Kinematically-consistent Real-time 3D Human Body Estimation for Physical and Social HRI

Lorenzo Ferrini
PAL Robotics
Barcelona, Spain
lorenzo.ferrini@pal-robotics.com

Séverin Lemaignan
PAL Robotics
Barcelona, Spain
severin.lemaignan@pal-robotics.com

Abstract—We present a software tool, fully integrated with ROS, that enables robots to perceive people full body in 3D. The system works either with a simple RGB camera, or a RGB-D camera for better 3D absolute position estimation. The system is based on Google Mediapipe, and runs at > 8Hz on CPU. The consistency of the human kinematic model is ensured by relying on a URDF-defined kinematic model, that could be adjusted to each person's anthropometric characteristics.

Index Terms—Physical Modeling, Pose Estimation, ROS4HRI

I. Introduction

In the context of Social Assistive Robots (SARs), and more in general in Human-Robot Interaction, it is fundamental for a robot to be aware of the 3D pose of the human it is interacting with. Obtaining good and efficient results in this direction represents a stepping stone toward complex supportive and collaborative tasks. We can not imagine robots fulfilling everyday assistive tasks as, for instance, walking support for elderly people, without a good geometric estimation of the person it is trying to help. The 3D human joint estimation should happen smoothly, avoiding high computational burden as this operation is usually just part of a greater modular system.

With this work, we are presenting the software developed to evaluate human joints and links poses through 3D skeleton estimation consistently with an explicit human kinematic model. This system exploits the 3D human pose estimation inferred from RGB images by Google Mediapipe's Holistic Model estimator [1], combining it with the human body kinematic model already defined in [2] as part of the ROS4HRI framework, to obtain joints and links poses actually consistent with the human movement capabilities. Finally, the system publishes the information about the poses as a ROS tf tree directly connected to the camera link frame. To enable everyone to test and run the system, it is possible to start using this with a simple RGB camera; however, including depth information from an RGB-D sensor permits to detect body distance and movements along the plane.

II. Related Work

Human body estimation is fundamental for almost any application involving physical or social interactions between humans and robots; over the years, researchers proposed different solutions, roughly split between invasive and non-invasive techniques.

In the former case, solutions usually provide good results in terms of both computational burden and joint angles estimation precision: this is the case of IMUs based approaches [3], where wearable sensors information gets processed and filtered to obtain precise joints angles estimations, and optical markers based solutions like OptiTrack, which precision made it a safe choice for robotic applications [4] as well as movie industry. These solutions pays the excellent estimation results provided in terms of ease-of-use, a fundamental property when it comes to SARs, where the user experience is a crucial aspect of the overall system performance evalutation [5].

On the other hand, non-invasive approaches have become more and more popular over the last years, thanks to the improvements achieved in 2D and 3D deep learning-based human pose estimation from RGB images, both in terms of precision and computational cost. In [6] authors compare the results obtained from a markerless deep-learning based 3D pose estimation method, OpenPose [7], with those produced by an optical-markers based technique: they show how the markerless technique is able to provide 3D pose estimation affected by a limited Mean Average Error when compared to the markers-based one, while offering a better experience in terms of ease-of-use.

Many proposed non-invasive 3D human pose estimation solutions rely on point clouds computation and access [8]. In [9], authors realize a ROS-based tool to estimate 3D skeleton pose starting from OpenPose 2D skeleton estimation and directly accessing the generated point cloud at the 2D landmarks to obtain their 3D coordinates in camera frame. A similar approach provides good results in [10], where authors specifically acknowledge [9] work: in this case, precision gets tested through Aruco markers-related measurements. In [11] authors describe a deep learning architecture aimed at estimating human 3D pose starting from 2D skeleton keypoints, inferred from the RGB image, and the depth image. Differently from what happens with 3D pose estimation directly computed from RGB images, this solution limits the number of assumptions needed: this technique implicitly takes into account the fact that humans have different heights and shapes.

None of the aforementioned works explicitly rely on a consistent kinematic model of the human body. Such a structure is a safe way to specify that humans are not capable to perform any given movement one could think of, since our joints have rotation limits: self-occlusions, for instance, could lead to estimation of poses that appear to be impossible for a human. Moreover, through model parameters, it is possible to use information about the human real size to obtain realistic 3D pose estimations starting from the joint angles inferred from RGB images.

III. Proposed Approach

Our work aims to provide a solution for real time 3D skeleton estimation, publishing the results as a tf tree originating from the camera link; tf is the ROS system to generate and manage frames. A stable and trustworthy 3D skeleton estimation for human-robot interaction must guarantee the consistency of the human kinematic model in order to avoid, for instance, self-occlusions related issues: a combination of 3D skeleton estimation from RGB images and the predefined ROS4HRI human kinematic model will address the problem.

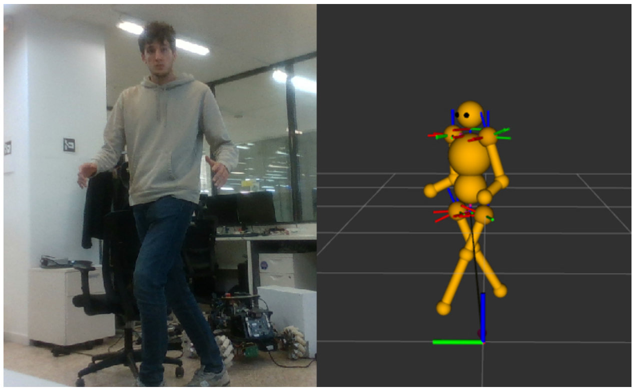

Fig. 1. Human body pose estimation; it is possible to see that, at this point, the system does not detect head rotation.

Thanks to Google's Mediapipe APIs [1], it is easy to implement in robotics (or any other kind of application) state-of-the-art 3D skeleton estimation from RGB images. The Mediapipe holistic model estimator applies BlazePose model to estimate 3D skeleton poses, providing 33 3D landmarks, with the person modeled as they were in a 2m x 2m x 2m cubic space, with the reference frame placed in the middle point between their hips and each axis range spanning between -1m and 1m. From the origin of the reference frame, Z is positive getting closer to the camera. The holistic model estimation also includes face mesh, 2D skeleton pose and hands pose, with the limit of being able to evaluate just one person in the scene. We use the obtained 3D skeleton pose values, expressed in cartesian space, to extract human joints angles through an inverse kinematic process; then, we generate and publish a jointstate message. A robot_state_publisher proceeds computing the human body kinematic estimation through a direct kinematic process, using the jointstate message values, possibly taking into account specific body shapes and dimensions that the Mediapipe estimation is not able to detect.

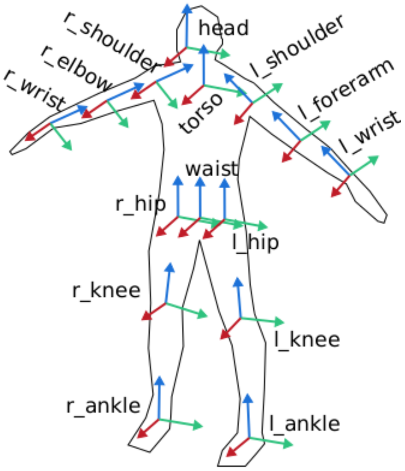

Fig. 2. ROS4HRI human body links and frames definition from [2].

The presence of an RGB-D sensor is useful to include depth and along-the-plane movement information, two values that provide an overall more informative estimation of the body position in the environment visible to the robot camera. We define a tf trasformation between the camera frame and the human model reference frame using the 3D coordinates of the Mediapipe reference frame origin point, without involving any point cloud in the process: starting from the 2D skeleton estimation, we evaluate the pixel location of the origin point in the RGB image and then, through a frame transformation process, we access the corresponding depth image pixel to compute the XYZ values. Eventually, we compute the body rotation along the vertical axis. This angle comes from the evaluation of the right hip rotation along the transversal plane.

$$\theta = \operatorname{arctan2}\left(\frac{z_{rh}}{x_{rh}}\right) \qquad (1)$$

Once we have gathered all this information, we define a specific transformation between the origin of the body model and the camera link using tf tools, completing the human body estimation and frame publishing process. One single ROS node manages all the operations described in this section, subscribing to 1 to 4 topics (which depends on the type of camera available):

- /rgb_image.
- /depth_image.
- /rgb_info, containing RGB camera useful parameters for RGB to depth frame transformation.
- /depth_info, containing depth camera parameters for RGB to depth frame transformation.

IV. Results

Rviz permits to visualize, through the RobotModel plugin, the human body model, updating it according to the new tf transformations computed by the robot state publisher (Fig. 1). This allowed us to directly check the results of the proposed approach and related software, performing movements and comparing these with the displayed body model; we have experienced flowing, realistic model movements, following our own real poses. The system, which runs without exploiting GPU acceleration, performs at around 8 Hz for 1 person detection on an Intel® Core™ i7-7700K CPU.

V. Future Work

We have achieved satisfactory results for a single human body estimation, both in terms of perceived movement precision and flowing pose transition. However, the limit to one person detected in the scene is too strict, and we need to move forward in this sense. There are two possible solutions: one is retraining the Mediapipe holistic model estimator to be able to detect more people inside of the image; otherwise, we could try to detect the people in the scene using an object detection system (e.g., YOLO) and perform, on each one of the detected people, an holistic model estimation. We will investigate further to understand which solution could fit better our research, considering that this tool is meant for real-time use. We are also currently not exploiting the height parameter available in the the human kinematic model xacro file; this reflects non-realistic equally tall representations for every person undergoing the pose estimation process. We are working on the height estimation process, which should take into account 2D skeleton estimation, landmarks visibility and XYZ coordinates computation. We must underline that currently this work does not address the existence of different body shapes, which is another limit that we need to overcome, as well as providing a robust automated testing solution for the estimation results.

VI. Acknowledgment

This work was supported by the PERSEO project, which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 955778.

References

[1] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee et al., "Mediapipe: A framework for building perception pipelines," arXiv preprint arXiv:1906.08172, 2019.

[2] Y. Mohamed and S. Lemaignan, "Ros for human-robot interaction," arXiv preprint arXiv:2012.13944, 2020.

[3] H. Ahmed and M. Tahir, "Improving the accuracy of human body orientation estimation with wearable imu sensors," IEEE Transactions on instrumentation and measurement, vol. 66, no. 3, pp. 535–542, 2017.

[4] B. Busch, G. Maeda, Y. Mollard, M. Demangeat, and M. Lopes, "Postural optimization for an ergonomic human-robot interaction," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 2778–2785.

[5] J. Lindblom and R. Andreasson, "Current challenges for ux evaluation of human-robot interaction," in Advances in ergonomics of manufacturing: Managing the enterprise of the future. Springer, 2016, pp. 267–277.

[6] N. Nakano, T. Sakura, K. Ueda, L. Omura, A. Kimura, Y. Iino, S. Fukashiro, and S. Yoshioka, "Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras," Frontiers in sports and active living, vol. 2, p. 50, 2020.

[7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," IEEE transactions on pattern analysis and machine intelligence, vol. 43, no. 1, pp. 172–186, 2019.

[8] T. Xu, D. An, Y. Jia, and Y. Yue, "A review: Point cloud-based 3d human joints estimation," Sensors, vol. 21, no. 5, p. 1684, 2021.

[9] M. Arduengo, S. Jorgensen, K. Hambuchen, L. Sentis, F. Moreno-Noguer, and G. Alenya, "Ros wrapper for real-time multi-person pose estimation with a single camera," Institut de Robotica i Informatica Industrial, CSIC-UPC, Tech. Rep, 2017.

[10] F. Lygerakis, A. C. Tsitos, M. Dagioglou, F. Makedon, and V. Karkaletsis, "Evaluation of 3d markerless pose estimation accuracy using openpose and depth information from a single rgb-d camera," in Proceedings of the 13th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2020, pp. 1–6.

[11] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 1986–1992.