2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Exploring 3D Human Pose Estimation and Forecasting from the Robot's Perspective: The HARPER Dataset

Andrea Avogaro[1*], Andrea Toaiari[1*], Federico Cunico[1*], Xiangmin Xu[2], Haralambos Dafas[2], Alessandro Vinciarelli[2], Emma Li[2] and Marco Cristani[1]

***Abstract*— We introduce HARPER, a novel dataset for 3D body pose estimation and forecasting in dyadic interactions between users and Spot, the quadruped robot manufactured by Boston Dynamics. The key-novelty of HARPER is its focus on the robot's perspective, *i.e.*, on the data captured by the robot's sensors. This makes 3D body pose analysis challenging, as being close to the ground results in only partial captures of humans. The scenario underlying HARPER includes 15 actions, of which 10 involve physical contact between the robot and users. The corpus contains recordings not only from Spot's built-in stereo cameras but also from a 6-camera OptiTrack system, with all recordings synchronized. This setup leads to ground-truth skeletal representations with a precision of less than a millimeter. Additionally, the corpus includes reproducible benchmarks for 3D Human Pose Estimation, Human Pose Forecasting, and Collision Prediction, all based on publicly available baseline approaches. This enables future HARPER users to rigorously compare their results with those provided in this work.**

I. Introduction

One of the main changes characterizing the transition from Industry 4.0 to Industry 5.0 is the shift from Human-Robot *Interaction* to Human-Robot *Collaboration* [1]. This shift requires robots to evolve into cobots—intelligent platforms equipped with capabilities like visual perception, action recognition, intent prediction, and safe online motion planning. These technologies empower cobots with awareness of their surroundings, enabling them to adapt their behaviour in real-time, a stark contrast to the rigid, pre-programmed routines of traditional robots [2]. In other words, understanding human behavior is a crucial requirement for a robot to become a cobot and, correspondingly, to be capable of adaptive and seamless interaction with its users [3].

Thus motivated, we propose *Human from an Articulated Robot Perspective* (HARPER), a new, publicly available dataset revolving around the interaction between human users and Spot, the quadruped robot manufactured by Boston Dynamics. Such a platform attracts increasingly more attention, and, not surprisingly, it was recently included in *Habitat 3.0* [4], one of the most advanced environments for simulating Human-Robot interactions. Moreover, Spot is an ideal cobot candidate for at least three reasons: 1) the four-legged design and the biologically inspired locomotion enable it to operate on diverse and challenging terrains (the robot can even climb stairs [5]), thus making Spot a potential companion in a wide range of settings [6]–[9]; 2) Spot is equipped with one of the most advanced self-balancing systems available on the market, significantly limits the risk of accidents during close physical interaction with users; and 3) Spot is equipped with a total of 5 greyscale + depth sensors mounted on its body and an RGB-D camera on its grasper arm (see Fig. 1). This is important because such a sensing apparatus makes Spot particularly suitable for analysis and understanding of human behavior, a critical step in the evolution from robot to cobot.

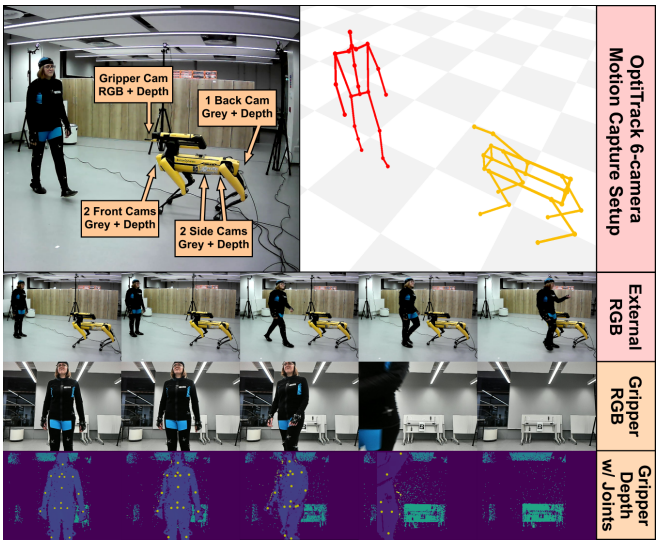


Fig. 1. HARPER Showcase. (Top-left) We exploit the Spot on-board equipment to let the robot perceive people. (Top-right) Thanks to a 6-camera OptiTrack setup we capture 3D human poses represented with 21-joints and 0.035 mm of error. (Second row) An additional external RGB camera shows the actions performed. (Third row) The gripper RGB camera Point of View. (Fourth row) The gripper depth camera Point of View, with the ground truth joints in yellow.

HARPER includes dyadic interactions between Spot and 17 human users, 5 females and 12 males, each performing 15 actions that require different degrees of collaboration with the robot (see Section III for more details). The data captured with the Spot sensors were enriched with the recordings of a 6-camera OptiTrack Motion Capture (MoCap) system

*Equal contribution
[1]University of Verona, Department of Engineering for Innovation Medicine, Italy. (e-mail: name.surname@univr.it)
[2]University of Glasgow, School of Computing Science, UK
Project page: https://intelligolabs.github.io/HARPER
This work has been partially financed within the PNRR research activities of the consortium iNEST (Interconnected North-Est Innovation Ecosystem) funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4 Componente 2, Investimento 1.5 – D.D. 1058 23/06/2022, ECS_00000043) and by the UKRI Centre for Doctoral Training in Socially Intelligent Artificial Agents, Grant Number EP/S02266X/1

capable of extracting skeletal models of the users. The joints were localized with a precision of less than one millimeter (see Fig. 1), thus providing highly accurate ground-truth information about the pose and position of the users. This is a significant advantage because Spot's sensors and MoCap cameras are synchronized. Therefore, skeletal models can be used to reliably validate approaches for human behavior analysis and understanding based solely on Spot sensors.

In addition to the above, skeletal representations enable one of the key novelties of HARPER, namely the possibility to train approaches capable of recognizing 3D body pose and movement when the Spot, due to its limited height, can "see" its users only partially, something that happens whenever the distance between them is small. To the best of our knowledge, this is one of the first datasets that allows the investigation of such a problem in 3D.

We asked the 17 HARPER participants to stage two types of physical contact with the robot, namely *unintentional* and *intentional*, according to the terminology proposed in [10]. The first type includes (staged) collisions, while the second includes punches, kicks and soft touches. We paid special attention to the first type because of the major role collisions play in scenarios based on co-located interactions. We enriched HARPER with benchmarks, *i.e.*, reproducible experimental protocols and baseline approaches designed to address three tasks relevant to the analysis of physical contact, namely 3D Human Pose Estimation (especially when Spot can "see" its users only partially), 3D Human Pose Forecasting and Collision Prediction. This allows researchers interested in HARPER to rigorously compare their results with those presented in this article (see Section IV).

Overall, the main contributions of the paper can be summarized as follows:

- We propose the first dataset that includes not only the "point of view" of the robot (the data captured with the Spot's sensors) but also a panoptic point of view (the data captured with the MoCap system) that provides accurate ground-truth information for position and pose of both users and robot;
- To the best of our knowledge, HARPER is the first dataset enabling the reconstruction of the human users' pose with the data captured with a quadruped robot, a problem which is challenging because Spot is short and the subject is usually close to it (hence, the cameras cannot capture the whole body of the user);
- HARPER allows, for the first time, the visual prediction of collisions between a mobile robotic platform and users.

The rest of this paper is organized as follows: Section II surveys previous work, Section III describes HARPER in detail, Section IV presents the benchmarks, and Section V draws the conclusions.

II. Related Work

Table I shows the main differences between HARPER and existing datasets of similar scope. Most available corpora are based on the analysis of people's trajectories. The THÖR dataset [11], a well-known example, contains the 2D trajectories of 9 human users moving together with a robot. Besides this, the data includes 6D head positions, LiDAR data from a stationary sensor, and the participants' orientations and eye gaze directions. THÖR-Magni [12], a significantly more extensive dataset from the authors of THÖR, introduces onboard sensors on the mobile robot and semantic attributes describing the roles and activities of detected people. Similarly, the JRDB [13] dataset aims to enable mobile robots to detect and track humans in both indoor and outdoor settings. The data includes stereo cylindrical RGB videos and LiDAR point clouds annotated with 2D and 3D bounding boxes. In addition, the dataset includes benchmarks for both 2D and 3D detection and tracking. A more recent version includes 2D human-pose skeletal annotations [19].

Other datasets provide information about objects that the robots can encounter while moving. For example, CODa [16] aims at both object detection and semantic segmentation. It was acquired with a wheeled robot, and it features sequences in indoor and outdoor settings of a university campus, as well as 3D semantic segmentation and 3D object detection benchmarks. In the case of FROG [15], based on LiDAR sensors placed on a wheeled robot at roughly the height of human knees, the problem is the detection of people in possibly crowded sites where humans can be confused with static and dynamic obstacles. A similar issue is at the core of the dataset proposed in [17], where the material is collected with an RGB-D camera mounted on a small mobile robot. The annotations include attributes such as, *e.g.*, the presence of static obstacles, illumination and humans' poses. An OptiTrack MoCap system provides information about the position of both the robot and users. The problem of navigating through an environment, possibly shared with humans, is the focus of HuRon [20]. The data was collected using a Roomba bot with LiDAR, bumper collision detectors, video, and odometry sensors. However, no pose annotation is provided about the people sharing the space with the robot.

HARPER shows major novelties with respect to the datasets above. The availability of 3D skeletons for both the robot and users provides unprecedentedly detailed information about the interaction between the two, especially when considering that the joints are captured with a precision of less than one millimeter. A similar acquisition precision is achieved with InHARD [18], an industrial HRI dataset featuring both RGB images and MoCap data of a person performing multiple manual tasks, captured with wearable devices. A robotic arm, mostly stationary, is the platform used for the experiments, and it never collides with the user, offering a looser type of interaction. This is not the case in HARPER, which includes physical contacts of different types. In [24], a mobile wheeled robot is employed to capture an HRI dataset in a retail environment. Multiple people navigate the room and perform picking and sorting actions while the robot moves along with them. Egocentric and scene videos, eye gaze directions, point clouds, and other data are collected. The human poses are collected through

TABLE I

MAIN HRI DATASETS REVOLVING AROUND HUMAN MOVEMENT AND ITS ANALYSIS. VALUES IN THE PARTICIPANTS COLUMN INDICATED WITH THE ASTERISK (*) REFER TO DATASETS CAPTURED IN UNCONTROLLED SCENARIOS.

Dataset	Participants	Actions	Mobile Robot	Robot POV	Human Skeleton	Human Joints	Marker-Based MoCap	Robot Skeleton	Collisions / Intended Contacts
THÖR [11]	9	13	✓	✗	✗	✗	✓	✗	✗
THÖR-Magni [12]	40	48	✓	✓	✗	✗	✓	✗	✗
JRDB [13]	3.5K*	N/A	✓	✓	✗	✗	✗	✗	✗
L-CAS Multisensor [14]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
FROG [15]	1M*	N/A	✓	✓	✗	✗	✗	✗	✗
CODa [16]	N/A*	N/A	✓	✓	✗	✗	✗	✗	✗
PTUA [17]	N/A	N/A	✓	✓	✗	✗	✓	✗	✗
InHARD [18]	16	14	✗	✗	3D	17	✗	✗	✗
JRDB-Pose [19]	5K*	N/A	✓	✓	2D	17	✗	✗	✗
HuRoN [20]	N/A* (5/17 for exp)	N/A	✓	✓	✗	✗	✗	✗	✓
NatSGD [21]	18	11	✗	✓	estim. 2D	25	✗	Arm	✗
CHICO [22]	20	7	✗	✗	2D, 3D	15	✗	Arm	✓
SCAND [23]	N/A* (14 for exp)	12	✓	✓	✗	✗	✗	Quadruped, Wheeled	✗
UF-Retail-HRI [24]	8	2	✓	✓	3D	23	✗	Arm	✗
HARPER	17	15	✓	✓	2D, 3D	21	✓	Quadruped	✓

an IMU-based MoCap device, which requires careful setup and calibration for every person. However, the Spot used in HARPER is a more advanced robotic platform, and its movement is significantly less constrained.

Skeleton representations were also used in other corpora. In [22], the scenario is a collaboration between a user and a robotic arm in an industrial setting. A MoCap system captures the user's skeleton from an external point of view, missing the robot's perspective. Furthermore, the acquisition is markerless, making the joint localisation less precise. In another dataset, the multimodal NatSGD [21], the goal is imitation learning, and the data includes human commands, such as speech and gestures, with a focus on robot behavior in the form of synchronized robot trajectories. However, the joint localization is, once again, less precise than in HARPER because it is performed by applying Openpose [25] to videos.

Finally, to the best of our knowledge, the only other dataset in which the robot Spot was actually involved is SCAND [23], where two robots, a wheeled one and the Spot, are teleoperated in human-populated environments. A large variety of data is acquired thanks to an additional LiDAR sensor mounted on the two robots. However, no skeletal models are considered for humans, a major difference with respect to HARPER. The dataset we propose appears to have distinctive characteristics with respect to those currently available in the literature.

III. THE HARPER DATASET

The main motivation behind the design of HARPER is to expand the research opportunities enabled by previous HRI datasets (see Section II), especially considering the transition from robots to cobots. The collection of the corpus involved 17 participants who were asked to perform 15 actions (the same for all participants). The data was captured with the sensors equipped on Spot: 5 greyscale + depth sensors and one RGB-D camera mounted on the gripper. Moreover, we used 6 MoCap sensors (OptiTrack) and one RGB camera capturing the full setting (see Fig. 2).

Overall, HARPER contains 607 sequences for a total of over 60000 RGB images, grayscale images, depth frames, and 3D data from multi-sensor recordings. In the following, we discuss the acquisition setup (Sec. III-A), we describe the actions we captured and their annotations (Sec. III-B), and, finally, we provide key statistics about the data (Sec. III-C).

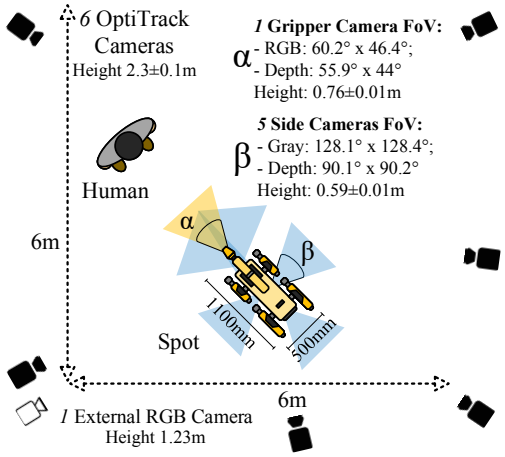


Fig. 2. A 6-camera OptiTrack system covers a 6×6 squared meters area where users and Spot can freely move. The external RGB camera's Field of View covers the setting. The 5 Spot on-body greyscale + depth cameras and the RGB-D frontal camera (gripper) cover the environment surrounding the robot.

A. Acquisition Setup

We collected all data in a laboratory (the layout is in Fig. 2). The 6 cameras of the OptiTrack MoCap system were arranged to cover a $6 \times 6\ m^2$ area, free of obstacles, in which the participants performed the 15 actions of the HARPER scenario. All participants wore a motion capture suit with 37 reflective markers distributed according to the OptiTrack *Baseline Marker Set* configuration. After calibration, the OptiTrack tracks the markers with a 0.035 mm error at a sampling frequency of 120Hz. Furthermore, thanks to the

configuration above, the OptiTrack software (Motive) automatically extracts a 21-joint skeleton representation based on the marker positions.

The robot involved in the experiment is the Boston Dynamics Spot, a 12-DoF (3 per leg) quadruped robot equipped with 5 stereo cameras (greyscale with depth) around its body and one RGB-D camera on the gripper. The Spot acts within the OptiTrack area described above, and its skeleton is obtained by applying forward kinematics to its internal motors state, acquired through the API provided by Boston Dynamics. The Spot skeleton is positioned in the same 3D scene as the participants' skeletons using a 4-marker rigid body mounted on its back and tracked by the OptiTrack.

The Spot cameras operate at about 10 FPS and their data is synchronised with the OptiTrack data. This ensures one of the most distinctive features of HARPER, i.e., the availability of two view points, one from the robot and a panoptic view that covers the whole scene. The synchronization is achieved by aligning the timestamps of the data, with a temporal alignment error of less than 2 milliseconds. It is worth noticing that the overlap between Spot cameras is limited to the 3 frontal cameras with a very partial overlap. As a reference, we added an external RGB camera, positioned outside the OptiTrack delimited area, to capture the whole scene (see bottom left part of Fig. 2). All the videos recorded with such a camera are provided with the dataset.

B. Actions and Annotations

We invited 17 university students to participate in the data collection (5 females and 12 males). They all signed an informed consent letter, and the information they provided, including the data collected during their participation, was handled according to the ethical regulations of the university in which the data was collected. Each participant interacted with the Spot individually in a session that followed a consistent sequence of steps. First, participants were assisted in wearing the suit necessary for marker tracking (see above), and then they were asked to display a T-pose for calibrating the skeleton extraction.

After calibrating the OptiTrack, we asked the participants to perform 15 actions designed to reproduce different situations (Table II), with the robot standing still for 8 actions and moving for 7 actions. In particular, the participants were instructed to act collisions as realistically as possible, *i.e.*, as if they were accidentally and unintentionally bumping into the Spot. The area covered by the OptiTrack is sufficiently broad to perform the actions comfortably. However, some participants moved inadvertently out of it, leading to missed markers in a few frames. Similarly, some occlusions prevented the OptiTrack from working properly for a few moments. These issues concerned no more than 3% of the total frames, and missing markers were effectively replaced through linear interpolation, ensuring that the skeleton representation was acquired with continuity and with the same precision at all times.

OptiTrack and Spot share the same reference system, which enables the projection of 3D skeletons onto the videos captured with the robot's cameras (greyscale and RGB). This allows for accurate annotation of the joint positions in the video. In addition, given that the robot's leg motor state is known, forward kinematics was applied to compute the position of the robot's joints in the 3D space, starting from the rigid body mounted on the Spot back. This allowed us to obtain a 21-joint representation of both the participants' skeletons and the robot's skeleton.

TABLE II

HARPER ACTIONS. THE EXPRESSION *Contact* MEANS THAT THE DISTANCE BETWEEN SPOT AND USER IS LOWER THAN 10 CM.

Action	Action Description	Robot Moving	Contact
A1 Walk+Crash Frontal	Human walks towards Spot oriented frontally then collides;		✓
A2 Walk+Crash 45°	Human walks towards Spot oriented at 45° then collides;		✓
A3 Walk+Crash Sideway	Human walks towards Spot oriented at 90° then collides;		✓
A4 Walk+Crash Backwards	Human walks towards Spot oriented backwards then collides;		✓
A5 Walk+Stop	Human walks towards Spot, then stops right before colliding;	✓	
A6 & A7 Walk+Avoid	Human and Spot walk towards each other avoiding collision at last second on the right (A6) / left (A7).	✓	
A8 Walk+Touch	Human walks towards Spot, then physically touch it;		✓
A9 Walk+Kick	Human walks towards Spot, then kicks it;		✓
A10 & A11 Walk+Punch	Human walks towards Spot oriented at 0° (A10) / 90° (A11), then punches it		✓
A12 Circular Walk	Human and Spot walk together in a circular path	✓	
A13 Circular Follow +Touch	Human follows Spot in a circle, then touches it with the hand	✓	✓
A14 Circular Follow + Avoid	Spot follows the human in a circle, then avoids contact	✓	
A15 Circular Follow + Crash	Spot follows the human in a circle, then a collision happen	✓	✓

C. Dataset Statistics

Fig. 3a shows, for all possible values of n, the number of frames in which exactly n human skeleton joints are visible to the robot. This information is important for understanding the difficulty level involved in addressing one of the new tasks that HARPER enables, namely analysis and understanding of human poses when they are only partially visible. Similar information is shown in Fig. 3b, where human joints are grouped according to five body parts, *i.e.* head (2 joints), torso (5 joints), left/right arm (3 joints), and left/right leg (4 joints). The figure reports the percentage of times these body parts are visible (one joint is sufficient for the part to be considered visible) to each camera. One of the main patterns

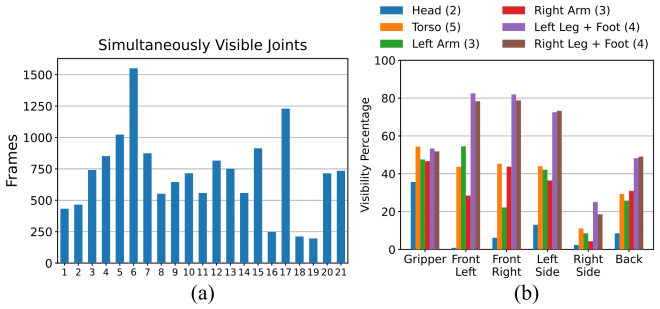


Fig. 3. Joints visibility from the robot's perspective. The left chart shows how many frames contain exactly n joints for $n = 1, \ldots, 21$. The right plot shows the percentage of frames in which the different parts of the skeleton are visible.

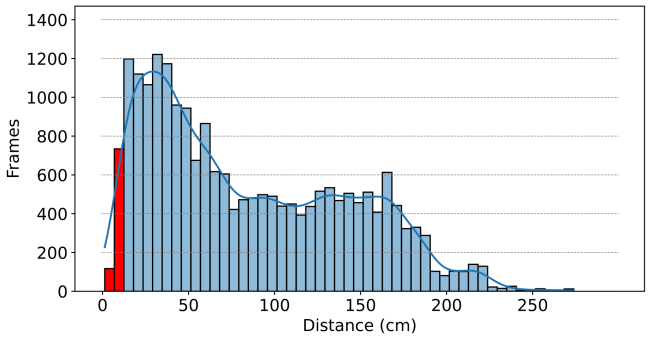


Fig. 4. Distribution of distances between Spot and users (the distance considers the two closest joints of human and robot). Red columns correspond to distances lower than 10 cm, considered as cases of physical contact.

is that the gripper camera is more likely to capture the upper part of the body and legs, but not the feet (*i.e.* the spike on 17 visible joints caused by the gripper camera field of view).

Regarding the interaction between Spot and the participants, Fig. 4 shows the histograms of the distances between their closest joints. Two modes appear: below and above a distance of 1.3 meters. Distances corresponding to physical contact are highlighted in red. A threshold distance of 10 cm was used to determine whether physical contact was happening (see details in Sec. IV-C).

IV. Experimental Evaluation

HARPER provides three benchmarks: *3D Human Pose Estimation* (3D-HPE), *3D Human Pose Forecasting* (3D-HPF), and *Collision Prediction* (CP). All benchmarks are from the *robot's perspective*, *i.e.*, they are based on data captured with the robot's sensors, which is one of the key novelties of HARPER. Participants S1-S12 were used for training (15984 frames), while participants S13-S17 were used for testing (5542 frames). For 3D-HPF, we sampled 10990 sequences (of 20 frames each) for the training set and 3502 for the test set, keeping the same distribution of participants. The sequences were sampled by using a rolling window with a 1-frame step. Sequences without visible joints were excluded from the test set.

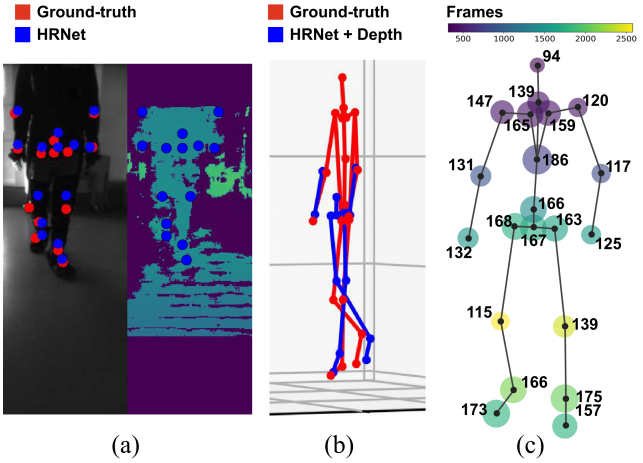


Fig. 5. Results of 3D human pose estimation from the robot perspective. (a) On the left, the predicted 2D joints (in blue) by HRNet [26] and the corresponding ground-truth joints (in red). On the right is the depth image with the same 2D detections. The depth will serve to do the lifting. (b) The lifted 3D poses alongside the complete OptiTrack skeletons. (c) MPJPE (in mm) for every visible joint (inside the depth Field of View) on the test set. The size of the blobs is proportional to the errors, while colors are related to the number of times a joint is visible from the robot's perspective.

A. 3D Human Pose Estimation

3D-HPE from robot's perspective involves finding the 3D coordinates of the visible human joints using greyscale images and synchronized depth maps captured by the robot's sensors. The main challenge is that humans may not be fully visible (Section III). The proposed baseline approach first uses a 2D pose estimator to find the position of the visible joints and then computes their 3D positions by exploiting the depth values, as shown in [27] (Fig. 5).

The 2D pose estimator is HRNet [26], trained on HARPER training data after resizing the images to 256×256 (no augmentation is applied). The depth sensors' Field of View (FoV) is narrower than that of the video cameras. Therefore, if an estimated joint is out of the depth FoV, it is considered non-visible. The positions of the joints, along with their corresponding depth values, can then be mapped into the 3D OptiTrack coordinate system. Once this task is completed, the 3D points inferred by the approach can be compared with those of the MoCap ground-truth skeleton.

We evaluated 2D pose estimation performance with the Percentage of Correct Keypoints (PCK) [28], *i.e.*, the fraction of correct predictions within a distance threshold τ (set to 0.5 on the predicted heatmaps). For the 3D joints estimation, we used the Mean Per Joint Position Error (MPJPE) [29], *i.e.*, the mean Euclidean distance between the visible estimated joints and the ground-truth OptiTrack ones. We obtained a PCK of 86.8% and an average MPJPE (computed only on the visible joints) of 151 mm on 2D and 3D poses, respectively (Fig. 5). The 2D baseline performs well, especially when considering that, in many cases, only one limb is visible or the participant is very close to the Spot. The 3D lifting shows some limitations due to the noise in the depth maps, especially when the participants are far from the Spot.

TABLE III

POSE FORECASTING ERRORS. WE PROVIDE THE MPJPE EXPRESSED IN MM WITH A PREDICTION HORIZON OF 400 AND 1000 MS. THE ERRORS ARE COMPUTED FOR THE PARTICULAR FRAME FOR EACH ACTION (FIRST NINE COLUMNS) AS WELL AS THE AVERAGE OVER ALL FRAMES (*Average*) AND THE AVERAGE OVER THE LAST FRAME OF EACH ACTION INSTANCE (*Last frame average*).

Actions		**A1-4**		**A5**		**A6-7**		**A8**		**A9**		**A10-11**		**A12**		**A13**		**A14**		**A15**		***Average***		***Last frame average***	
Time (ms)		*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*	*400*	*1000*
STS-GCN	GT	115	169	104	150	99	134	116	129	143	218	144	213	148	229	179	268	112	148	150	224	120	171	147	260
	GT+R	182	229	121	153	281	340	126	145	148	213	154	224	156	234	193	288	262	302	306	338	190	242	212	326
	HRNet+D+R	414	469	188	249	559	674	276	306	197	283	257	340	204	294	337	430	646	684	560	606	382	453	400	538
SiMLPe	GT	63	150	60	140	42	99	31	72	66	145	77	183	89	208	99	221	46	108	90	216	60	141	98	264
	GT+R	158	231	100	174	261	351	87	121	82	163	114	199	112	226	137	260	231	302	280	353	158	237	193	359
	HRNet+D+R	436	528	213	275	674	917	290	368	232	349	310	451	205	307	412	610	681	820	785	1170	441	595	511	790
EqMotion	GT	42	110	37	91	25	61	23	60	40	104	60	139	66	174	73	164	35	96	68	168	41	104	69	197
	GT+R	147	200	81	125	255	326	79	108	61	124	102	169	92	188	112	185	221	286	267	328	146	205	171	299
	HRNet+D+R	419	488	194	234	569	660	274	299	187	237	262	319	187	285	340	407	632	678	613	628	386	446	422	552

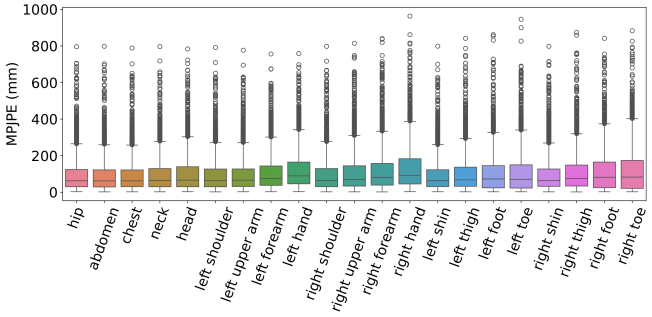


Fig. 6. MPJPE for each joint using EqMotion with GT as input and a forecasting horizon of 1000 ms.

However, the performance was sufficient to address 3D-HPF and CP, both from the robot's perspective.

B. 3D Human Pose Forecasting

3D-HPF from the robot's perspective involves predicting the future poses of human usere with the robot's sensors. The pose at time t can be denoted as $X_t \in \mathbb{R}^{D \times J_h}$, where $D = 3$ is the dimension of the space and $J_h = 21$ is the total number of joints in the human skeleton (X_t is the set of all joint positions in 3D). Correspondingly, 3D-HPF means predicting $X_{t+1:t+K}$ based on $X_{t-T+1:t}$, where $X_{i:j} = X_i, X_{i+1}, \ldots, X_j$, and K is the *horizon*. In line with widely-used experimental protocols [30], [31], we set $T = 10$ and $K = 4$ (roughly 400 ms) or $K = 10$ (roughly 1000 ms), two cases referred to as *short-term* and *long-term* forecasting, respectively. We used average MPJPE over the K predicted frames (average MPJPE) or MPJPE over the K^{th} predicted frame (final MPJPE) as performance metrics. The pose forecasting baselines we applied are STS-GCN [31], SiMLPe [32] and EqMotion [33]. All three were trained using MPJPE as a loss function without applying augmentation. The training was performed using the 21-joint poses obtained with the OptiTrack sensor as a ground-truth.

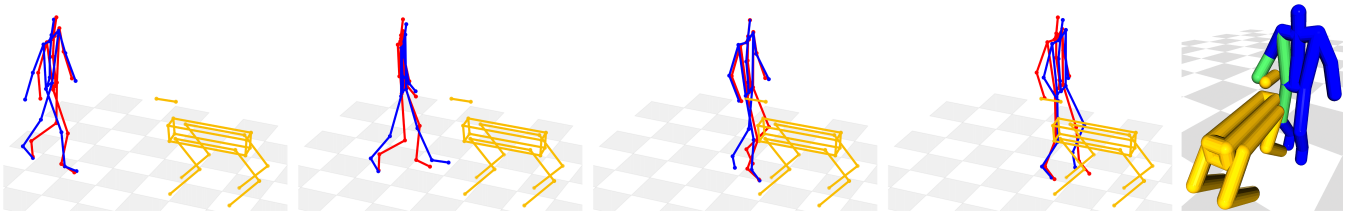
Each baseline has three variants based on different assumptions about the input data: *GT* assumes that the robot can access all ground-truth joints in the human skeleton, (*GT+R*) assumes that the robot can access only the joints visible to its sensors, and (*HRNet+D+R*) represents the 3D pose as described in Section IV-A. The GT+R and HRNet+D+R baselines deal with input sequences of incomplete poses. These cannot be processed with the forecasting baselines above and, in general, with any of the approaches in the literature. Therefore, we used a diffusion-based time series imputation model, CSDI [34], built on a cascade of transformer blocks with skip connections. This model takes a sequence of incomplete poses as input and uses them to condition the generation of a complete pose, reconstructing the position of missing joints.

Table III shows the results for the three variants of every baseline. GT achieves the best performance, while HRNet+D+R, corresponding to the most challenging task, performs the worst. EqMotion [33] delivers the best absolute results *when in the presence of GT data*: 43 mm, on average, over the 400 ms horizon, and 70 mm over the 1000 ms horizon. However, STS-GCN [31] bridges the performance gap with EqMotion when the data is noisier like, *e.g.*, in the HRNet+D+R case: the best average MPJPE is 313 mm over the 400 ms horizon and 332 mm over the 1000 ms horizon, while EqMotion achieves an MPJPE of 309 mm over the 400 ms horizon, and of 333 mm over the 1000 ms horizon.

Finally, we computed the MPJPE for each joint using the baseline with the smallest average error, *i.e.*, EqMotion [33], with GT as input (Fig. 6). We also estimated the correlation r between these errors and the average velocity of each human joint with the Pearson coefficient (r=0.79, p=1.79e-05), noticing that the faster a joint moves, the harder it is to predict its trajectory in the future.

C. Collision Prediction

CP from robot's perspective is the task of predicting whether the user and the robot will have physical contact, regardless of intent. Our focus is on the contacts or collisions caused by humans with the robot, due to the intricate challenges associated with predicting human movements, especially when the body is partially visible. Table II shows that HARPER includes four types of physical contact, all



(a) (b)

Fig. 7. Qualitative results for the pose forecasting with the 1000 ms horizon. (a) shows the human pose forecasted in blue and the ground-truth in red. At the end of the sequence, an accidental collision occurs. In (b), the collision (highlighted in green) is detected as explained in Sec. IV-C. The forecasting approach used is EqMotion [33] on the GT data.

TABLE IV

PERFORMANCE OF THE DIFFERENT COLLISION PREDICTION METHODS WITH A 1000 MS HORIZON IN TERMS OF ACCURACY, SENSITIVITY, AND SPECIFICITY SCORE. THE EVALUATION IS DIVIDED INTO THE FOUR CATEGORIES OF CONTACTS REPRESENTED IN THE HARPER DATASET.

Method	Input Type	Unintended			Touch			Kick			Punch		
		Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑	Acc.↑	Sen.↑	Spec.↑
STS-GCN [31]	GT	0.93	0.92	0.94	0.92	0.91	0.93	0.85	0.78	0.90	0.78	0.83	0.71
SiMLPe [32]	GT	0.93	0.93	0.92	0.94	0.92	0.97	0.81	0.83	0.79	0.78	0.89	0.63
EqMotion [33]	GT	0.95	0.96	0.94	0.95	0.96	0.94	0.93	0.88	0.96	0.82	0.83	0.82
Depth-based	D	0.49	0.25	0.90	0.53	0.33	0.87	0.71	0.39	0.94	0.60	0.61	0.59
EqMotion [33]	HRNet+D+R	0.74	0.61	0.90	0.89	0.88	0.91	0.80	0.73	0.86	0.67	0.66	0.69

acted out to the best of the participants' abilities. Since they differ significantly in terms of energy and limbs involved, we addressed them as different cases in the experiments. The CP process takes as input a sequence of human poses $X_{t+1:t+K}$ (see above for the notation) and a sequence of robot's poses $Y_{t+1:t+K} = (Y_{t+1}, \dots, Y_{t+K})$, where $Y_t \in \mathbb{R}^{D \times J_r}$ ($D = 3$ and J_r is the number of joints of the robot). The sequence $Y_{t+1:t+K}$ is assumed to be known because the robot plans its actions in advance. The goal is to determine whether $X_{t+1:t+K}$ and $Y_{t+1:t+K}$ indicate a *physical contact*, defined as two cylinders of radius $r = 5.0$ cm centered around the skeletal links of Spot and user are closer than a threshold $d_h = 10.0$ cm (Fig. 7b). Performance metrics used are accuracy, sensitivity, and specificity [35]. Sensitivity is $\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ (measuring how effectively the system avoids False Positives), while specificity is $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$ (measuring how effectively the system predicts True Negatives).

We started the CP experiments by feeding the methods of Sec. IV-B with the OptiTrack ground-truth data. This gave us an upper bound of the performance and showed that punches and kicks are the most difficult to predict (Table II), due to the speed and energy involved. In contrast, touch, with the lowest speed and energy showed the best performance. Then, we replaced the ground-truth data with the pose forecasts output by EqMotion [33] in its HRNet+D+R variant, completed by the CSDI [34] diffusion process (Section IV-B). Table II shows a performance decrease, but not significantly.

Finally, we evaluated a straightforward baseline called *Depth-Based* in Table IV, showing that CP requires more sophisticated approaches. The baseline is a linear regression over the future K depth frames given T previous frames. This allowed us to test whether any points are predicted to get closer than d_h. We set $T = 10$, $K = 10$, and $d_h = 10$ cm, as used in the pose forecasting baselines. As expected, the Depth-Based method performed worse than the other methods, except for kicks, where accuracy and specificity were higher, possibly because the robot's cameras capture users' legs more easily than other body parts.

V. CONCLUSIONS

We present HARPER, the first dataset focused on how quadruped robots "see" their users. The data includes 1) video and depth streams captured with the sensors of Spot, and 2) skeleton representations of users and Spot in interaction captured with an OptiTrack MoCap. The interaction scenarios were designed around specific problems (Section IV). However, the data enables one to address a much wider spectrum of problems, including, *e.g.*, proxemic behavior [36] and action recognition [37] (the list is not exhaustive). In all cases, the key novelty is that the Spot sensors can capture only part of the user's body. This leaves open the challenging problem of reconstructing the full 3D skeleton of the users while having at disposition only a partial 2D image of them. To the best of our knowledge, this is the first corpus revolving around such a problem, and therefore, we enriched the data with benchmarks, including reproducible protocols and baseline approaches. In this way, the experiments we presented can be replicated and the results of future works can be rigorously compared with those of this paper.

VI. ACKNOWLEDGEMENT

The authors would like to thank the CSI group in the School of Engineering at UofG for their support with hardware, including the SPOT and OptiTrack systems. We would also like to thank all the participants who volunteered for the

experiment and data acquisition, as well as Abdulrahman Alshememry, who assisted with part of the data annotation.

REFERENCES

[1] P. K. R. Maddikunta, Q.-V. Pham, B. Prabadevi, N. Deepa, K. Dev, T. R. Gadekallu, R. Ruby, and M. Liyanage, "Industry 5.0: A survey on enabling technologies and potential applications," *Journal of Industrial Information Integration*, vol. 26, p. 100257, 2022.

[2] S. El Zaatari, M. Marei, W. Li, and Z. Usman, "Cobot programming for collaborative industrial tasks: An overview," *Robotics and Autonomous Systems*, vol. 116, pp. 162–180, 2019.

[3] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3D human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021.

[4] X. Puig, E. Undersander, A. Szot, M. D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A. W. Clegg, M. Hlavac, S. Y. Min *et al.*, "Habitat 3.0: A co-habitat for humans, avatars and robots," *arXiv preprint arXiv:2310.13724*, 2023.

[5] P. Biswal and P. K. Mohanty, "Development of quadruped walking robots: A review," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 2017–2031, 2021.

[6] W. Merkt, V. Ivan, Y. Yang, and S. Vijayakumar, "Towards shared autonomy applications using whole-body control formulations of locomanipulation," in *Proceedings of the IEEE International Conference on Automation Science and Engineering*, 2019, pp. 1206–1211.

[7] M. Guertler, L. Tomidei, N. Sick, M. Carmichael, G. Paul, A. Wambsganss, V. H. Moreno, and S. Hussain, "When is a robot a cobot? moving beyond manufacturing and arm-based cobot manipulators," *Proceedings of the Design Society*, vol. 3, pp. 3889–3898, 2023.

[8] S. Halder, K. Afsari, E. Chiou, R. Patrick, and K. A. Hamed, "Construction inspection & monitoring with quadruped robots in future human-robot teaming: A preliminary study," *Journal of Building Engineering*, vol. 65, p. 105814, 2023.

[9] S. S. Mohammadi, N. F. Duarte, D. Dimou, Y. Wang, M. Taiana, P. Morerio, A. Dehban, P. Moreno, A. Bernardino, A. Del Bue *et al.*, "3dsgrasp: 3d shape-completion for robotic grasp," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3815–3822.

[10] F. Franzel, T. , Eiband1, and D. Lee, "Detection of collaboration and collision events during contact task execution," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2020, pp. 376–383.

[11] A. Rudenko, T. P. Kucner, C. S. Swaminathan, R. T. Chadalavada, K. O. Arras, and A. J. Lilienthal, "THÖR: Human-robot navigation data collection and accurate motion trajectories dataset," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 676–682, 2020.

[12] T. Schreiter, T. R. de Almeida, Y. Zhu, E. G. Maestro, L. Morillo-Mendez, A. Rudenko, T. P. Kucner, O. M. Mozos, M. Magnusson, L. Palmieri *et al.*, "The magni human motion dataset: Accurate, complex, multi-modal, natural, semantically-rich and contextualized," *arXiv preprint arXiv:2208.14925*, 2022.

[13] R. Martin-Martin, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[14] Z. Yan, L. Sun, T. Duckctr, and N. Bellotto, "Multisensor online transfer learning for 3d lidar-based human detection with a mobile robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 7635–7640.

[15] F. Amodeo, N. Pérez-Higueras, L. Merino, and F. Caballero, "Frog: A new people detection dataset for knee-high 2D range finders," *arXiv preprint arXiv:2306.08531*, 2023.

[16] A. Zhang, C. Eranki, C. Zhang, J.-H. Park, R. Hong, P. Kalyani, L. Kalyanaraman, A. Gamare, A. Bagad, M. Esteva *et al.*, "Towards robust robot 3D perception in urban environments: The UT campus object dataset," *arXiv preprint arXiv:2309.13549*, 2023.

[17] X. Zhang, A. Ghimire, S. Javed, J. Dias, and N. Werghi, "Robot-person tracking in uniform appearance scenarios: A new dataset and challenges," *IEEE Transactions on Human-Machine Systems*, 2023.

[18] M. Dallel, V. Havard, D. Baudry, and X. Savatier, "Inhard-industrial human action recognition dataset in the context of industrial collaborative robotics," in *Proceedings of the IEEE International Conference on Human-Machine Systems*, 2020, pp. 1–6.

[19] E. Vendrow, D. T. Le, J. Cai, and H. Rezatofighi, "JRDB-pose: A large-scale dataset for multi-person pose estimation and tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4811–4820.

[20] N. Hirose, D. Shah, A. Sridhar, and S. Levine, "Sacson: Scalable autonomous control for social navigation," *IEEE Robotics and Automation Letters*, 2023.

[21] S. Shrestha, Y. Zha, S. Banagiri, G. Gao, Y. Aloimonos, and C. Fermuller, "Natsgd: A dataset with speech, gestures, and demonstrations for robot learning in natural human-robot interaction," *arXiv preprint arXiv:2403.02274*, 2024.

[22] A. Sampieri, G. M. D. di Melendugno, A. Avogaro, F. Cunico, F. Setti, G. Skenderi, M. Cristani, and F. Galasso, "Pose forecasting in industrial human-robot collaboration," in *European Conference on Computer Vision*. Springer, 2022, pp. 51–69.

[23] H. Karnan, A. Nair, X. Xiao, G. Warnell, S. Pirk, A. Toshev, J. Hart, J. Biswas, and P. Stone, "Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 807–11 814, 2022.

[24] Y. Chen, Y. Luo, C. Yang, M. O. Yerebakan, S. Hao, N. Grimaldi, S. Li, R. Hayes, and B. Hu, "Human mobile robot interaction in the retail environment," *Scientific Data*, vol. 9, no. 1, p. 673, 2022.

[25] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[26] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[27] P.-L. Liu and C.-C. Chang, "Simple method integrating OpenPose and RGB-D camera for identifying 3D body landmark locations in various postures," *International Journal of Industrial Ergonomics*, vol. 91, p. 103354, 2022.

[28] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2012.

[29] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh, "Panoptic studio: A massively multiview system for social motion capture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3334–3342.

[30] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[31] T. Sofianos, A. Sampieri, L. Franco, and F. Galasso, "Space-time-separable graph convolutional network for pose forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 209–11 218.

[32] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda-Pineda, and F. Moreno-Noguer, "Back to MLP: A simple baseline for human motion prediction," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4809–4819.

[33] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang, "EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1410–1420.

[34] Y. Tashiro, J. Song, Y. Song, and S. Ermon, "CSDI: Conditional score-based diffusion models for probabilistic time series imputation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 804–24 816, 2021.

[35] Y. J. Heo, D. Kim, W. Lee, H. Kim, J. Park, and W. K. Chung, "Collision detection for industrial collaborative robots: A deep learning approach," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 740–746, 2019.

[36] J. Mumm and B. Mutlu, "Human-robot proxemics: physical and psychological distancing in human-robot interaction," in *Proceedings of the International Conference on Human-Robot Interaction*, 2011, pp. 331–338.

[37] A. Chrungoo, S. Manimaran, and B. Ravindran, "Activity recognition for natural human robot interaction," in *International Conference on Social Robotics*, 2014, pp. 84–94.