WM-SBA: Weighted Multibody Sparse Bundle Adjustment

Kai Cordes, Mark Hockner, Hanno Ackermann, Bodo Rosenhahn, and Jörn Ostermann
Institut für Informationsverarbeitung (TNT), 30167 Hannover, Germany
{cordes, hockner, ackermann, rosenhahn, ostermann}@tnt.uni-hannover.de

Abstract

Sparse bundle adjustment (SBA) is the state of the art method for simultaneously optimizing a set of camera poses and 3D points. The multibody bundle adjustment optimizes the static scene and the moving rigid object(s). The result is one camera path representing the main camera motion and virtual camera path(s) for each of the independently moving objects in the scene. The bundle adjustment for the multibody problem is performed in a joint optimization. Main motion (static scene) and object motion(s) are included in the optimization such that the sparse algorithm of SBA can still be applied, even when enforcing the constraint that main camera and object camera share the same intrinsic parameters. The joint optimization approach enables weighting the resulting error for each of the motion models and therefore influences the optimization process. Our experiments with synthetic and natural image data show that an appropriate weighting leads to more accurate camera parameters.

1 Introduction

Structure and motion (*SAM*) recovery consists of feature detection, correspondence analysis, outlier elimination, and bundle adjustment. The idea of bundle adjustment is to minimize the distance between the reprojection of an estimated 3D object point and the measured feature point for each camera, in which the 3D point is visible. Bundle adjustment uses a statistical error model which is equivalent to a maximum likelihood estimator and simultaneously estimates the camera parameters and the 3D positions of feature points [7, 10]. Sparse bundle adjustment (SBA) is the standard optimization method for this problem and several extensions have been proposed to improve the performance [5, 7] or the applicability [1, 2, 6, 9].

In [4], a multibody SAM recovery approach is introduced. One motion model represents the static scene and another one represents a rigidly moving object. By using the constraint that the intrinsic camera parameters are identical for both motion models, the estimation of the focal length is improved.

We introduce a formulation of the multibody bundle adjustment which enables the weighting of different motion models. Additionally, it allows for a sparse optimization of the resulting multibody Jacobian matrix. It is shown that the reconstruction improves by the weighting, if the motion models have a different representation quality, i.e. a different amount of noise. For many applications, e.g. the three-dimensional reconstruction of scene and moving objects from a driving car, this assumption is well justified [8]. For the evaluation, synthetic as well as natural data is used.

In the following Sect. 2, the reference bundle adjustment and the new approach for the joint optimization of the multibody problem are explained. In Sect. 3, the experimental conditions are defined. Sect. 4 shows the results on synthetic data while Sect. 5 demonstrates the application with natural images. In Sect. 6, the paper is concluded.

2 Multibody Bundle Adjustment

In [4], it is shown that the estimation of the variant focal length improves by enforcing the constraint that both cameras (the one which observes the static geometry and the virtual camera for the moving object) share the same focal length.

We extend the sparse bundle adjustment (SBA) provided by the authors of [7] to the new approach of *weighted multibody sparse bundle adjustment* (WM-SBA). In contrast to [4], our formulation uses the sparse structure of the Jacobian [7] for optimizing the static scene and moving objects. Additionally, we introduce a weighting scheme for the reprojection errors. As we extended the SBA package [7], we adopt their notation.

2.1 Bundle Adjustment

The bundle adjustment aims at minimizing the following term [7]:

$$\min_{\mathbf{a_j},\mathbf{b_i}} \sum_{i=1}^{n} \sum_{j=1}^{m} ||Q(\mathbf{a_j},\mathbf{b_i}) - \mathbf{x_{ij}}||^2 \tag{1}$$

Here, the camera parameters are represented by the vectors $\mathbf{a_j}$ and the 3D object points are represented by the vectors $\mathbf{b_i}$. The observation values are the feature points $\mathbf{x_{ij}}$ in the images. The function $Q(\mathbf{a},\mathbf{b})$ is determined by the projection of the 3D object point $\mathbf{b}$ onto the image plane of the camera represented by its camera parameters $\mathbf{a}$. The extrinsic parameters consist of translation $\mathbf{T}$ and rotation $\mathbf{R}$, which are used to map the 3D object point $\mathbf{b_i}$ from the camera coordinate system to the world coordinate system:

The projection is determined by the intrinsic parameters of the camera. In [7], five intrinsic parameters are assumed (focal length f, principal point (u_0, v_0), aspect ratio α, and skew s). The intrinsic parameters build the calibration matrix K,

$$\mathbf{K} = \begin{pmatrix} f & s & u_0 \\ 0 & \alpha f & v_0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

Additionally, up to 5 parameters are allowed for the radial distortion coefficients. All intrinsic parameters can be determined by a calibration (if they are fixed throughout the image sequence) or are optimized by the bundle adjustment. The projection is visualized in Fig. 1 with the green cameras and object points. The

reference method [7] builds the Jacobian Matrix J for the optimization of (1). Here is a small example for $m = 2$ images and $n = 3$ 3D points:

$$J = \begin{pmatrix} A_{11} & 0 & B_{11} & 0 & 0 \\ 0 & A_{12} & B_{12} & 0 & 0 \\ A_{21} & 0 & 0 & B_{21} & 0 \\ 0 & A_{22} & 0 & B_{22} & 0 \\ A_{31} & 0 & 0 & 0 & B_{31} \\ 0 & A_{32} & 0 & 0 & B_{32} \end{pmatrix} \quad (3)$$

To reduce the computation time, the Levenberg Marquardt based optimization exploits the sparse structure of the Jacobian matrix [7].

2.2 WM-SBA: Weighted Multibody SBA

Our formulation of the multibody bundle adjustment preserves the sparsity of the Jacobian matrix and enables the joint optimization by describing the 3D object points of the moving object OBJ relative to the MAIN camera coordinate system. The extrinsic parameters of the static scene geometry are given as $\mathbf{R}^{(0)}$ and $\mathbf{T}^{(0)}$ (with additional $^{(0)}$ to indicate motion model 0). The extrinsic parameters of the moving object, $\mathbf{R}^{(1)}$ and $\mathbf{T}^{(1)}$ are described relative to motion model 0. A 3D object point $\mathbf{b_i}'^{(1)}$is rotated and translated such that it can be projected into the camera image with the camera parameters of motion model 0:

$$\mathbf{b_i}'^{(1)} = \mathbf{R}^{(0)} \cdot \left(\mathbf{R}^{(1)} \cdot (\mathbf{b_i}^{(1)} - \mathbf{T}^{(0)} - \mathbf{T}^{(1)}) \right) \quad (4)$$

Note, that $\mathbf{R}^{(1)}$ and $\mathbf{T}^{(1)}$ hold the same number of parameters as in the reference optimization approach. The second motion model with parameters $\mathbf{a_j}^{(1)}$ and observation values $\mathbf{b_i}^{(1)}$ is visualized with orange color in Fig. 1. Now, the Jacobian matrix J' is formulated by alternating the object points of the different motion models on the right hand side. This leads to 2×3 sized sub matrices $B_{ij}^{(00)}$ (MAIN) and $B_{ij}^{(11)}$ (OBJ):

$$J' = \begin{pmatrix} A_{11}^{(00)} & 0 & 0 & 0 & B_{11}^{(00)} & 0 & 0 & 0 & 0 & 0 \\ A_{11}^{(10)} & A_{11}^{(11)} & 0 & 0 & 0 & B_{11}^{(11)} & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{12}^{(00)} & & B_{12}^{(00)} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & A_{12}^{(10)} & A_{12}^{(11)} & 0 & B_{12}^{(11)} & 0 & 0 & 0 & 0 \\ A_{21}^{(00)} & 0 & 0 & 0 & 0 & 0 & B_{21}^{(00)} & 0 & 0 & 0 \\ A_{21}^{(10)} & A_{21}^{(11)} & 0 & 0 & 0 & 0 & 0 & B_{21}^{(11)} & 0 & 0 \\ 0 & 0 & A_{22}^{(00)} & 0 & 0 & 0 & B_{22}^{(00)} & 0 & 0 & 0 \\ 0 & 0 & A_{22}^{(10)} & A_{22}^{(11)} & 0 & 0 & 0 & B_{22}^{(11)} & 0 & 0 \\ A_{31}^{(00)} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{31}^{(00)} & 0 \\ A_{31}^{(10)} & A_{31}^{(11)} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & B_{31}^{(11)} \\ 0 & 0 & A_{32}^{(00)} & 0 & 0 & 0 & 0 & 0 & B_{32}^{(00)} & 0 \\ 0 & 0 & A_{32}^{(10)} & A_{32}^{(11)} & 0 & 0 & 0 & 0 & 0 & B_{32}^{(11)} \end{pmatrix} \quad (5)$$

The submatrices $A_{ij}^{(00)}$ have the same entries as the matrices A_{ij} for motion model 0 (cf. eq. (3)). The submatrices $A_{ij}^{(11)}$ now consist of the relative parameters of the moving object with respect to the MAIN motion model. It follows, that the submatrices $A_{ij}^{(10)}$ depend on the MAIN motion camera parameters like $A_{ij}^{(00)}$.

By subsuming each block $A'_{ij} = \begin{pmatrix} A_{ij}^{(00)} & 0 \\ A_{ij}^{(10)} & A_{ij}^{(11)} \end{pmatrix}$ and each block $B'_{ij} = \begin{pmatrix} B_{ij}^{(00)} & 0 \\ 0 & B_{ij}^{(11)} \end{pmatrix}$, the global structure of the matrix with its zeros is preserved (cf. eq. (3)). Hence, the technique of sparse optimization [7] can still be applied. Only the sizes of the matrices and their entries changed. If five intrinsic parameters are optimized additionally to the six extrinsic parameters, the WM-SBA builds 4×17 sized submatrices instead of 2×11 sized matrices for each motion model. The resulting reprojection error of the WM-SBA can be separated into the two parts:

$$c^{(0)} = \sum_{i=1}^{n} \sum_{j=1}^{m} ||Q(\mathbf{a_j}^{(0)}, \mathbf{b_i}^{(0)}) - \mathbf{x_{ij}}^{(0)}||^2, \quad (6)$$

$$c^{(1)} = \sum_{i=1}^{n'} \sum_{j=1}^{m'} ||Q'(\mathbf{a_j}^{(0)}, \mathbf{a_j}^{(1)}, \mathbf{b_i}^{(0)}, \mathbf{b_i}^{(1)}) - \mathbf{x_{ij}}^{(1)}||^2$$

A constant λ weights the costs of motion models:

$$\min_{\mathbf{a_j}^{(0)}, \mathbf{b_i}^{(0)}, \mathbf{a_j}^{(1)}, \mathbf{b_i}^{(1)}} c^{(0)} + \lambda \cdot c^{(1)} \quad (7)$$

Practically, the weighting is implemented by the additional λ multiplied to the covariance matrices of the 2D points for the corresponding motion model. Due to the joint optimization approach, the weighting influences the optimization and leads to a different minimum. The method can easily be extended to an arbitrary number of motion models.

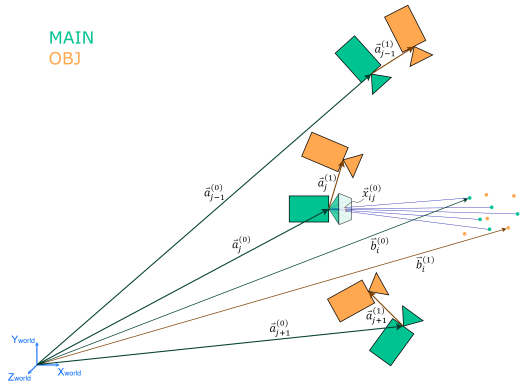


Figure 1: Projection of 3D points into the cameras. The green cameras and points belong to the real camera motion (MAIN), the orange ones belong to the 3D points of the moving object and the (virtual) camera path (OBJ).

3 Experiments

In our experiments, two motion models are constructed, a main motion model (MAIN) and an object motion model (OBJ). Both motion models observe

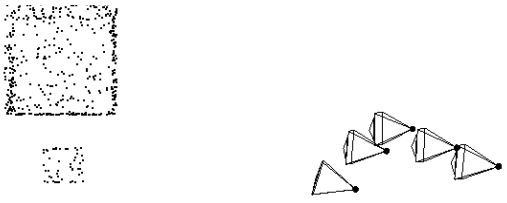
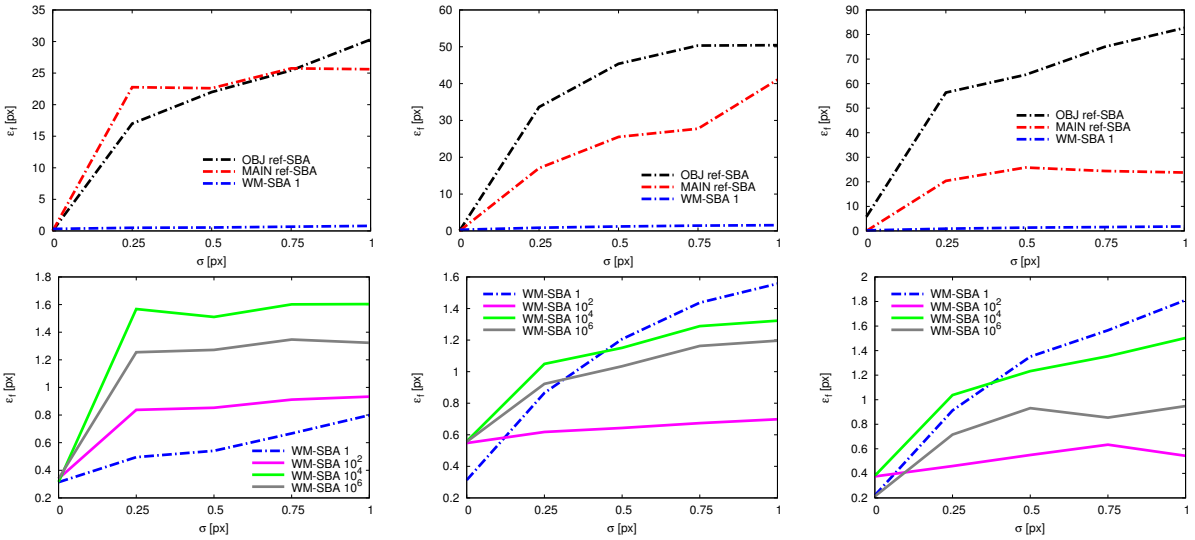


Figure 2: Dataset *Syn3* with a smaller OBJ volume compared to MAIN; both camera paths start at the same position.



(a) *Syn1*: 250-250 pts. (b) *Syn2*: 250-250 pts. (c) *Syn3*: 500-50 pts.

Figure 3: Results for the synthetic data. We show the mean errors ε_f for 500 experiments of the focal length, assumed to be unknown and varying (ground truth [px]: 1406). In *Syn1* both motion models have the same representation. In *Syn2* and *Syn3*, the OBJ points have more noise due to the smaller size of the cube.

points on cubes. For the evaluation, noise is added to the initial camera parameters ($\sigma_T^2 = 0.1$, $\sigma_R^2 = 10^{-5}$), the 3D points ($\sigma_P^2 = 1$), and the 2D points. To account for the usually larger uncertainty in the direction of the optical axis [11], the noise for the 3D points in this direction is amplified by a factor of 5. Throughout one experiment, the noise on the camera parameters and the 3D points remains constant while the noise on the 2D points is varied. We construct a data set *Syn1* with equal amount of noise for MAIN and OBJ (identical size of cubes, equal number of points), and data sets *Syn2* and *Syn3* with more noise on the OBJ model, constructed by a smaller OBJ cube. In *Syn3*, the cubes have different numbers of points (500 and 50 instead of 250 both). *Syn3* is visualized in Fig. 2.

The intrinsic parameters are initialized with their ground truth values. To obtain stable results, 500 runs with the same ground truth are constructed for each 2D point noise level. All evaluated methods get exactly the same initial input positions of cameras, 3D points, and 2D points. The compared methods are ref-SBA and WM-SBA with different values for the weighting coefficient λ, $\lambda \in \{1, 10^2, 10^4, 10^6\}$. The results for WM-SBA with the weighting coefficient 1 should match the approach proposed in [4]. We choose 3 cameras for the experiments because SBA applications start with an initial reconstruction of 2 or 3 cameras no matter wether the sequential [9] or the hierarchical [3] scheme is used. We show the results for the focal lengths in Sect. 4. The results for the principle points are similar since its optimization is closely related to the estimation of the focal length. The computational time using WM-SBA instead of ref-SBA scales by a factor of 3.

4 Results

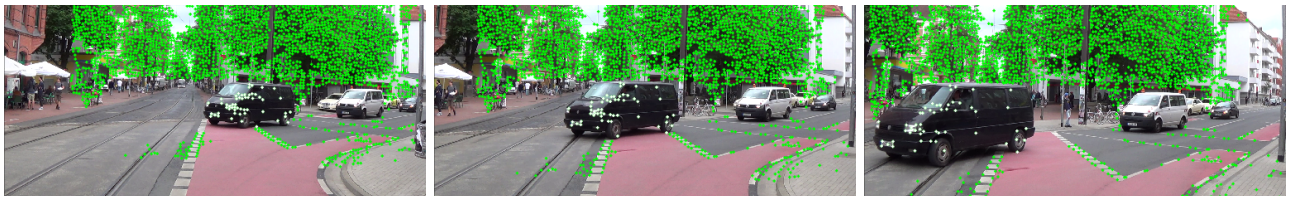
Due to the constrained intrinsic parameters, the joint optimization show a large gain compared to the separated optimization (top row of Fig. 3). For *Syn2* and *Syn3*, the OBJ provides less accuracy than MAIN as it was expected. For the smaller set of points for OBJ (*Syn3*), the error increases further. The joint optimization WM-SBA provides much better results as shown for $\lambda = 1$. The bottom row of Fig. 3 shows a comparison of different weightings λ. For *Syn1*, the weighting $\lambda = 1$ leads to the best results. For a noisier OBJ model compared to MAIN, the weighting with $\lambda = 10^2$ performs best (*Syn2*). For the additional uneven representation with different numbers of points (*Syn3*), the results are comparable to *Syn2* and $\lambda = 10^2$ is still the best choice. For both sets, *Syn2* and *Syn3*, $\lambda = 1$ leads to the largest errors. The optimal λ does not depend significantly on the unevenly distributed numbers of points, but on their spatial distribution. The lower depth of field of OBJ induces an amplification of the noise on the 3D points compared to the better distributed points of MAIN.

5 Application

In Fig. 4, examples for the application scenario are presented. The goal is the accurate multibody reconstruction of all cameras and 3D object points. The feature points are selected with the Harris corner detector and tracked with the KLT tracker. The resulting feature tracks are manually grouped into static scene (MAIN model) and moving object (OBJ model). The evaluated methods optimize extrinsic and intrinsic (variant focal length and principal point) parameters, like in Sect. 3. Initial parameters are estimated using standard SBA and manually set, constant intrinsic parameters. The results are shown in the Fig. 5.

In both sequences, the movement of the moving object OBJ is not correctly captured by the ref-SBA approach. The movement towards the observer is interpreted as an increase in focal length f. Instead, it should be explained by a movement of the virtual OBJ camera towards the observer with constant f. The correct interpretation is achieved by the proposed constrained WM-SBA which assumes identical intrin-

(a) *Nat1* (8 frames) with 2206 trajectories on the static scene and 72 trajectories on the moving object.



(b) *Nat2* (21 frames) with 2652 trajectories on the static scene and 52 trajectories on the moving object (black car).

Figure 4: Natural image sequences: the green points determine the MAIN motion model, the white points on the bus determine the moving object OBJ. In *Nat1*, only the black car in considered as moving object.

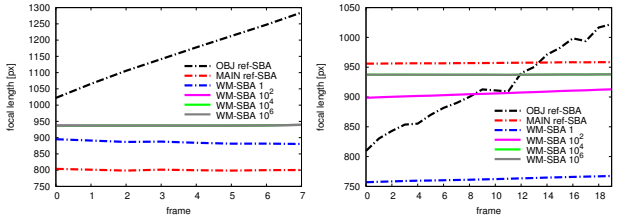


Figure 5: Resulting focal lengths. In *Nat1* (left), WM-SBA with $\lambda \in \{10^2, 10^4, 10^6\}$ lead to the same result. In *Nat2* (right), $\lambda \in \{10^4, 10^6\}$ give the same result.

sic parameters for MAIN and OBJ for each frame.

Some selections of λ lead to the same result (cf. Fig. 5). Together with the result in Sect. 3, we can infer that the points of OBJ provide valuable scene information in the joint optimization of the WM-SBA. Although ground truth values for the natural sequences are not available, the computed focal lengths for WM-SBA with $\lambda = 1$ in Fig. 5, left, appear to be too small. The results of WM-SBA with $\lambda \in \{10^2, 10^4\}$ provide reasonable results for both sequences.

6 Conclusions

We propose a new multibody sparse bundle adjustment approach. It allows for joint optimization of static scene and moving objects. It is possible to weight costs arising from different motion models. It is shown with synthetic and natural data that appropriate weightings lead to more accurate camera parameters compared to the common multibody bundle adjustment. It turns out that the optimal weighting depends on the spatial distribution of the observed points in the scene.

The presented application reconstructs the static scene and moving objects in traffic situations observed from a moving car. Although the moving object in the video is represented with very few 3D points, the proposed approach results in a reliable estimation of the intrinsic camera parameters while the reference fails. WM-SBA gives reliable results using the proposed weighting.

Future works shall investigate how to correctly choose the λ for optimal results. The spatial distribution will be exploited to infer suitable values of λ.

References

[1] K. Cordes. *Occlusion Handling in Scene Reconstruction from Video*, volume 10. VDI Verlag, 2014.

[2] K. Cordes, B. Scheuermann, B. Rosenhahn, and J. Ostermann. Learning object appearance from occlusions using structure and motion recovery. In *Asian Conference on Computer Vision*, volume 7726 of *LNCS*, pages 611–623. Springer, 2012.

[3] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *European Conference on Computer Vision*, volume 1406 of *LNCS*, pages 311–326. Springer, 1998.

[4] A. W. Fitzgibbon and A. Zisserman. Multibody structure and motion: 3-D reconstruction of independently moving objects. In *European Conference on Computer Vision*, pages 891–906. Springer, 2000.

[5] K. Konolige. Sparse sparse bundle adjustment. In *British Machine Vision Conference*, 2010.

[6] C. Kurz, T. Thormählen, and H.-P. Seidel. Bundle adjustment for stereoscopic 3D. In *Computer Vision / Computer Graphics Collaboration Techniques*, volume 6930 of *LNCS*, pages 1–12. Springer, 2011.

[7] M. A. Lourakis and A. Argyros. SBA: A Software Package for Generic Sparse Bundle Adjustment. *ACM Trans. Math. Software*, 36(1):1–30, 2009.

[8] K. Ozden, K. Schindler, and L. Van Gool. Simultaneous segmentation and 3d reconstruction of monocular image sequences. In *IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[9] M. Pollefeys, L. V. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004.

[10] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon. Bundle adjustment - a modern synthesis. In *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, ICCV '99, pages 298–372. Springer, 2000.

[11] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–152, 1989.