data

MDPI

Data Descriptor

UNIPD-BPE: Synchronized RGB-D and Inertial Data for Multimodal Body Pose Estimation and Tracking

Mattia Guidolin [1,*], Emanuele Menegatti [2] and Monica Reggiani [1]

[1] Department of Management and Engineering, University of Padova, 35122 Padova, Italy
[2] Department of Information Engineering, University of Padova, 35122 Padova, Italy
* Correspondence: mattia.guidolin@unipd.it

Abstract: The ability to estimate human motion without requiring any external on-body sensor or marker is of paramount importance in a variety of fields, ranging from human–robot interaction, Industry 4.0, surveillance, and telerehabilitation. The recent development of portable, low-cost RGB-D cameras pushed forward the accuracy of markerless motion capture systems. However, despite the widespread use of such sensors, a dataset including complex scenes with multiple interacting people, recorded with a calibrated network of RGB-D cameras and an external system for assessing the pose estimation accuracy, is still missing. This paper presents the University of Padova Body Pose Estimation dataset (*UNIPD-BPE*), an extensive dataset for multi-sensor body pose estimation containing both single-person and multi-person sequences with up to 4 interacting people. A network with 5 Microsoft Azure Kinect RGB-D cameras is exploited to record synchronized high-definition RGB and depth data of the scene from multiple viewpoints, as well as to estimate the subjects' poses using the Azure Kinect Body Tracking SDK. Simultaneously, full-body Xsens MVN Awinda inertial suits allow obtaining accurate poses and anatomical joint angles, while also providing raw data from the 17 IMUs required by each suit. This dataset aims to push forward the development and validation of multi-camera markerless body pose estimation and tracking algorithms, as well as multimodal approaches focused on merging visual and inertial data.

Dataset: https://doi.org/10.17605/OSF.IO/YJ9Q4

Dataset License: CC0

Keywords: markerless motion capture; inertial motion capture; body pose estimation; RGB-D; IMU

Citation: Guidolin, M.; Menegatti, E.; Reggiani, M. *UNIPD-BPE*: Synchronized RGB-D and Inertial Data for Multimodal Body Pose Estimation and Tracking. *Data* **2022**, *7*, 79. https://doi.org/10.3390/data7060079

Academic Editors: Filipe Meneses and Joaquín Torres-Sospedra

Received: 6 May 2022
Accepted: 2 June 2022
Published: 9 June 2022

1. Summary

Human motion analysis commonly relies on optoelectronic systems that track small retroreflective markers attached to the subject's body. These systems, although extremely accurate, are characterized by high costs and complex setups. Such characteristics constrain their use to specific applications that are confined in a dedicated laboratory (e.g., clinical analyses or animation industry motion capture). However, real-time human pose estimation could benefit a variety of fields, ranging from human–robot interaction, Industry 4.0, autonomous driving, surveillance, and telerehabilitation. In such contexts, the deployment of optoelectronic systems is usually not feasible, and markerless analyses are a promising tool to address this issue.

Markerless body pose estimation (BPE) has been a topic of intensive research for decades in the computer vision community. Despite the improvements achieved in the latest years thanks to the advances enabled by data-driven approaches [1–4], the accurate assessment of human motion without relying on any sensor or marker attached to the body is still an open challenge. Limited fields of view of the cameras and occlusions due to the environment, but also self-occlusions of the human body, limit the accuracy of such systems.

One possible solution to reduce the impact of the aforementioned limitations consists of exploiting a distributed camera network to acquire data of the same scene from multiple viewpoints. By fusing the partial information obtained from each camera, it is possible to reduce the effect of occlusions and, at the same time, increase the overall system's accuracy.

In recent years, the development of portable and easy-to-use low-cost 3D cameras (e.g., the Microsoft Kinect, Microsoft Corp., Redmond, WA, USA) has further pushed the interest in markerless BPE [5–8]. The main advantage of these devices is the possibility to retrieve real-time synchronized RGB and depth data of the scene, up to 30 Hz. However, despite the widespread use of such sensors and the variety of available human motion datasets, only a small number of public datasets include RGB-D data and even less offer multiple calibrated RGB-D views. In fact, to the best of the authors' knowledge, a comprehensive dataset including complex scenes with multiple people, RGB, and depth data from a significant calibrated RGB-D camera network, together with ground truth body poses for all the recorded sequences, is still missing. All the most used markerless motion capture datasets (either focused on BPE or on action recognition) lack at least one of the aforementioned features.

HumanEva [9] is one of the first and most used datasets recorded for benchmarking markerless human pose estimation algorithms. The dataset includes 6 actions of daily living (ADLs) recorded by 4 different actors using 4 grayscale cameras, 3 RGB cameras, and a marker-based optoelectronic system as a ground truth. No information on the depth of the scene is available, and each sequence only involves a single person.

Human3.6M [10], on the other hand, offers depth data of the scene using a single Time-of-Flight (ToF) sensor. Also in this case, ground truth poses are acquired via marker-based motion capture, while visual data are recorded using 4 RGB cameras. The dataset includes a predefined set of 16 ADLs performed by 11 actors. Even in this case, no interactions among subjects are available.

Our previous work, the *IAS-Lab Action Dataset* [11], was one of the first to include RGB-D sensors in the acquisition setup. This dataset consists of 15 ADLs performed by 12 people. RGB and depth data are provided, as well as the persons' body poses estimated by exploiting a markerless BPE algorithm. However, data are recorded using a single Kinect v1 camera. Additionally, no ground truth poses are available, nor are sequences with multiple people.

Berkeley MHAD [12] is one of the first datasets to include accelerometers in the acquisition setup. Eleven ADLs performed by 12 actors are recorded using marker-based motion capture, 12 RGB cameras, 2 Kinect v1 cameras, and 6 accelerometers. However, similarly to the previous works, the focus is on estimating single persons' actions, and no interactions are taken into account.

TUM Shelf [13] is among the most used datasets for benchmarking markerless BPE algorithms. It includes 5 RGB cameras to record a group of 4 people disassembling a shelf. Severe occlusions and unbounded motion of the persons are the main challenges of this dataset. However, since no other sensing devices are involved, the dataset offers only sparse manually annotated poses as a ground truth. The same authors also released the *TUM Campus* dataset [13]. The particularity of this dataset is that it is captured outdoors. The recorded scenes depict 3 people interacting on campus grounds. Similar to *TUM Shelf*, only 3 RGB cameras are used. Thus, the same limitations apply.

CMU Panoptic [14] is a large-scale dataset that includes 480 VGA cameras, 31 HD cameras, and 10 Kinect v2 cameras. A variety of actions (including both single-person and multi-person activities) are recorded inside a custom-built dome accommodating all the hardware. However, since vision is the only modality used to retrieve data, the recorded poses are only computed via triangulation based on a 2D BPE algorithm that runs on each camera, without any external ground truth.

Another public dataset including multiple depth views is the *NTU RGB+D* dataset [15]. Forty subjects were recorded performing a set of 60 actions that include ADLs, mutual activities, and health-related movements. The sensors used to extract the persons' poses

were 3 Kinect v2 cameras. However, since the focus is on the validation of action recognition algorithms, no ground truth poses are provided, but only labels indicating the type of actions being performed.

All the aforementioned datasets mainly focused on vision, including markerless and marker-based motion capture. *UTD-MHAD* [16], on the other hand, introduced the use of one inertial measurement unit (IMU), in conjunction with a Kinect v1 camera. Eight subjects were individually recorded while performing a set of 27 predefined actions ranging from sports, hand gestures, ADLs, and training exercises. Similarly to the previous work, however, the focus is on action recognition. Thus, the available ground truth is limited to manually annotated labels describing the actions being performed.

Total Capture [17] is a widely used dataset and one of the first to introduce the usage of a full-body inertial suit consisting of 13 IMUs, alongside 8 RGB cameras and marker-based motion capture. Five subjects are recorded performing a set of 5 actions selected from range of motion activities, walking, acting, running, and freestyle. Ground truth poses are computed via marker-based motion capture. However, the dataset does not include interactions between subjects, and no information on the depth of the scene is available.

AndyData-lab [18], similarly to the previous work, includes data from marker-based motion capture, a full-body inertial suit, 2 RGB cameras, while also adding finger pressure sensors. Since this work focuses on human motion analysis in industrial settings, 13 subjects are recorded while performing 6 industrial tasks, including screwing at different heights and manipulating loads. As in the previous work, neither interactions among subjects nor information on the depth of the scene are available.

Finally, *Human4D* [19] includes data from an optoelectronic system and 4 Intel RealSense RGB-D cameras (Intel Corp., Santa Clara, CA, USA). Four actors are recorded, both individually and in pairs, while performing a set of 14 single-person ADLs and 5 two-person activities in a professional motion capture studio. Ground truth poses are collected via marker-based motion capture, and both RGB and depth recordings of the scene are available. However, during the recordings, all actors needed to wear a full-body black suit to accommodate the body markers required by the optoelectronic system during the entire trial. These artificial clothes can hinder the performance of RGB-based markerless BPE algorithms, potentially decreasing their accuracy, since they do not constitute a realistic scenario.

This paper presents the University of Padova Body Pose Estimation dataset (*UNIPD-BPE*), an extensive dataset for multi-sensor BPE containing a large number of single-person and multi-person sequences with up to 4 people interacting. Full-body poses, as well as raw data from each sensor, are recorded both by means of a calibrated network with 5 RGB-D cameras (i.e., Microsoft Azure Kinect, Microsoft Corp., Redmond, WA, USA) and by exploiting up to 2 highly accurate full-body inertial suits (i.e., Xsens MVN Awinda, Xsens Technologies, Enschede, Netherlands). All recorded data are publicly available under the Creative Commons CC0 license at https://doi.org/10.17605/OSF.IO/YJ9Q4.

The Azure Kinect is the latest RGB-D camera developed by Microsoft, with improved performance compared to the previous model (Kinect v2). As demonstrated in [20], the Azure Kinect standard deviation is reduced by more than 50 % with respect to the Kinect v2, while also achieving a depth estimation error lower than 11 mm. For these reasons, the Azure Kinect is a promising device with a wide range of uses including object recognition, people tracking and detection, and human–computer interaction. This dataset is the first to include high-definition RGB, depth, and BPE data from 5 calibrated Azure Kinect cameras. Videos and point clouds are recorded both at a resolution of 1920 $\times$ 1080 pixels @ 30 Hz and 640 $\times$ 576 pixels @ 30 Hz (native resolution of the depth sensor). Moreover, all subjects' body poses are estimated via markerless motion capture by exploiting the Azure Kinect Body Tracking SDK [21], offering baseline data to develop and benchmark different BPE and tracking algorithms. The high number of cameras allows us to assess the impact of different camera network configurations on the accuracy achieved

by markerless BPE algorithms, while the high-resolution recordings allow us to quantify how different image resolutions can impact a specific algorithm.

The *UNIPD-BPE* dataset also contains full-body inertial motion capture data, collected by up to 2 Xsens MVN Awinda suits. Each suit consists of 17 MTw Awinda trackers, including a 3-axis gyroscope, a 3-axis accelerometer, and a 3-axis magnetometer. As demonstrated in [22], these sensors are extremely accurate for inertial BPE. Each tracker has a dynamic accuracy of 0.75 °RMS for roll and pitch, and 1.5 ° RMS for the heading estimation, constituting a flexible and reliable tool for capturing human motion [23]. The proposed dataset includes both the raw data from each tracker, and detailed data describing each subject's body kinematics, computed by exploiting the MVN Analyze software. Such software combines the data of all motion trackers with a biomechanical model of the human, allowing to obtain an accurate and drift-free estimate of the body pose [24]. The hardware/software combination used on this work allowed to record raw IMU data (estimated orientations, angular velocities, linear accelerations, magnetic fields) for all the trackers required by each suit @ 60 Hz, as well as 3D positions, orientations, velocities, accelerations of the 23 segments defining the Xsens biomechanical model, anatomical joint angles of 22 joints plus 6 additional joint angles targeted to ergonomic analyses, and the body center of mass location throughout all the sequences.

No optoelectronic data are included in this dataset because the required markers attached to the body are highly reflective, resulting in a strong distortion in the Kinects' depth and, consequently, in a poor estimation of the body pose. While it is possible to properly synchronize the two systems to avoid interference, this solution still degrades the Azure Kinect's performance. Therefore, to ensure maximum accuracy of the recorded markerless data, we chose to employ an inertial motion capture system in place of the optoelectronic one. The software used for the estimation of the body poses (Xsens MVN Analyze), coupled with the chosen hardware (Xsens MVN Awinda), allows us to obtain an accuracy comparable to state-of-the-art optoelectronic systems, as demonstrated in [24].

All the cameras and inertial suits used in this work are hardware synchronized, while the relative poses of each camera with respect to the inertial reference frame are calibrated before each sequence to ensure maximum overlap of the two sensing systems outputs. The proposed setup allowed to record synchronized 3D poses of the persons on the scene both via Xsens' inverse kinematics algorithm (inertial motion capture) and by exploiting the Azure Kinect Body tracking SDK (markerless motion capture), simultaneously. The additional raw data (RGB, depth, camera network configuration) allow the user to assess the performance of any custom markerless motion capture algorithm (based on RGB, depth, or both). Further analyses can be progressed by varying the number of cameras being used and/or their resolution and frame rate. Moreover, raw angular velocities, linear accelerations, magnetic fields, and orientations from each IMU allow to develop and test multimodal BPE approaches focused on merging visual and inertial data. Finally, the precise body dimensions of each subject are provided. They include body height, weight, and segment lengths measured before the beginning of a recording session. They were used to scale the Xsens biomechanical model, and also constitute a ground truth for assessing the markerless BPE accuracy on estimating each subject's body dimensions.

The recorded sequences include 15 participants performing a set of 12 ADLs (e.g., walking, sitting, and jogging). The actions were chosen to present different challenges to BPE algorithms, including different movement speeds, self-occlusions, and complex body poses. Moreover, multi-person sequences, with up to 4 people performing a set of 7 different actions, are provided. Such sequences offer challenging scenarios where multiple self-occluded persons move and interact in a restricted space. They allow assessing the accuracy of multi-person tracking algorithms, focused on maintaining frame-by-frame consistent IDs of each detected person. To this end, the proposed dataset has already been used to validate our previous work, describing a real-time open-source framework for multi-camera multi-person tracking [25]. A total of 13.3 h (over 1,400,000 frames) of RGB, depth, and markerless BPE data from 5 RGB-D cameras are present in the dataset, while the

inertial motion capture system allowed to record 3 h (over 600,000 frames) of human poses, corresponding to 51.2 h of raw IMU data from all the sensors used in each suit.

The remainder of the paper is organized as follows. Section 2 describes the content and organization of the dataset. Section 3 presents the methods applied for data collection and describes how to replicate the setup used for the acquisitions. Finally, Section 4 concludes the article, addressing possible uses of the dataset in different research fields.

2. Data Description

The *UNIPD-BPE* dataset contains: (1) high definition videos and point clouds from each RGB-D camera, (2) positions, orientations, and confidences of the body joints estimated via markerless motion capture, (3) raw IMU data from each tracker used in the inertial suits, (4) full-body kinematics and anatomical joint angles obtained via inertial motion capture. Table 1 summarizes all available data, while Sections 2.1 and 2.2 describe in detail the recordings obtained by each RGB-D camera and by the inertial suits, respectively.

Table 1. Content of the *UNIPD-BPE* dataset.

Source	Typology	Details
Calibration	Transforms	Relative poses among cameras
Camera network	Video	1920 × 1080 pixels @ 30 Hz (native resolution)
	Video	640 × 576 pixels @ 30 Hz (reprojection on the depth)
	Depth	640 × 576 pixels @ 30 Hz (native resolution)
	Depth	1920 × 1080 pixels @ 30 Hz (reprojection on the RGB)
	BPE	3D positions, orientations, confidences of 32 joints
Inertial suit	IMU data	Orientations and raw IMU data @ 60 Hz
	BPE	3D positions, orientations, velocities, accelerations of 23 segments
	Joint angles	Anatomical joint angles of 22 joints
	Center of mass	3D position of the person's center of mass

2.1. Microsoft Azure Kinect

The camera network used in this work consists of 5 Azure Kinect cameras (labeled *k01*, *k02*, *k03*, *k04*, *k05*). Details on the spatial configuration of the sensors can be found in Section 3.1. Each camera includes a 1 MP Time-of-Flight depth sensor, a 12 MP CMOS rolling shutter RGB sensor, a 6-DoF IMU, and a 7-microphone circular array. A factory calibration process provides intrinsic and extrinsic calibrations of the sensors.

The *UNIPD-BPE* dataset contains the following data, captured from each of the 5 cameras:

- video recordings (1920 × 1080 pixels @ 30 Hz (native resolution) and 640 × 576 pixels @ 30 Hz (reprojected on the depth));
- depth recordings (1920 × 1080 pixels @ 30 Hz (reprojected on the RGB) and 640 × 576 pixels @ 30 Hz (native resolution));
- 3D positions, orientations, confidences of 32 body joints defined in the Azure Kinect Body Tracking SDK model (Appendix A.1).

Data are recorded at the maximum frame rate allowed by the system. The video resolution was chosen to provide high-definition captures, while also maintaining the dataset size as manageable.

2.2. Xsens MVN Awinda

The Xsens MVN Awinda suit used in this work consists of 17 MTw Awinda trackers placed on the head, chest, shoulders, upper arms, forearms, hands, pelvis, thighs, shanks, and feet. Each tracker includes a 3-axis gyroscope, a 3-axis accelerometer, a 3-axis magnetometer, and has a dynamic accuracy of 0.75 ° RMS for roll and pitch, and 1.5 ° RMS for the heading estimation [23].

Before each sequence, the body model used to estimate the motion was specifically scaled to each participant's characteristics. All subjects' body dimensions and general information (sex, age, weight, height) are annotated in dedicated files included in the dataset.

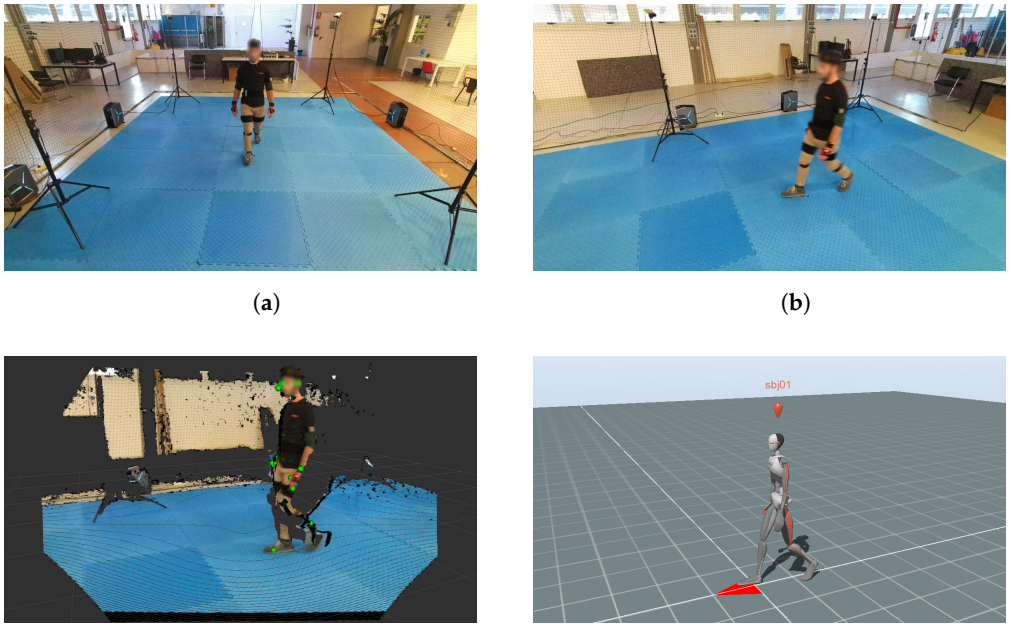
The *UNIPD-BPE* dataset contains the following data, captured for up to 2 subjects simultaneously:

- orientations, angular velocities, linear accelerations, magnetic fields of 17 MTw Awinda trackers @ 60 Hz;
- 3D positions, orientations, linear and angular velocities, linear and angular accelerations of 23 body segments defined in the Xsens MVN Analyze model (Appendix A.2);
- anatomical joint angles (flexion/extension, abduction/adduction, internal/external rotation) of 22 body joints, plus 6 additional joint angles calculated for ergonomic analyses;
- 3D position of the body center of mass.

Data are recorded at 60 Hz (maximum frame rate allowed by the system) using the Xsens MVN Analyze software (version 2021.0.1).

2.3. Dataset Structure

A total of 13.3 h of RGB, depth, and markerless BPE data are present in the dataset, corresponding to over 1,400,000 frames obtained from a calibrated network with 5 RGB-D cameras. The inertial suits, on the other hand, allowed to record 3 h of inertial motion capture data, corresponding to a total of over 600,000 frames recorded by each of the 17 IMUs used by every suit. Figure 1 shows an example frame of the available data recorded during a walking sequence.



(**a**) (**b**)

(**c**) (**d**)

Figure 1. Sample data during a walking sequence: (**a**) RGB frame from *k01*, (**b**) RGB frame from *k02*, (**c**) depth and markerless pose estimation from *k02*, (**d**) inertial pose estimation from MVN Analyze.

The dataset is divided into 2 folders: *single_person*, containing all the sequences where a single subject is recorded, and *multi_person*, containing all the sequences with multiple subjects. Sections 2.3.1 and 2.3.2 explain the organization of the dataset for single-person and multi-person sequences, respectively.

2.3.1. Single-Person Sequences

To make the data easily accessible, the single-person sequences are organized as follows. The *single_person* folder contains the data recorded from 15 subjects performing

the 12 actions described in Table 2, with 4 repetitions each. Thus, it contains 15 folders, named *sbj<xx>*, where *<xx>* indicates the subject's ID. Each *sbj<xx>* folder contains the data recorded by the cameras, the inertial motion capture data, and a *yaml* file (named *sbj<xx>_info.yaml*) including the subject's ID, sex, age, weight, and the body dimensions used for inertial BPE.

The recorded data are stored in 6 subfolders: 5 folders containing the camera network data, named after the convention *k<yy>*, where *<yy>* indicates the camera's ID, and an additional folder containing the inertial suit data, named *xsens*. Each *k<yy>* folder contains 4 repetitions for each of the 12 actions, resulting in 48 files (one per sequence), named following the convention *sbj<xx>_<action_name><zz>.bag*, where *<zz>* indicates the recorded repetition. For each recorded sequence, the *xsens* folder contains 3 sets of files, named *sbj<xx>_<action_name><zz>.(mvnx | bvh | c3d)*. Each single-person action has an average duration of approximately 13 s. The complete list of recorded actions is reported in Table 2.

Table 2. List of actions performed during single-person sequences.

Index	Action Name	Description
0	t_pose	T-pose to be used for calibration purposes
1	n_pose	N-pose to be used for calibration purposes
2	walk	Walking at self-selected speed
3	squat	Squatting
4	bend	Bending down
5	sit	Sitting on a chair
6	jog	Jogging in place
7	jump	Jumping in place
8	cross_arms	Crossing arms
9	point	Pointing to different directions
10	wave	Waving hands
11	throw	Pretending to throw an object

The *bag* file format indicates a bag file, commonly used in ROS [26] (Robot Operating System) to store ROS message data. This format was chosen since it allows to store and distribute heterogeneous streams of synchronized data. By using *bag* files, it is also possible to play the recorded data simulating a real-time acquisition. Additionally, the content of a bag file can be exported in different formats by exploiting one of the many open-source tools developed by the ROS community. ROS bags, in fact, play an important role in ROS, and a variety of tools have been written to allow storage, processing, analysis, and visualization of the stored data.

All the *bag* files in this dataset contain RGB captures and depth point clouds from each camera, information on the camera network calibration, positions, orientations, and confidences of each participant's joints estimated via markerless motion capture.

The *mvnx* extension (MVN Open XML format) refers to Xsens' proprietary format. It is a human-readable XML format that can be imported into various software programs, including MATLAB and Microsoft Excel. This format contains information on sensor data, segment kinematics, and joint angles, as well as the subject's body dimensions. The *bvh* format (BioVision Hierarchical data) embeds captured motion data in ASCII format and is typically used in animation applications. It requires a hierarchical structure, such that only relative joint angles can be exported into this file format. Finally, *c3d* (Coordinate 3D) is a format used in optical systems and only contains 3D point coordinates. Therefore, the stored data are limited to the bony landmarks calculated from the estimated virtual marker set.

2.3.2. Multi-Person Sequences

Multi-person sequences include the 7 actions described in Table 3, repeated with 2, 3, and 4 people simultaneously on the scene. The only exception is the action labeled *eight*, where the persons are walking forming an eight, which required the presence of 4 people. The actions were selected to challenge different aspects typical of markerless BPE. As a

result, there are different actions where multiple people are in close proximity, with partial and/or full occlusions, and with people exiting and reentering the scene.

The *multi_person* folder contains data recorded from all the sequences that include multiple subjects. It contains 3 folders, named <*xx*>*people*, where *xx* indicates the number of subjects present in each sequence. Similarly to the *single_person* sequences, each folder contains a *yaml* file (named <*xx*>*people_info.yaml*), the data recorded by the cameras, and the inertial motion capture data. In this case, however, the *yaml* file stores the IDs of all the subjects on the scene. At the beginning of each sequence, in fact, all participants stand in front of the master camera (*k01*). To allow for the correct assignment of each subject's body dimensions, the *yaml* file contains the IDs of all the participants ordered from left to right, as seen by the master camera. The body dimensions can be retrieved by accessing the corresponding *sbj*<*zz*>*_info.yaml* file, where <*zz*> indicates the ID assigned for the single-person sequences.

The recorded data are stored in 6 subfolders: 5 folders containing the camera network data, named after the convention *k*<*yy*>, where <*yy*> indicates the camera's ID, and an additional folder containing the inertial suit data, named *xsens*. Each *k*<*yy*> folder can contain 6 or 7 files (depending on the number of people interacting), named following the convention <*xx*>*people_*<*action_name*>*.bag*. Each *bag* file includes the same typology of data recorded for single-person sequences. In this case, however, no repetitions are available, since the focus is on providing relevant data for the assessment of multi-person skeletal tracking, and being each sequence the summation of the actions performed by multiple people simultaneously. For each recorded sequence, the *xsens* folder contains 6 sets of files, named <*xx*>*people_*<*action_name*>*_sbj*<*yy*>*.(mvnx | bvh | c3d)*, and <*xx*>*people_*<*action_name*>*_sbj*<*zz*>*.(mvnx | bvh | c3d)*, being inertial data available for up to two subjects simultaneously. Each multi-person sequence has an average duration of approximately 27.5 s. The complete list of recorded actions is reported in Table 3.

Table 3. List of actions performed during multi-person sequences.

Index	Action Name	Description
0	static	Static poses to be used for calibration purposes
1	free_static	Free movements while remaining in the same place
2	free_dynamic	Free movements while changing positions
3	circle	Walk in a circle
4	cross	Switch positions while walking in a circle
5	in_out	Enter and exit from the cameras field of view
6	eight	Walk forming an eight

3. Methods

This section describes the experimental setup, the methodology used, and the characteristics of the participants. All data were recorded in a laboratory environment, to allow accurate calibration of the RGB-D camera network and proper alignment of markerless and inertial motion capture.

3.1. Experimental Setup

The experimental setup (Figure 2) includes 5 RGB-D cameras and up to 2 full-body inertial suits. Each camera is connected to a dedicated desktop PC, while the IMUs communicate wirelessly to a receiver (Awinda station) connected to a PC that acts as a master. All PCs are connected to the same local network. Software time synchronization among PCs is obtained using the NTP protocol, whereas sensors synchronization is performed by exploiting the onboard hardware offered by the two sensing systems. More details on hardware synchronization are reported in Section 3.4.

Figure 2. Experimental setup used for the acquisition of the *UNIPD-BPE* dataset. The 5 RGB-D cameras are highlighted in red, while the inertial suits' master receiver is highlighted in green.

The cameras are placed at a height of 2 m, in the configuration shown in Figure 3. They are approximately placed in a circle with a radius of 3 m. This allows to cover an area of approximately 4×4 m^2 where most cameras have full visibility of the persons in the scene. The pose of each camera with respect to a common global reference frame was estimated prior to the recordings using an internally developed calibration algorithm.

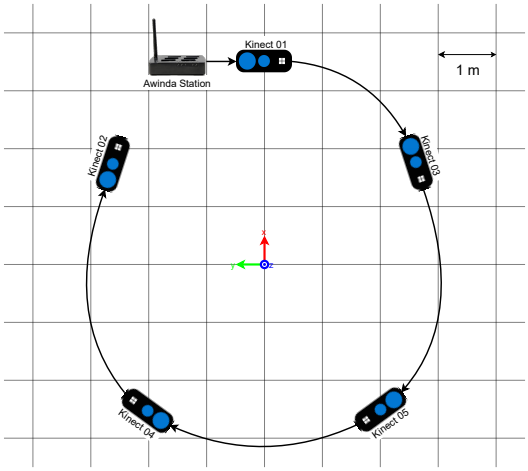


Figure 3. Spatial organization of the sensors used during the acquisitions. The axes define the global reference frame of both the camera network and the inertial suits, while the arrow lines indicate the cabled connections required for the hardware synchronization.

The recorded data were acquired using the Microsoft Azure Kinect ROS Driver [27] under ROS Noetic (Ubuntu 20.04 LTS). The driver allows to publish each person's detected poses as standard ROS messages. However, it does not include information on the detection confidence. For this reason, the driver has been customized to also include information on the estimated joints' confidence in the messages. The mapping between the confidence levels assigned by the markerless BPE algorithm and the corresponding values stored in the messages is reported in Table 4.

Table 4. Azure Kinect Body Tracking SDK confidence mapping.

Confidence Value	Description
0	The joint is out of range (too far from depth camera)
1	The joint is not observed (likely due to occlusion), predicted joint pose
2	Medium confidence in joint pose
3	High confidence in joint pose

3.2. Participants

A total of 15 participants were recruited for data collection (11 men, 4 women). The average age was 23.7 ± 2.7 years (min: 21 years, max: 29 years), the average weight 65.8 ± 12.7 kg (min: 48 kg, max: 92 kg), and the average height 1.75 ± 0.11 m (min: 1.57 m, max: 1.98 m). All participants gave written informed consent before data collection. Table 5 shows in detail each participant's characteristics. The ID assigned to each subject is the same for all experiments. Additionally, in the sequences where multiple people interact, a person's IDs corresponds to the one used in their individual sequence.

Table 5. Characteristics of the participants.

ID	Sex	Age [Years]	Weight [kg]	Height [m]
01	m	22	72	1.75
02	m	22	55	1.70
03	m	21	60	1.65
04	m	22	92	1.90
05	m	21	84	1.81
06	m	22	63	1.72
07	m	22	75	1.98
08	m	22	78	1.88
09	f	23	48	1.65
10	m	28	68	1.75
11	f	23	52	1.57
12	m	27	72	1.75
13	f	29	56	1.70
14	f	24	48	1.58
15	m	27	64	1.79

3.3. Acquisition Protocol

Before each session, the 17 IMU trackers were placed on the participants' head, chest, shoulders, upper arms, forearms, hands, pelvis, thighs, shanks, and feet, following the Xsens protocol. The body model used for the motion estimation was then specifically scaled to each participant's characteristics. MVN Analyze was configured in the *Single level* scenario, since all tasks were executed on a fixed-level ground, without elevation changes. The system was then calibrated with the *N-pose and walk procedure*, and the world frame aligned with the camera network's global reference frame. To maximize the overlap between markerless and inertial BPE, the suit's world frame was realigned to the cameras' global frame before each sequence recording.

For single-person sequences, the participants were asked to perform one of the actions described in Table 2 while facing a different cardinal direction in each repetition. Except for walking, where the start and end positions were fixed, the participants had maximum freedom regarding how to perform the actions.

Multi-person sequences include the actions reported in Table 3, each performed with a varying number of subjects ranging from 2 to 4. Inertial data are recorded for up to 2 subjects per sequence simultaneously. Sensors placement and software configuration are the same as for single-person sequences.

3.4. Time Synchronization

This section describes the synchronization procedure followed for the acquisition of the dataset. In fact, since the dataset includes information from heterogeneous sources and a distributed camera network, all sensors must be time-synchronized.

Each Azure Kinect camera includes two synchronization ports (Sync in and Sync out). In this work, all cameras are synchronized through a daisy-chain configuration (Figure 3). To avoid interference among infrared projectors, the captures were offset from each other by 160 µs, as suggested in Microsoft's documentation. Therefore, the maximum delay between two cameras in the network is equal to 640 µs, which is negligible with respect to the maximum frame rate of 30 Hz (<2% of the δt between two consecutive frames).

The Xsens MTw Awinda station includes 4 synchronization ports (2 Sync in and 2 Sync out). In this work, the Awinda station was used as a master device to synchronize inertial and markerless motion capture. A custom cable was built to allow the Awinda station to send synchronization pulses to the master Kinect (*k01* in Figure 3). The chosen configuration allowed Xsens to properly synchronize the Kinect cameras by sending a triggering signal when a recording session was started. Thus, the *Start recording* command in MVN Analyze also triggered the streaming of the camera network data (RGB frames, depth frames, and markerless body tracking).

4. Conclusions

This paper presented *UNIPD-BPE*, an extensive dataset for single- and multi-person body pose estimation. Single-person sequences include 15 participants performing a set of 12 activities of daily living, while multi-person sequences include 7 actions with 2 to 4 persons interacting in a confined area.

The dataset includes 13.3 h of high definition RGB and depth data (corresponding to over 1,400,000 frames) recorded by a calibrated RGB-D camera network of 5 synchronized Azure Kinect cameras, as well as each subject's full-body poses estimated using the Azure Kinect Body Tracking SDK. This allows to assess the impact of exploiting different numbers and/or configurations of cameras on the accuracy achieved by markerless BPE algorithms. The provided markerless body poses can be used as a baseline, while the raw recorded data (RGB, depth, and camera network configuration) allow the dataset user to assess the performance and accuracy of any custom markerless BPE algorithm (based on RGB, depth, or both).

Furthermore, 3 h of inertial motion capture poses were obtained by exploiting highly accurate Xsens MVN Awinda full-body suits, corresponding to a total of over 600,000 frames recorded by each of the 17 IMUs used by every suit. All sensors are hardware-synchronized, with the Xsens MVN Awinda system acting as a master to trigger the acquisitions. The relative poses of each camera with respect to the inertial reference frame are accurately calibrated before each sequence to ensure the best overlap of the two systems' outputs. This allows inertial motion capture estimates to be used to further investigate the accuracy of different markerless BPE algorithms. Since the raw IMU data are also available, the dataset can also be used to develop novel sensor fusion algorithms, aiming at improving the performance of both markerless motion capture, by increasing the achievable accuracy, and inertial motion capture, by limiting possible drifting phenomena.

The multi-person sequences offer challenging scenarios where multiple partially occluded persons move and interact in a restricted space. This allows us to investigate the performance of multi-person tracking algorithms, both regarding the accuracy of the pose estimation in cluttered environments, and the ability to maintain frame-by-frame consistent IDs of each detected person in the scene.

The proposed dataset also presents some limitations. Due to the hardware used in the RGB-D camera network, no optoelectronic data could be included. This would offer an additional source of information, also allowing us to assess the accuracy of inertial motion capture. Moreover, the main focus of the dataset is on the validation of different BPE algorithms. As a result, all recordings were acquired in a laboratory environment,

with a limited amount of background clutter, to ensure the best overlap between markerless and inertial body poses.

To conclude, the *UNIPD-BPE* dataset aims to push forward the development of markerless BPE and tracking algorithms, enabling a variety of applications where unobtrusive accurate knowledge of human motion is of paramount importance. The dataset in fact includes data both for single-person RGB- and depth-based human motion estimation, for multi-person BPE and tracking, and for visual and inertial sensor fusion. The high-definition videos and point clouds, recorded by 5 calibrated and synchronized RGB-D cameras, allow simulating a variety of different scenarios (e.g., a pure RGB camera network, a pure depth camera network, an uncalibrated camera network, etc.). Finally, the included markerless and inertial body poses are useful for the development and testing of different multimodal sensor fusion and people tracking algorithms, without the necessity of expensive hardware and bulky acquisition setups.

Author Contributions: Conceptualization, M.G., E.M., and M.R.; data curation, M.G.; writing—original draft preparation, M.G.; writing—review and editing, E.M. and M.R.; supervision, E.M. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: Part of this work was supported by MIUR (Italian Minister for Education) under the initiative "Departments of Excellence" (Law 232/2016).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in OSF at https://doi.org/10.17605/OSF.IO/YJ9Q4 under the CC0 license.

Acknowledgments: The authors would like to thank Daria Battini and the Ergo-Lab group for providing one of the inertial suits, as well as for the precious support during the acquisition of this dataset.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Body Joint Definitions and Hierarchy

Appendix A.1. Microsoft Azure Kinect Body Tracking

The Microsoft Body Tracking SDK allows to process Azure Kinect captures to generate body tracking results. A skeleton includes 32 joints. Each connection (bone) links the parent joint with a child joint. As demonstrated in [28], the mean joint estimation error has an average value of 8 mm and a standard deviation of 6 mm in static conditions. Table A1 lists the joint connections. Additional information can be found in [29].

Table A1. Azure Kinect Body Tracking joint definitions and hierarchy.

ID	Joint Name	Parent Joint	ID	Joint Name	Parent Joint
0	PELVIS	-	16	HANDTIP_RIGHT	HAND_RIGHT
1	SPINE_NAVAL	PELVIS	17	THUMB_RIGHT	WRIST_RIGHT
2	SPINE_CHEST	SPINE_NAVAL	18	HIP_LEFT	PELVIS
3	NECK	SPINE_CHEST	19	KNEE_LEFT	HIP_LEFT
4	CLAVICLE_LEFT	SPINE_CHEST	20	ANKLE_LEFT	KNEE_LEFT
5	SHOULDER_LEFT	CLAVICLE_LEFT	21	FOOT_LEFT	ANKLE_LEFT
6	ELBOW_LEFT	SHOULDER_LEFT	22	HIP_RIGHT	PELVIS
7	WRIST_LEFT	ELBOW_LEFT	23	KNEE_RIGHT	HIP_RIGHT
8	HAND_LEFT	WRIST_LEFT	24	ANKLE_RIGHT	KNEE_RIGHT
9	HANDTIP_LEFT	HAND_LEFT	25	FOOT_RIGHT	ANKLE_RIGHT
10	THUMB_LEFT	WRIST_LEFT	26	HEAD	NECK
11	CLAVICLE_RIGHT	SPINE_CHEST	27	NOSE	HEAD
12	SHOULDER_RIGHT	CLAVICLE_RIGHT	28	EYE_LEFT	HEAD
13	ELBOW_RIGHT	SHOULDER_RIGHT	29	EAR_LEFT	HEAD
14	WRIST_RIGHT	ELBOW_RIGHT	30	EYE_RIGHT	HEAD
15	HAND_RIGHT	WRIST_RIGHT	31	EAR_RIGHT	HEAD

Appendix A.2. Xsens MVN Analyze

The Xsens MVN Analyze software features a scalable biomechanical model and offers real-time 3D animation, graphs, and data streaming. A skeleton includes 23 segments connected by 22 joints. As demonstrated in [24], the inertial body poses show a RMSE lower than 5° for the estimation of the anatomical joint angles in the sagittal plane. Table A2 contains the list of the body segments defining a skeleton, its joints, and the trackers used to estimate human motion. Additional information can be found in [30].

Table A2. Xsens MVN joint definitions and hierarchy.

ID	Segment Label	Tracker	Joint
0	Pelvis	Pelvis	jL5S1
1	L5	T8	jL4L3
2	L3	Head	jL1T12
3	T12	RightShoulder	jT9T8
4	T8	RightUpperArm	jT1C7
5	Neck	RightForeArm	jC1Head
6	Head	RightHand	jRightC7Shoulder
7	Right Shoulder	LeftShoulder	jRightShoulder
8	Right Upper Arm	LeftUpperArm	jRightElbow
9	Right Forearm	LeftForeArm	jRightWrist
10	Right Hand	LeftHand	jLeftC7Shoulder
11	Left Shoulder	RightUpperLeg	jLeftShoulder
12	Left Upper Arm	RightLowerLeg	jLeftElbow
13	Left Forearm	RightFoot	jLeftWrist
14	Left Hand	LeftUpperLeg	jRightHip
15	Right Upper Leg	LeftLowerLeg	jRightKnee
16	Right Lower Leg	LeftFoot	jRightAnkle
17	Right Foot	-	jRightBallFoot
18	Right Toe	-	jLeftHip
19	Left Upper Leg	-	jLeftKnee
20	Left Lower Leg	-	jLeftAnkle
21	Left Foot	-	jLeftBallFoot
22	Left Toe	-	-

References

1. Toshev, A.; Szegedy, C. DeepPose: Human Pose Estimation via Deep Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1653–1660.
2. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
3. Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; Sun, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7103–7112.
4. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef] [PubMed]
5. Shotton, J.; Fitzgibbon, A.; Cook, M.; Sharp, T.; Finocchio, M.; Moore, R.; Kipman, A.; Blake, A. Real-Time Human Pose Recognition in Parts from Single Depth Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 1297–1304.
6. Alabbasi, H.; Gradinaru, A.; Moldoveanu, F.; Moldoveanu, A. Human Motion Tracking & Evaluation using Kinect V2 Sensor. In Proceedings of the 2015 E-Health and Bioengineering Conference (EHB), Iasi, Romania, 19–21 November 2015; pp. 1–4.
7. Kim, J.; Lee, I.; Kim, J.; Lee, S. Implementation of an Omnidirectional Human Motion Capture System Using Multiple Kinect Sensors. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2015**, *98*, 2004–2008. [CrossRef]
8. Bilesan, A.; Behzadipour, S.; Tsujita, T.; Komizunai, S.; Konno, A. Markerless Human Motion Tracking Using Microsoft Kinect SDK and Inverse Kinematics. In Proceedings of the 2019 12th Asian Control Conference (ASCC), Kitakyushu, Japan, 9–12 June 2019; pp. 504–509.
9. Sigal, L.; Balan, A.O.; Black, M.J. HumanEva: Synchronized Video and Motion Capture Dataset and Baseline Algorithm for Evaluation of Articulated Human Motion. *Int. J. Comput. Vis.* **2010**, *87*, 4–27. [CrossRef]
10. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339. [CrossRef] [PubMed]
11. Munaro, M.; Ballin, G.; Michieletto, S.; Menegatti, E. 3D flow estimation for human action recognition from colored point clouds. *Biol. Inspired Cogn. Archit.* **2013**, *5*, 42–51. [CrossRef]

12. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Berkeley MHAD: A Comprehensive Multimodal Human Action Database. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Clearwater Beach, FL, USA, 15–17 January 2013; pp. 53–60.
13. Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilic, S. 3D Pictorial Structures for Multiple Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1669–1676.
14. Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; Sheikh, Y. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3334–3342.
15. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
16. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 168–172.
17. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In Proceedings of the 28th British Machine Vision Conference, London, UK, 4–7 September 2017; pp. 1–13.
18. Maurice, P.; Malaisé, A.; Amiot, C.; Paris, N.; Richard, G.J.; Rochel, O.; Ivaldi, S. Human movement and ergonomics: An industry-oriented dataset for collaborative robotics. *Int. J. Robot. Res.* **2019**, *38*, 1529–1537. [CrossRef]
19. Chatzitofis, A.; Saroglou, L.; Boutis, P.; Drakoulis, P.; Zioulis, N.; Subramanyam, S.; Kevelham, B.; Charbonnier, C.; Cesar, P.; Zarpalas, D.; et al. HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media. *IEEE Access* **2020**, *8*, 176241–176262. [CrossRef]
20. Tölgyessy, M.; Dekan, M.; Chovanec, L.; Hubinskỳ, P. Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2. *Sensors* **2021**, *21*, 413. [CrossRef] [PubMed]
21. Azure Kinect Body Tracking SDK Documentation. Available online: https://microsoft.github.io/Azure-Kinect-Body-Tracking/release/1.1.x/index.html (accessed on 27 May 2022).
22. Guidolin, M.; Petrea, R.A.B.; Oboe, R.; Reggiani, M.; Menegatti, E.; Tagliapietra, L. On the Accuracy of IMUs for Human Motion Tracking: A Comparative Evaluation. In Proceedings of the 2021 IEEE International Conference on Mechatronics (ICM), Kashiwa, Japan, 7–9 March 2021; pp. 1–6.
23. Paulich, M.; Schepers, M.; Rudigkeit, N.; Bellusci, G. *Xsens MTw Awinda: Miniature Wireless Inertial-Magnetic Motion Tracker for Highly Accurate 3D Kinematic Applications*; Xsens: Enschede, The Netherlands, 2018.
24. Schepers, M.; Giuberti, M.; Bellusci, G. *Xsens MVN: Consistent Tracking of Human Motion Using Inertial Sensing*; Xsens: Enschede, The Netherlands, 2018; pp. 1–8.
25. Guidolin, M.; Tagliapietra, L.; Menegatti, E.; Reggiani, M.Hi-ROS: Open-Source Multi-Camera Sensor Fusion for Real-Time People Tracking. *Comp. Vis. Image Understand.* 2022, *submitted*.
26. Quigley, M.; Conley, K.; Gerkey, B.; Faust, J.; Foote, T.; Leibs, J.; Wheeler, R.; Ng, A.Y. ROS: An open-source Robot Operating System. In *Proceedings of the ICRA Workshop on Open Source Software*, Kobe, Japan, 12–17 May 2009; Volume 3, p. 5.
27. Azure Kinect ROS Driver. Available online: https://github.com/microsoft/Azure_Kinect_ROS_Driver (accessed on 27 May 2022).
28. Romeo, L.; Marani, R.; Malosio, M.; Perri, A.G.; D'Orazio, T. Performance Analysis of Body Tracking with the Microsoft Azure Kinect. In Proceedings of the 2021 29th Mediterranean Conference on Control and Automation (MED), Online, 22–25 June 2021; pp. 572–577.
29. Azure Kinect Body Tracking Joints. Available online: https://docs.microsoft.com/azure/kinect-dk/body-joints (accessed on 27 May 2022).
30. Xsens MVN User Manual. Available online: https://www.xsens.com/hubfs/Downloads/usermanual/MVN_User_Manual.pdf (accessed on 27 May 2022).