Biomechanically Accurate Vision and Geometry based human mesh recovery with a single RGBD sensor

Victor Nogues[1], Hugo Rodet[1], Julia Irvine[2], Elise K. Laende[3], Manuela Kunz[4], and Lama Seoud[1]

[1] Polytechnique Montréal, Canada
[2] Queen's University, Canada
[3] University of Waterloo, Canada
[4] National Research Council Canada

1 Introduction

Human pose and shape estimation is fundamental to medical applications. Traditionally, motion capture relied on costly marker-based infrared systems requiring extensive manual processing and specialized clinical environments. Although affordable RGB-D sensors have become ubiquitous, recent computer vision research has largely favoured RGB-only human mesh recovery, side-lining depth modalities despite their clinical potential. However, controlled settings like gait laboratories do not require complex "in the wild" algorithms and are well within the functional range of commercial RGB-D sensors. To bridge the gap toward biomechanical usability, this work leverages the SKEL [13] body model to regress biomechanically valid joint angles directly from the body surface, evaluating this approach on a lower-limb joint angle prediction task using a custom-collected dataset.

In this work, we exploit the explicit geometric information provided by the depth of RGBD sensors to propose an efficient kinematics and pose and shape recovery pipeline.

2 Related works

Markerless motion capture with RGB sensors. Several multi-camera markerless motion capture solutions like Pose2Sim and MAMMA offer robust OpenSim-based inverse kinematics (IK); others use joint triangulation or fitted skin landmarks. However, they all remain costly and complex to deploy. Consequently, monocular systems (e.g., CameraHMR) are increasingly explored as alternatives in clinical settings. Because SMPL joint angles lack biomechanical relevance, recent methods bridge this gap by applying synthetic surface markers to avatars for standard OpenSim IK processing.

Depth in human-centric tasks. While consumer RGB-D cameras (e.g., Kinect, Intel RealSense) garnered a lot of interest when they were first introduced, recent progress has largely favoured RGB-only methods [12,15] due to the availability of massive datasets and powerful vision backbones. Nonetheless, depth remains crucial for physical interaction tasks (e.g., robotics, hand gesture recognition [20,21]) and improves 3D metrics when effectively fused with RGB data [5,18]. However, literature on RGB-D-based human pose **and shape** estimation is scarce. Existing approaches either rely on deprecated tracking APIs [2] or slow, noise-sensitive fitting techniques (e.g., SegFit [14] using Human3D segmentation [16]).

Efficient exploitation of depth information. Beyond early or late fusion strategies [5, 18], depth can be converted into point clouds and processed via specialized architectures like PointNet [8] or Point Transformer [23].

3 Datasets

We train on a set of concatenated datasets, both synthetic or real. The datasets are presented in Tab. 1.

Dataset	Train	Validation	Test	Data type
BEDLAM [4, 17]	826,411	-	-	Synthetic, subset
HuMMan [6]	40,868	-	-	Real, subset
BEHAVE [3]	38,904	6,484	18,216	Real, full dataset
EgoBody [22]	845,708	248,506	617,330	Real, full kinect subset
InterCap [11]	-	-	400,249	Real, full dataset

Table 1: Description of the datasets used to train, validate and benchmark our method

4 Methods

We propose an optimization-free pipeline, which directly infers joint angles q, shape β and camera translation t_{cam} from a single RGBD image. Unlike conventional RGB models (e.g., CameraHMR [15], HSMR [19]) that rely solely on a ViT backbone and a single query, our architecture introduces a multimodal design combining RGB and geometric features through cross-attention with a set of per-parameter learned queries.

Image encoder. We use a ViT-s [10] backbone pretrained on an ImageNet [9] classification task. We simply remove the classification head.

Geometry encoder. A shallow CNN extracts local feature maps, which are then concatenated pixel-wise with the RGBD channels. These are lifted to 3D metric space using camera intrinsics to form a fixed-size point cloud, which is then processed by a 3-layer PointNet-like network.

Decoder. Inspired by DETR [7], a set of learned queries (representing joints, shape, and translation) iteratively cross-attend to both image tokens and geometric tokens, sharing information via self-attention layers before projecting to continuous rotation, shape, and metric translation spaces.

5 Results

Pose and shape. Our method is compared against CameraHMR and HSMR using standard errors (MPJPE, PA-MPJPE, and MVE) alongside additional metrics: Mean Translation Error (MTE) and Mean Joint Limits Threshold Crossing (MJLTC, which counts frames for which any predicted joint-angle is off-limits, derived from the ones presented in [19]). As indicated, MJLTC is omitted for CameraHMR (which uses SMPL), and MTE is omitted for HSMR (which uses weak perspective).

Dataset	Method	MPJPE	PA-MPJPE	MVE	MTE	MJLTC (%)
BEHAVE [3]	CameraHMR [15]	51.9	38.2	64.3	83.6	-
	HSMR [19]	94.6	51.3	108.0	-	11.3
	Ours	**45.5**	**36.5**	**56.8**	**38.3**	**10.5**
EgoBody [22]	CameraHMR	48.8	34.3	57.8	144.3	-
	HSMR	92.3	48.2	109.6	-	23.5
	Ours	**42.2**	**33.9**	**52.7**	**73.8**	**8.9**
InterCap [11]	CameraHMR	51.2	**32.3**	62.5	116.2	-
	HSMR	82.5	48.9	95.9	-	18.1
	Ours	**45.7**	32.8	**57.4**	**55.0**	**16.9**

Table 2: Results on three test datasets. Our method is compared to CameraHMR (with ground truth intrinsics provided) and HSMR. All metrics are in mm unless specified, best results are in bold, lower is better. MPJPE is computed on all 24 SMPL joints.

Tab. 2 shows that our method consistently matches or outperforms RGB-based state-of-the-art approaches across three public datasets, including the unseen InterCap dataset. Key advantages include: high performance by remaining completely optimization-free; lightweight execution by combining explicit depth geometry with a small RGB backend; and consistent runtimes by employing a fixed patch grid sampling step to ensure execution speed remains independent of the point cloud's density.

Kinematics. A dataset was collected using two Femto Bolt (Orbbec, USA), in a controlled lab environment. We refer to this dataset as AzureTheia. It comprises recordings of various motions (walking with eyes open, closed, and with a cane) using a dual-camera setup. 5 subjects (1 male, 4 females) with different ages and body types are recorded simultaneously by our dual-camera setup and Theia3D [1] reference system. We proceed to compare hips and knees flexion from both the Kinect Azure Body Tracking Toolkit (Kinect ABT) and our method to the ground truth from the Theia3D system.

We evaluate hip, and knee flexion accuracy using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Intraclass Correlation Coefficient (ICC), alongside 95% confidence intervals (2.5th and 97.5th percentiles). MAE and RMSE are first averaged per-sequence prior to final statistical computation. Tab. 3 targets situations where the participant faces the camera, whereas Tab. 4 contains sequences from the camera parallel to the patient's sagittal plane.

Side-camera results show excellent hip (ICC 0.94) and knee (ICC 0.92 to 0.94) reliability, outperforming Kinect ABT across all joints. Front-camera knee scores moderately decrease (ICC 0.84 to 0.86) due to depth ambiguity in the sagittal plane.

	Kinect ABT			Ours		
Flexion angle	MAE	RMSE	$ICC_{3,1}$	MAE	RMSE	$ICC_{3,1}$
Left hip	27.08	32.96	0.75 [0.73, 0.77]	10.17	11.54	0.91 [0.91, 0.92]
Right hip	29.35	35.15	0.76 [0.74, 0.77]	10.16	11.43	0.92 [0.92, 0.93]
Left knee	8.17	12.36	0.70 [0.68, 0.72]	6.45	8.38	0.84 [0.83, 0.85]
Right knee	8.44	13.14	0.68 [0.66, 0.70]	6.18	8.11	0.86 [0.85, 0.87]

Table 3: Comparative analysis on lower limbs flexion angles accuracy from the **front camera**. MAE and RMSE are in degrees. Ground truth angles are retrieved from a Theia3D markerless system. Limits of agreement at 95% are provided in brackets.

	Kinect ABT			Ours		
Flexion angle	MAE	RMSE	$ICC_{3,1}$	MAE	RMSE	$ICC_{3,1}$
Left hip	28.64	34.27	0.75 [0.72, 0.77]	9.31	10.12	0.94 [0.93, 0.95]
Right hip	30.66	36.08	0.73 [0.71, 0.76]	9.84	10.64	0.94 [0.93, 0.94]
Left knee	7.91	11.58	0.77 [0.74, 0.79]	4.64	5.86	0.94 [0.94, 0.95]
Right knee	7.57	10.83	0.79 [0.77, 0.81]	4.75	6.18	0.92 [0.91, 0.93]

Table 4: Comparative analysis on lower limbs flexion angles accuracy from the **side camera**. MAE and RMSE are in degrees. Ground truth angles are retrieved from a Theia3D markerless system. Limits of agreement at 95% are provided in brackets.

6 Conclusion

We present a compact, optimization-free RGBD framework for human pose and shape recovery using a dual-branch encoder and cross-attention decoder. By integrating metric depth directly, our method achieves state-of-the-art benchmark accuracy.

A proof of concept in a real lab setting against a Theia3D reference system shows excellent ICC for hip and knee flexion angles, substantially outperforming the Azure Kinect Body Tracking baseline. Future work will expand clinical validation to diverse populations and pathological gait analysis.

References

1. Theia Markerless - Markerless Motion Capture Redefined. https://www.theiamarkerless.ca/
2. Bashirov, R., Ianina, A., Iskakov, K., Kononenko, Y., Strizhkova, V., Lempitsky, V., Vakhitov, A.: Real-time RGBD-based Extended Body Pose Estimation. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 2806–2815. IEEE, Waikoloa, HI, USA (Jan 2021). https://doi.org/10.1109/WACV48630.2021.00285
3. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and Method for Tracking Human Object Interactions. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15914–15925. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01547
4. Black, M.J., Patel, P., Tesch, J., Yang, J.: BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8726–8737. IEEE, Vancouver, BC, Canada (Jun 2023). https://doi.org/10.1109/CVPR52729.2023.00843
5. Boudreault-Morales, G.E., Marquez-Chin, C., Liu, X., Zariffa, J.: The effect of depth data and upper limb impairment on lightweight monocular RGB human pose estimation models. BioMedical Engineering OnLine **24**(1), 12 (Feb 2025). https://doi.org/10.1186/s12938-025-01347-y
6. Cai, Z., Ren, D., Zeng, A., Lin, Z., Yu, T., Wang, W., Fan, X., Gao, Y., Yu, Y., Pan, L., Hong, F., Zhang, M., Loy, C.C., Yang, L., Liu, Z.: HuMMan: Multi-modal 4D Human Dataset for Versatile Sensing and Modeling. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, vol. 13667, pp. 557–577. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20071-7_33
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I. pp. 213–229. Springer-Verlag, Berlin, Heidelberg (Aug 2020). https://doi.org/10.1007/978-3-030-58452-8_13
8. Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J.: PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 77–85. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.16
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (Jun 2021)
11. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint Markerless 3D Tracking of Humans and Objects in Interaction. In: Andres, B., Bernard, F., Cremers, D., Frintrop, S., Goldlücke, B., Ihrke, I. (eds.) Pattern Recognition, vol. 13485, pp. 281–299. Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-031-16788-1_18

12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-End Recovery of Human Shape and Pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7122–7131 (Jun 2018). https://doi.org/10.1109/CVPR.2018.00744
13. Keller, M., Werling, K., Shin, S., Delp, S., Pujades, S., Liu, C.K., Black, M.J.: From Skin to Skeleton: Towards Biomechanically Accurate 3D Digital Humans. ACM Transactions on Graphics **42**(6), 1–12 (Dec 2023). https://doi.org/10.1145/3618381
14. Lascheit, K.: Robust Human Registration with Body Part Segmentation on Noisy Point Clouds
15. Patel, P., Black, M.J.: CameraHMR: Aligning People with Perspective. In: 2025 International Conference on 3D Vision (3DV). pp. 1562–1571 (Mar 2025). https://doi.org/10.1109/3DV66043.2025.00146
16. Takmaz, A., Schult, J., Kaftan, I., Akçay, M., Leibe, B., Sumner, R., Engelmann, F., Tang, S.: 3D Segmentation of Humans in Point Clouds with Synthetic Data (Aug 2023)
17. Tesch, J., Becherini, G., Achar, P., Yiannakidis, A., Kocabas, M., Patel, P., Black, M.J.: BEDLAM2.0: Synthetic humans and cameras in motion. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
18. Tziafas, G., Kasaei, H.: Early or Late Fusion Matters: Efficient RGB-D Fusion in Vision Transformers for 3D Object Recognition. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9558–9565. IEEE, Detroit, MI, USA (Oct 2023). https://doi.org/10.1109/IROS55552.2023.10341422
19. Xia, Y., Zhou, X., Vouga, E., Huang, Q., Pavlakos, G.: Reconstructing Humans with a Biomechanically Accurate Skeleton (Mar 2025). https://doi.org/10.48550/arXiv.2503.21751
20. Yao, Y., Fu, Y.: Real-Time Hand Pose Estimation from RGB-D Sensor. In: 2012 IEEE International Conference on Multimedia and Expo. pp. 705–710 (Jul 2012). https://doi.org/10.1109/ICME.2012.48
21. Yuan, S., Garcia-Hernando, G., Stenger, B., Moon, G., Chang, J.Y., Lee, K.M., Molchanov, P., Kautz, J., Honari, S., Ge, L., et al.: Depth-based 3d hand pose estimation: From current achievements to future goals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2636–2645 (2018)
22. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: EgoBody: Human Body Shape and Motion of Interacting People from Head-Mounted Devices. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022, vol. 13666, pp. 180–200. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20068-7_11
23. Zhao, H., Jiang, L., Jia, J., Torr, P., Koltun, V.: Point Transformer. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16239–16248 (Oct 2021). https://doi.org/10.1109/ICCV48922.2021.01595