{proj}} = \frac{1}{s} (\mathbf{q}_i^{-1} \mathbf{y}_j \mathbf{q}_i - \mathbf{o}_i) \quad (3)$$

and only using the x and y coordinate of  $\mathbf{x}_j^{\text{proj}}$  to calculate

$$\mathbf{r}_j = \rho \left( \mathbf{x}_j - \left[ \mathbf{x}_{j,x}^{\text{proj}}, \mathbf{x}_{j,y}^{\text{proj}} \right]^T \right). \quad (4)$$

The residuals  $\mathbf{r}_j$  are iteratively calculated during each optimization step.

### 3.5. Registration

After running the coarse pose estimation and local bundle adjustment on a group of three cameras, the group needs to be registered into the same (global) coordinate system as all previously reconstructed groups. Apart from the initial three cameras, each group always contains two cameras that have been part of at least one other group. By rotating these two cameras onto their previously registered counterparts, the local poses can be transformed into the global coordinate system. In our implementation this is done by calculating the rotation from a unit quaternion  $\mathbf{q}_i$  to another unit quaternion  $\mathbf{q}_j$  using the equation  $\mathbf{q}_r = \mathbf{q}_i^* \cdot \mathbf{q}_j$ .

## 4. RESULTS

In Section 4.1, we evaluate our proposed pipeline on synthetic datasets especially created for this purpose. Furthermore, we use subsets of these datasets to compare our implementation against the state-of-the-art solution by Julià et al. [4]. Finally, the pipeline was applied to real datasets captured with a telecentric lens in Section 4.2.

### 4.1. Synthetic Datasets



**Fig. 2:** The synthetic datasets used for testing the orthographic SfM pipeline.

We evaluate our algorithm based on three synthetic models: Blender’s mascot *Suzanne* (Fig. 2a), two linked tori (Fig. 2b) and the *XYZ RGB Dragon* (Fig. 2c). The models are procedurally textured with semi-random noise patterns to ensure a large number of unique features to be detected by the feature detector.

We evaluate our approach on three different camera positioning setups: cameras placed on a perfect circle around the object to mimic real world capturing using a turntable (*Circle*), three circles of different heights to test the algorithms generalization capabilities (*Vertical*) and a camera setup with random noise to the rotation direction compared to the *Vertical* setup in order to test our approaches robustness (*Rotated*).

**Evaluation Metric** The deviation of two camera orientations can be evaluated using quaternions [14]. If the representation of the estimated camera orientation  $\mathbf{q}_E$  as well as the corresponding ground truth  $\mathbf{q}_{GT}$  are known, the rotation  $\mathbf{q}_R$  that is necessary to rotate  $\mathbf{q}_E$  onto the ground truth is calculated as  $\mathbf{q}_R = \mathbf{q}_E^* \cdot \mathbf{q}_{GT}$ . By converting  $\mathbf{q}_R$  into an axis-angle representation we can calculate a single angle  $\alpha = 2 |\arccos(\mathbf{q}_{R,w})|$  that represents the difference between estimated orientation and the ground truth. In the following, this is used as a metric to evaluate the accuracy of our computed orientations.

**Comparison** We compare our proposed pipeline to the algorithm of Julià et al. [4], which is to our best knowledge the state-of-the art solution for SfM on (pseudo)-orthographic image data. Their approach is only able to handle three images as input. Therefore, we selected three images from each of the datasets and calculated the feature tracks using our pipeline. These are subsequently used to estimate the camera poses

Model		Suzanne			Linked Tori			Dragon		
Pose Set		Circle	Vertical	Rotated	Circle	Vertical	Rotated	Circle	Vertical	Rotated
Julià et al. [4]	<b>Mean Error</b> $\bar{\alpha}$ [°]	<b>0.025969</b>	<b>4.338321</b>	<b>14.65476</b>	<b>0.013215</b>	<b>4.342366</b>	<b>14.655496</b>	<b>0.010675</b>	<b>4.350638</b>	<b>21.187649</b>
	Standard deviation [°]	0.022264	6.112474	11.951518	0.01083	6.119286	11.95721	0.009367	6.109137	23.838683
	Mean Pose Est. Runtime [s]	335.12658	2.943899	1.524096	168.473194	0.595378	1.826083	107.009562	0.719133	0.381937
Our Algorithm	<b>Mean Error</b> $\bar{\alpha}$ [°]	<b>0.004297</b>	<b>0.008302</b>	<b>0.009632</b>	<b>0.005631</b>	<b>0.005045</b>	<b>0.008495</b>	<b>0.008546</b>	<b>0.042155</b>	<b>0.025316</b>
	Standard deviation [°]	0.003448	0.008441	0.007153	0.006472	0.003637	0.006157	0.007417	0.030072	0.033767
	Mean Pose Est. Runtime [s]	1.672996	0.921427	0.381716	1.593956	0.403308	0.331142	1.337982	0.587157	0.245496
Our Algorithm (all images)	<b>Mean Error</b> $\bar{\alpha}$ [°]	<b>0.006979</b>	<b>0.002698</b>	<b>0.003835</b>	<b>0.003399</b>	<b>0.004</b>	<b>0.004246</b>	<b>0.010588</b>	<b>0.009142</b>	<b>0.016325</b>
	Standard deviation [°]	0.003107	0.001192	0.002225	0.001689	0.00196	0.004032	0.004122	0.004289	0.006373
	Mean Runtime [s]	45.193533	387.337364	292.109408	266.983307	311.944935	377.671413	186.77068	178.321606	189.716136
	Image count	16	48	48	36	48	48	36	48	48

**Table 1:** Orientation error of the estimated poses on synthetic datasets with three images as well as the full synthetic datasets.

for both algorithms in order to ensure a fair comparison of both approaches. The system specifications used in our experiments are an *Intel Core i5-9600K* processor and *2x 8GB Crucial CT8G4DFS8266 DDR4* memory.

The experimental results (see Tab. 1) show that our algorithm outperforms the state-of-the-art solution on all datasets and camera setups, usually by several orders of magnitude. However, our results also imply that the algorithm of Julià et al. [4] only performs as expected on the *circle* dataset. A comparably low execution time and high mean error on the other datasets indicate that their approach terminates early on the multi-circle datasets. But even a restriction to the *circle* camera setup leads to an outperformance of one order of magnitude regarding the mean angle error on the datasets (*Suzanne* and *Linked Tori*), while maintaining a significantly lower execution time. Furthermore, our experiments suggest that our pipeline scales well to larger datasets (see lower half of Tab. 1).

#### 4.2. Real Datasets



(a) Valdivia Figurine (b) Elephant Statuette

**Fig. 3:** The captured datasets used for testing the orthographic SfM pipeline.

We further evaluate the proposed system on two real datasets. The datasets are captured with a *0.28X CobaltTL* telecentric lens every ten degrees using a *LSDH-100WS* turntable with a resolution of  $4.5''$  and a repetition accuracy of less than  $18''$  [24]. This allows us to evaluate the results based on the aforementioned ten degrees as ground truth. As test objects we use a small *Valdivia Figurine* made of stone (Fig. 3a) and a *Wooden Elephant Statuette* (Fig. 3b).

Our experiments result in a mean angular error of  $0.984^\circ$  (std  $0.463^\circ$ ) on the *valdivia* dataset and a mean angular error of  $0.234^\circ$  (std  $1.397^\circ$ ) for the *elephant*.

In comparison to the synthetic datasets, this error is significantly higher. While this could signal that our pipeline is less robust on real data, it is also likely that the actual poses during acquisition do not match the assumed ground truth due to measurement irregularities like a slight roll of the camera. For this reason, we omitted the comparison to Julià et al. [4], as no accurate comparison would have been possible.

## 5. CONCLUSION

In our work we presented a novel SfM pipeline that uses the factorization method by Tomasi and Kanade and a specialized bundle adjustment implementation to enhance the accuracy of SfM on orthographic image data. This method enables the deployment of systems that are able to operate in unstructured capturing environments but retain a high degree of precision while circumventing the foreshortening effect.

Our results show that the proposed specialized pipeline has a mean angular error that is at least one order of magnitude lower than the state-of-the-art while maintaining a superior run-time performance and robustness. On real-world datasets with circular positioned orthographic cameras, we measured errors between  $0.234^\circ$  and  $0.984^\circ$  when optimizing the angle of our turntable setup.

**Source Code** The source code and datasets for this paper are available at <https://github.com/kai-neumann/OrthoSfM>.

## Acknowledgements

Part of the research in this paper was funded by DFG (Deutsche Forschungsgemeinschaft) project 407 714 161 and the AVIF project “Robust Fracture Deformation Parameters”. We thank the anonymous reviewers whose comments helped improve this manuscript.

## References

- [1] C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” in *International Journal of Computer Vision*, vol. 9, 1992, p. 137–154, DOI: 10.1007/BF00129684.
- [2] R. I. Hartley, “Euclidean reconstruction from uncalibrated views,” in *Joint European-US workshop on applications of invariance in computer vision*. Springer, 1993, pp. 235–256, DOI: 10.1007/3-540-58240-1\_13.
- [3] M. Pollefeys, R. Koch, and L. Van Gool, “Self-calibration and metric reconstruction inspite of varying and unknown intrinsic camera parameters,” *International Journal of Computer Vision*, vol. 32, no. 1, pp. 7–25, 1999, DOI: 10.1023/A:1008109111715.
- [4] L. F. Julià, P. Monasse, and M. Pierrot-Deseilligny, “The Orthographic Projection Model for Pose Calibration of Long Focal Images,” *Image Processing On Line*, vol. 9, pp. 231–250, 2019, DOI: 10.5201/ipol.2019.248.
- [5] S. Ullman, “The interpretation of structure from motion,” *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979, DOI: 10.1098/rspb.1979.0006.
- [6] B. Bennett, D. Hoffman, J. Nicola, and C. Prakash, “Structure from two orthographic views of rigid motion,” *JOSA A*, vol. 6, no. 7, pp. 1052–1069, 1989, DOI: 10.1364/JOSAA.6.001052.
- [7] C. Harris, “Structure-from-motion under orthographic projection,” in *European Conference on Computer Vision*. Springer, 1990, pp. 118–123, DOI: 10/d8px4k.
- [8] J. J. Koenderink and A. J. Van Doorn, “Affine structure from motion,” *JOSA A*, vol. 8, no. 2, pp. 377–385, 1991, DOI: 10.1364/JOSAA.8.000377.
- [9] P. Sturm and B. Triggs, “A factorization based algorithm for multi-image projective structure and motion,” in *European conference on computer vision*. Springer, 1996, pp. 709–720, DOI: 10.1007/3-540-61123-1\_183.
- [10] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: exploring photo collections in 3d,” in *ACM siggraph 2006 papers*, 2006, pp. 835–846, DOI: 10.1145/1179352.1141964.
- [11] C. Wu, “Towards linear-time incremental structure from motion,” in *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 2013, pp. 127–134, DOI: 10.1109/3DV.2013.25.
- [12] S. Fuhrmann, F. Langguth, and M. Goesele, “MVE - A Multi-View Reconstruction Environment,” in *Eurographics Workshop on Graphics and Cultural Heritage*, R. Klein and P. Santos, Eds. The Eurographics Association, 2014, DOI: 10.2312/gch.20141299.
- [13] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113, DOI: 10.1109/CVPR.2016.445.
- [14] S. Bianco, G. Ciocca, and D. Marelli, “Evaluating the performance of structure from motion pipelines,” *Journal of Imaging*, vol. 4, no. 8, p. 98, 2018, DOI: 10.3390/jimaging4080098.
- [15] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, “Bundle adjustment—a modern synthesis,” in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372, DOI: 10/czhk74.
- [16] K. F. Blonquist and R. T. Pack, “A bundle adjustment approach with inner constraints for the scaled orthographic projection,” *ISPRS journal of photogrammetry and remote sensing*, vol. 66, no. 6, pp. 919–926, 2011, DOI: 10.1016/j.isprsjprs.2011.07.001.
- [17] M. Oskarsson, “Two-view orthographic epipolar geometry: Minimal and optimal solvers,” *Journal of Mathematical Imaging and Vision*, vol. 60, no. 2, pp. 163–173, 2018, DOI: 10.1007/s10851-017-0753-1.
- [18] V. Larsson, M. Pollefeys, and M. Oskarsson, “Orthographic-perspective epipolar geometry,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5570–5578.
- [19] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004, DOI: 10/bqrmsp.
- [20] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, “Speeded-up robust features (surf),” *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008, DOI: 10.1016/j.cviu.2007.09.014.
- [21] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, “A survey of structure from motion,” *Acta Numerica*, vol. 26, p. 305–364, 2017, DOI: 10/gbgx2g.
- [22] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz, “Multicore bundle adjustment,” in *CVPR 2011*, 2011, pp. 3057–3064, DOI: 10.1109/CVPR.2011.5995552.
- [23] J. Traa, “Least-squares intersection of lines,” 2013.
- [24] V. Knauthe, M. von Buelow, S. Guthe, M. Adam, C. Kontermann, and M. Oechsner, “High precision orthographic specimen scanning system for the evaluation of creep rupture parameters,” in *44. Vortragsveranstaltung zum Langzeitverhalten warmfester Stähle und Hochtemperaturwerkstoffe*. FVWHT, Nov. 2021.