BodySLAM: Joint Camera Localisation, Mapping, and Human Motion Tracking

Dorian F. Henning[1], Tristan Laidlow[1], and Stefan Leutenegger[1,2]

[1] Imperial College London, UK
{d.henning,t.laidlow15}@imperial.ac.uk
[2] Technische Universität München, Germany
stefan.leutenegger@tum.de

Appendix A Dataset Comparison

To fully evaluate BodySLAM, we required video sequences of human motion captured by a moving camera where camera intrinsics, ground truth camera trajectories and ground truth human body parameters were available. As shown in Table 1, none of the existing human motion datasets include all of these required elements. For this reason, we captured our own dataset of video sequences where the camera and human subject are both in motion, and used a motion capture system to obtain ground truth values.

	Supplied Information								
	RGB	stereo	multi-view	depth	video	dynamic camera	2D pose gt	3D pose gt	not synthetic	calibration	SMPL parameters (from MoSh)
Human3.6m [5]	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✗
MPI-INF-3DHP [10]	✓	✗	✓	✗	✓	✗	✓	✓	✓	✓	✓
3DPW [9]	✓	✗	✗	✗	✓	✓	✓	✓	✓	✓	✓
AMASS [8]	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✓
Berkeley MHAD [13]	✓	✗	✓	✓	✓	✗	✓	✓	✓	✓	✗
MPII HPE [1]	✓	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗
Inria Stereo [2]	✓	✓	✗	✓	✓	✓	✓	✗	✓	✗	✗
JTA [3]	✓	✗	✗	✗	✓	✗	✓	✓	✗	✓	✗
Mannequin [6]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
SHPED [7]	✓	✓	✗	✗	✓	✓	✓	✗	✓	✗	✗
KTP [11,12]	✓	✗	✗	✓	✓	✓	✗	✗	✓	✓	✗
BinoPerfCap [15]	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗
TartanAir [14]	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
KITTI [4]	✓	✓	✓	✓	✓	✓	✗	✗	✓	✓	✗

Table 1: Dataset overview

Appendix B Camera and Human Trajectory Errors

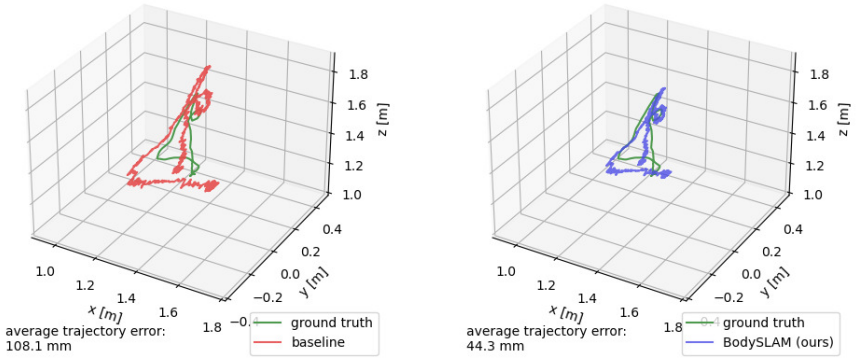
In general, there are many possible ways of aligning the estimated camera and human centre trajectories with ground truth to compute the (average) trajectory errors, but we consider the following four as the most sensible methods:

1. align the camera and human centre trajectories independently in $SE(3)$,
2. align *only* the camera trajectory in $SE(3)$ and compute the human centre trajectory based on this alignment,
3. jointly align both human and camera trajectories in $SE(3)$, or
4. align only the first frame of the camera poses, such that the initial condition is set to ground truth, and compute both the human and camera trajectories based off this measurement.

In our paper, we use the first method to compute the errors, as this method allows for the fairest comparison with methods that focus only on either the camera trajectory error or the human centre trajectory error. We use the fourth method for visualisation in the supplementary video where both trajectories must be shown simultaneously, as this method aligns the first estimates and shows the accumulating error over time. Using other methods might confuse the viewer as the initial poses would not be aligned.

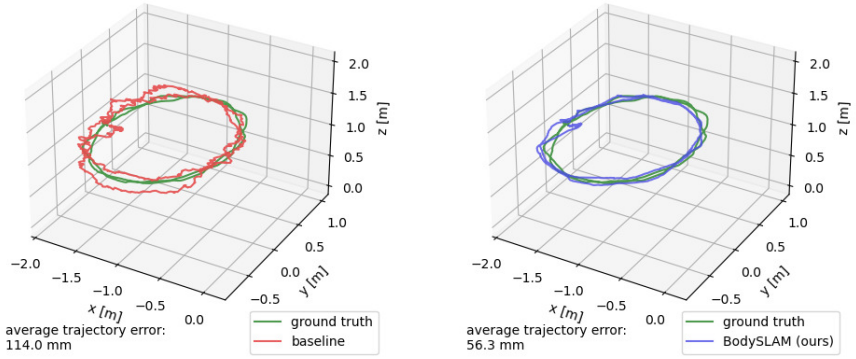
In the following sections, we provide examples of the camera and human centre trajectories estimated by BodySLAM and the baseline approach. As discussed in the main paper, the baseline method optimises the camera trajectory via classic monocular bundle adjustment using just the camera poses and the 3D landmarks. The human centre trajectory estimation is done using only the bundle adjusted camera poses and unary OpenPose measurements with no motion model. Section B.1 shows the trajectory estimates when aligning the camera and human centre trajectories independently in $SE(3)$ (method 1), and Section B.2 shows the trajectory estimates when aligning only the camera trajectory in $SE(3)$ (method 2).

B.1 Trajectories after Independent $SE(3)$ Alignment



(a) Baseline **(b)** BodySLAM

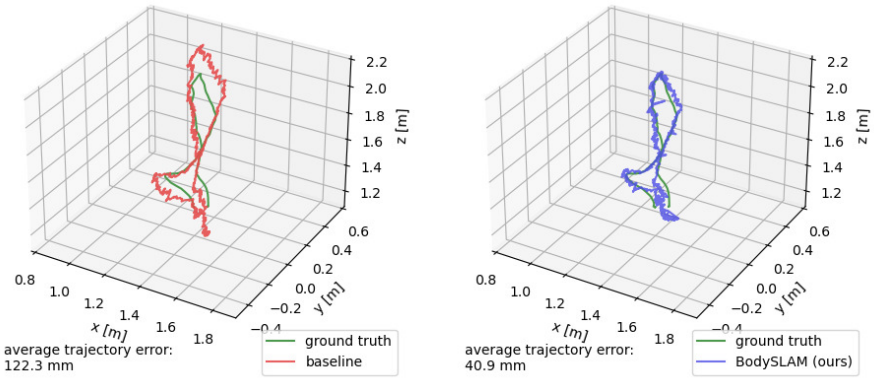
Fig. 1: Estimated camera trajectories for Sequence E2 after $SE(3)$ alignment with the ground truth camera trajectory. The baseline method optimises the camera trajectory via monocular bundle adjustment using just the camera poses and 3D landmarks. By using a human motion model to temporally constrain the optimisation problem, BodySLAM is able to accurately estimate the metric scale of the camera trajectory.



(a) Baseline **(b)** BodySLAM

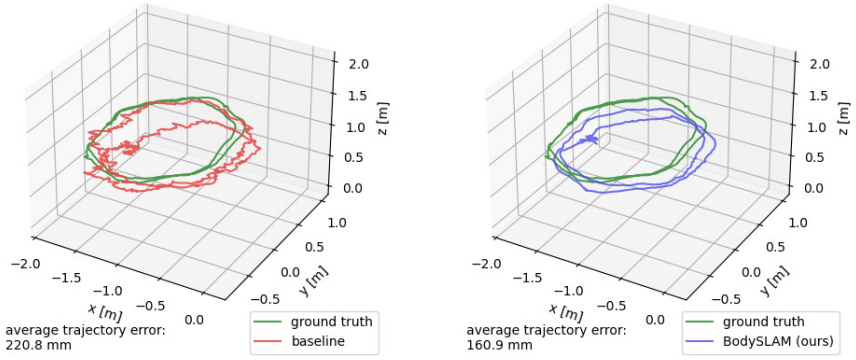
Fig. 2: Human centre trajectories for Sequence E2 after $SE(3)$ alignment with the ground truth human centre trajectory. The baseline method uses the bundle adjusted camera poses and unary OpenPose measurements to estimate human motion. In BodySLAM, the human centre poses and posture parameters are constrained by a motion model leading to smoother and more accurate trajectories at metric scale.

B.2 Trajectories after $SE(3)$ Alignment of Camera Poses



(a) Baseline **(b)** BodySLAM

Fig. 3: Estimated camera trajectories for Sequence E4 after $SE(3)$ alignment with the ground truth camera trajectory. The baseline method optimises the camera trajectory via monocular bundle adjustment using just the camera poses and 3D landmarks. By using a human motion model to temporally constrain the optimisation problem, BodySLAM is able to accurately estimate the metric scale of the camera trajectory.



(a) Baseline **(b)** BodySLAM

Fig. 4: Human centre trajectories for Sequence E4 after an $SE(3)$ alignment of the estimate and ground truth *camera* trajectories. The baseline method uses the bundle adjusted camera poses and unary OpenPose measurements to estimate human motion. In BodySLAM, the human centre poses and posture parameters are constrained by a motion model leading to smoother and more accurate trajectories at metric scale.

Appendix C Joint Optimisation and Structural Landmarks

To evaluate the influence and necessity of having a joint camera and human trajectory optimisation, we performed an ablation experiment to assess the influence of the structural landmarks to the overall performance. In the 2-step pipeline that we used throughout our paper, where we first estimated an initial trajectory of the camera, and then added the human state to the factor graph for the joint optimisation, we introduced an additional perturbation of the camera trajectory after the first step. This caused the average camera trajectory error to increase as seen in Table 2 after perturbation. When optimising the factor graph without structural landmarks, the mean error decreased, relying only on the human keypoint measurements and the motion model error term. However, we also show that including the structural landmarks in the second step of the optimisation, and therefore running a joint optimisation of all elements of the factor graph, improved the estimation accuracy of the camera trajectory.

We did not include the human and joint trajectory errors, since any change here in performance does only depend on the increased accuracy of the camera trajectory, rendering those results redundant.

Table 2: Optimisation of the camera trajectory with and without structural landmarks after perturbation with ± 100 mm and ± 0.01 rad.

	C-ATE [mm] SE(3)			
Seq.	original	perturbed	optimised no landmarks	optimised landmarks
E 2	193.0	213.1	139.69	**138.52**
E 3	122.3	136	78.03	**75.56**
M 4	203.1	225.2	143.03	**139.4**
D 1	310.4	340.7	223.08	**219.1**
D 2	316.0	349.1	209.48	**205.28**
mean	228.96	252.82	158.662	**155.572**

Appendix D Scale Estimation after Perturbation

In this ablation study, we analysed the ability of the motion model to recover the correct scale of the human and camera motion after perturbation with a scale factor. The camera trajectory was perturbed after the initial estimation by a scale factor ranging values from 0.1 to 5.0. The joint optimisation of the full factor graph was then performed, and the final optimised trajectory was aligned to ground truth in Sim(3). The estimated scale s should equal the inverse of the perturbation scale s', and additionally the deviation is reported in Table 3.

Table 3: Scale estimation after perturbation. Values closer to the perturbation are better. The deviation is the estimation error relative to the perturbation. This Table contains the numerical results from Figure 4 in the paper.

perturbation $1/s'$ **[-]**	5.000	2.000	1.000	0.500	0.200	0.100
estimated scale s **[-]**	3.673	2.106	1.051	0.528	0.224	0.124
deviation [%]	26.5	5.3	5.1	5.6	12.0	24.0

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
2. Ayvaci, A., Raptis, M., Soatto, S.: Sparse Occlusion Detection with Optical Flow. International Journal of Computer Vision (IJCV) (2011)
3. Fabbri, M., Lanzi, F., Calderara, S., Palazzi, A., Vezzani, R., Cucchiara, R.: Learning to detect and track visible and occluded body joints in a virtual world. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
4. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
5. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (2014)
6. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
7. López-Quintero, M.I., Marín-Jiménez, M.J., Muñoz-Salinas, R., Madrid-Cuevas, F.J., Medina-Carnicer, R.: Stereo Pictorial Structure for 2D articulated human pose estimation. Machine Vision and Applications (2015)
8. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
9. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
10. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: Proceedings of the International Conference on 3D Vision (3DV) (2017)
11. Munaro, M., Basso, F., Menegatti, E.: Tracking people within groups with rgb-d data. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2012)
12. Munaro, M., Menegatti, E.: Fast RGB-D people tracking for service robots. Journal on Autonomous Robots, Springer (2014)
13. Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., Bajcsy, R.: Berkeley MHAD: A comprehensive Multimodal Human Action Database. In: Proceedings of the IEEE Workshop on Applications of Computer Vision (WACV) (2013)
14. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam (2020)
15. Wu, C., Stoll, C., Valgaerts, L., Theobalt, C.: On-set performance capture of multiple actors with a stereo camera. In: ACM Transactions on Graphics (2013)