Machine Vision and Applications (2021) 32:4
https://doi.org/10.1007/s00138-020-01131-z

ORIGINAL PAPER

A system for the generation of in-car human body pose datasets

João Borges[1] · Sandro Queirós[1,2,3] · Bruno Oliveira[1] · Helena Torres[1] · Nelson Rodrigues[1] · Victor Coelho[4] · Johannes Pallauf[5] · José Henrique Brito[6] · José Mendes[1] · Jaime C. Fonseca[1]

Received: 3 December 2019 / Revised: 18 June 2020 / Accepted: 15 September 2020

Abstract

With the advent of autonomous vehicles, detection of the occupants' posture is crucial to tackle the needs of infotainment interaction or passive safety systems. Generative approaches have been recently proposed for human body pose in-car detection, but this type of approaches requires a large training dataset for a feasible accuracy. This requirement poses a difficulty, given the substantial time required to annotate such large amount of data. In the in-car scenario, this requirement risk increases even further, since a robust human body pose ground-truth system capable of working in it is needed but inexistent. Currently, the gold standard for human body pose capture is based on optical systems, requiring up to 39 visible markers for a plug-in gait model, which in this case are not feasible given the occlusions inside the car. Other solutions, such as inertial suits, also have limitations linked to magnetic sensitivity and global positioning drift. In this paper, a system for the generation of images for human body pose detection in an in-car environment is proposed. To this end, we propose to smartly combine inertial and optical systems to suppress their individual limitations: By combining the global positioning of 3 visible head markers provided by the optical system with the inertial suit's relative human body pose, we obtain an occlusion-ready, drift-free full-body global positioning system. This system is then spatially and temporally calibrated with a time-of-flight sensor, automatically obtaining in-car image data with (multi-person) pose annotations. Besides quantifying the inertial suit inherent sensitivity and accuracy, the feasibility of the overall system for human body pose capture in the in-car scenario was demonstrated. Our results quantify the errors associated with the inertial suit, pinpoint some sources of the system's uncertainty and propose how to minimize some of them. Finally, we demonstrate the feasibility of using system generated data (which was made publicly available), independently or mixed with two publicly available generic datasets (not in-car), to train 2 machine learning algorithms, demonstrating the improvement in their algorithmic accuracy for the in-car scenario.

Keywords Automotive applications · Supervised learning · Dataset generation · Human body pose estimation

1 Introduction

With current development and deployment of advanced driver-assistance systems (ADAS), higher levels of car automation will be available. With them, the human factor inside the car will also change. The need for occupants' advanced detection systems becomes even more relevant, opening the possibility to understand how passengers behave or how they interact with the car itself. These new possibilities can interface with specific in-car use cases, such as passive safety or comfort, and to tackle these needs, the system must be able to monitor the occupant's body pose. Several approaches have been described in the literature,

Bruno Oliveira and Helena Torres have contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00138-020-01131-z) contains supplementary material, which is available to authorized users.

✉ João Borges
jpbsilva@algoritmi.uminho.pt

[1] Algoritmi Center, University of Minho, Guimarães, Portugal

[2] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal

[3] ICVS/3B's - PT Government Associate Laboratory, Braga, Guimarães, Portugal

[4] Bosch, Braga, Portugal

[5] Bosch, Abstatt, Germany

[6] 2Ai - Polytechnical Institute of Cávado and Ave, Barcelos, Portugal

Published online: 08 October 2020

from layout mapping [37] to human body pose detection [2,5,9–12,16,25,26,30,31,36,38], both in RGB and depth images. From the different classes of methods, discriminative ones (machine learning [ML]) present the best results for an in-car contextualization, due to their generalization and low computational cost. This method comes with an important requirement—the need for a large and generic dataset for training. Real datasets are the primary choice to be used to train every detector. This decision comes from the fact that a real dataset is recorded with real sensors data, which will later be used to infer the pose. This datum not only gives the required information for pose inferring, but also gives information about the sensor's inherent noise model, allowing to generate a training dataset with similar characteristics to the one where the human pose must be detected. Usually, creation of a real dataset consists in recording the body pose with a motion capture system, while synchronously acquiring an image using an image sensor (e.g., RGB or time of flight [ToF]). The pose is then referenced from one system to the other to be correctly projected into the visual frame. This entire procedure is very time-consuming and requires a lot of manual interaction, hampering the task of creating a large and generic dataset. Despite some datasets related to human body pose being publicly available (with RGB and depth sensors), these motion capture datasets do not focus on the in-car scenario, mostly due to the fact that standard vision-based motion capture systems are not able to function properly in this scenario. With this in mind, the main goal of this work is to present a user-friendly system to generate human body pose datasets in an in-car environment, with the ground-truth system itself being also evaluated. These datasets may be used to train human body pose detection algorithms for an in-car scenario. This system combines a ToF sensor, inertial suit and the Vicon system (Vicon, Oxford, UK). The ToF sensor provides image data; the inertial suit provides a relative human body pose, and the Vicon system provides the global positioning inside the vehicle using 3 head markers only. Inertial-based limitations such as global positioning drift are suppressed with the optical system head tracking, with this joint being chosen given its high visibility and reduced tissue displacement. Optical-based limitations, such as occlusions, are suppressed with the inertial system relative human body pose tracking. The output data are comprised of human body poses given with respect to the camera's coordinate system in 2D and 3D for the ToF's amplitude/depth image frame and tridimensional point-cloud, respectively.

The rest of this paper is organized as follows. In Sect. 2, the related work on body pose datasets generation and motion capture systems is summarized. In Sect. 3, the overall system methodology is presented. The system's evaluation and its potential interest are presented in Sect. 4 and discussed in Sect. 5. The main conclusions are given in Sect. 6.

2 Related work

The generation of real data for ML algorithms is an important task in a high variety of areas. The acquisition of motion data in the in-car scenario is a difficult task given the lack of motion capture systems that can reliably work in it. Although there are accurate motion capture systems available, they are not focused in heavily occluded scenarios. The alternative can involve electromagnetic or inertial-based systems, solving the occlusion problem, but adding new limitations (i.e., magnetic distortion sensitivity for electromagnetic, and global drift for inertial).

2.1 Body pose datasets

To satisfy the development needs of human body pose detection algorithms, several datasets were generated and made available to the research community. The CMU Graphics Lab Motion Capture Database [8] is by far one of the most extensive datasets of publicly available motion capture data that focuses in human skeleton data and RGB image frames. Unfortunately, the dataset was generated based on the Vicon system, where the scenes are focused in non-car scenarios. Human3.6M [13] is also based on Vicon motion capture, with added depth and body scan image sensors. Temporal synchronization is achieved through hardware and software triggering. Another Vicon-based dataset is the HumanEva [32], where the main difference is the fact that RGB image frames were captured using an external video source, bringing the need to apply a synchronization method during data capture. Other non-car datasets, such as ITOP [12] and NTU RGB+D [29], are based on depth image sensors and body pose ground-truth data from Shotton et al. [30]. Manually annotated datasets such as the HMDB [18] are based on publicly available RGB images, where the labels are human made and prone to error. The biggest advantage on such dataset is its size and variability, giving the possibility of increasing algorithmic performance on human action recognition. Within the in-car scenario, Borghi et al. [4] used the Pandora dataset for the POSEidon head and shoulder pose estimator. The Pandora dataset is generated in a laboratory environment with minimal occlusion; this is achieved with different subjects performing similar driving behaviors while seating on a chair. Head and shoulder orientation were captured through inertial sensors. Alternatively, hybrid datasets offer an opportunity to utilize large synthetic datasets in conjunction with real ones, aiding models' generalization. Chen et al. [7] created an automated toolchain to synthesize RGB images from 3D poses. Regarding depth features, Martinez et al. [21] combined multi-person synthetic depth data with real sensor backgrounds.

2.2 Motion capture systems

2.2.1 Electromagnetic

Systems such as Polhemus (Polhemus, Vermont, USA) are electromagnetic based, and despite being highly accurate, they suffer from electromagnetic disturbances from external sources, such as metals or electronic devices. Mitobe et al. [23] were able to track finger movement of pianists with a spatial resolution of 3.8 μm, although it also showed the complexity of the setup and the need for placing a single wired sensor for each tracked joint. Other wireless solutions are available through the Polhemus Liberty latus [19,22], allowing full-body motion capture, but still suffering from the same sources of uncertainty.

2.2.2 Optical

Optical-based motion capture systems are used across several R&D fields [8,31,32] and can be separated in two types: (1) marker-based with high accuracy, illumination immunity and high setup time; (2) markerless with fast setup time but with reduced accuracy and higher sensitivity to light conditions. Although marker-based systems, such as the Vicon, are the gold standard for motion capture, they suffer from the fact that they need to have the markers in line of sight to guarantee accurate tracking. Rahmatalla et al. [27] circumvented the occlusion problem by adding virtual (calculated) markers while tracking a seated operator in-lab. However, these researchers only focused on the lateral pelvis' markers. Considering a markerless approach, Joo et al. [14,15] implemented a multi-view system, comprised by 480 synchronized video streams and 5 Kinects, allowing for the extraction of multi-person 3D anatomical landmarks. However, this system still requires a confined and controlled space that does not resemble the in-car scenario. Considering the technical limitations of the most robust motion capture systems, it is not feasible to use them alone for the in-car environment.

2.2.3 Inertial and magnetic measurement units (IMMUs)

IMMUs-based systems are an alternative to optical systems since they do not need the subject to be in line of sight. Theoretically, they are able to infer body segment orientation as well as joints' positions, although they are prone to errors caused by drift or magnetic sensitivity. Another issue for the estimation of full-body kinematics is related to the need of a biomechanical model and its initial calibration. Current biomechanical models are proposed by the Internal Society of Biomechanics (ISB) [35]. For example, Xsens (Xsens, Enschede, the Netherlands) uses a modified version with 23 segments and 22 joints [28]. During calibration, the subject needs to stand in a calibration posture, known as the N-Pose or T-Pose, which assumes that all segments are aligned and the coordinate systems of all joints are parallel to one another. This initial assumption adds a systematic error that offsets the segments' orientations and joints' positions. In its thesis, Orozco [24] compared MVN BIOMECH Awinda from Xsens against Vicon Nexus through the study of gait kinematics with calibration postures that deviate from the standard N-Pose, showing that the error introduced could be considered as a shift in joint angle values, while the shape was not affected. It was also possible to understand that this error could be corrected with the information of the true body posture captured during the calibration procedure. Two correction approaches were proposed (orientation correction and planar angle correction), giving the possibility to achieve an initial calibration procedure for human subjects that are not able to attain a N-Pose.

3 System overview

With the purpose of developing a system capable of generating ToF images with associated human body pose ground-truth for the in-car scenario, several systems are required:

- a ToF sensor for image capture (C being the position of the camera's optical center);
- an inertial suit (namely a MVN BIOMECH Awinda) for relative human body pose ground-truth (A being the head joint, and J each one of the remaining body joints);
- a global object positioning system, such as the Vicon system (W being the Vicon's global coordinate system, and O the subject's head object tracked by it);
- a car testbed.

As Fig. 1 illustrates, there is a certain complexity with the added systems. With it, there is a need to spatially and temporally align the data, to correctly project the human body pose information into the ToF camera's perspective.

To satisfy these requirements, our system implements the pipeline illustrated in Fig. 2.

3.1 Recording

Recording is done for all relevant systems: Image data are recorded from a ToF sensor; relative body pose data are recorded from the MVN Awinda inertial suit; and global positioning data are recorded with Vicon Nexus through the creation of a virtual head object for the human subject. Note that this object was selected for two main reasons: (1) low soft tissue-related errors for both Vicon and inertial suit's markers/sensors; (2) best joint visibility for the Vicon system in an in-car environment, generating more valid data.

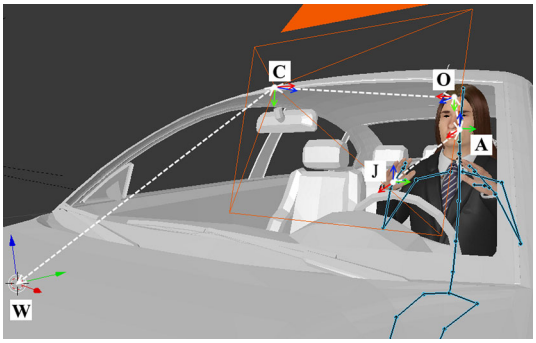


Fig. 1 3D representation of coordinate systems when recording a real dataset. W: Vicon global coordinate system; C: ToF optical center; O: subject's head object tracked by the Vicon system; A: inertial suit head joint; J: inertial suit joints

3.2 Alignment

In order to achieve proper temporal synchronization, a temporal alignment between systems is required (Sect. 3.2.3). Considering the ToF camera as master to avoid image blur from interpolations, the other two systems have to be synchronized [34] and interpolated to each master timestamp (Sects. 3.2.1 and 3.2.2).

Considering the goal of generating a human body pose ground-truth projected into the ToF coordinate system, a spatial calibration is required (Sect. 3.2.6). In this regard, there is the need to align the Vicon system with the ToF camera (Sect. 3.2.4), as well as the Awinda suit with the Vicon system (Sect. 3.2.5).

3.2.1 Vicon-to-ToF temporal alignment

To temporally align the ToF camera with the Vicon system, a marker was used as a pendulum while being visible by both systems. Two sets of points' coordinates were thus obtained in the camera's perspective: (1) one in the amplitude frame $amp_{x,y}$, since each pixel information reflects the object's light intensity projected on the camera's XY plane, the pendulum position (a light emitter) in the image frame, p_C, is detected by finding the pixel with maximum intensity (Eq. 1); (2) the projection of the Vicon's marker 3D coordinates, $\tilde{P}^W_{vm}$, in the camera's image frame, p^C_{vm}, through Eq. 2.

$$p_C = \underset{x,y}{\text{argmax}}(amp_{x,y}) \tag{1}$$

$$\begin{aligned} \tilde{P}^C_{vm} &= T^C_W \cdot \tilde{P}^W_{vm} \\ \therefore \tilde{p}^C_{vm} &= \tilde{M} \cdot \tilde{P}^C_{vm} \end{aligned} \tag{2}$$

Equation 2 relies on the relationship between image and space coordinates (Eq. 3) and two calibrations: (1) the camera's pre-calibrated intrinsic matrix $\tilde{M}$ (Eq. 4, using the strategy in [39]), with f_x, f_y being the focal length in pixels, c_x, c_y being the optical center in pixels, and s being the camera axes skew angle; (2) the Vicon-to-ToF spatial calibration T^C_W (Sect. 3.2.4).

$$\begin{aligned} &\tilde{\boldsymbol{p}} = \tilde{\boldsymbol{M}} \cdot \tilde{\boldsymbol{P}} \\ &\tilde{\boldsymbol{P}} = [X, Y, Z, 1]^\intercal, \boldsymbol{P} = [X, Y, Z]^\intercal \\ &\tilde{\boldsymbol{p}} = [U, V, S]^\intercal, \quad \boldsymbol{p} = (x, y), \quad x = \frac{U}{S}, \quad y = \frac{V}{S} \end{aligned} \tag{3}$$

$$\tilde{M} = \begin{bmatrix} f_x & s & c_x & 0 \\ 0 & f_y & c_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{4}$$

Both virtual markers must exist for each timestamp, generating an oscillating wave in the ToF x-axis, as shown in Fig. 3. This projection allows to estimate the time delay $t(0)^{\text{ToF}}_{\text{Vicon}}$ between both systems through the maximum of the cross-correlation of the two discrete-time sequences, p_C and p^C_{vm}.

3.2.2 Awinda-to-Vicon temporal alignment

To determine the temporal mismatch between the Vicon system and the Awinda suit, both systems head joint quaternions

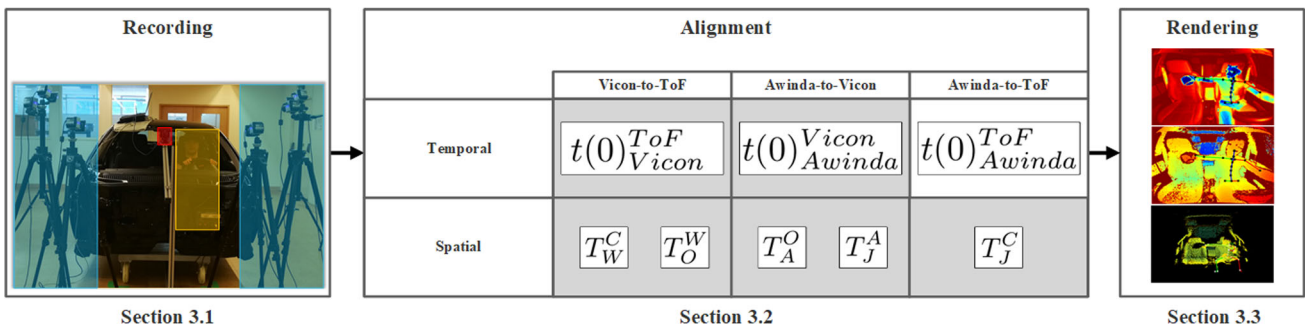


Fig. 2 Overview of the system pipeline. Recording: blue highlight represents the Vicon system; orange highlight represents the MVN BIOMECH Awinda system; red highlight represents the ToF camera. Alignment: gray highlight corresponds to the developed algorithms, with the other variables being determined by concatenation of the algorithms' outputs

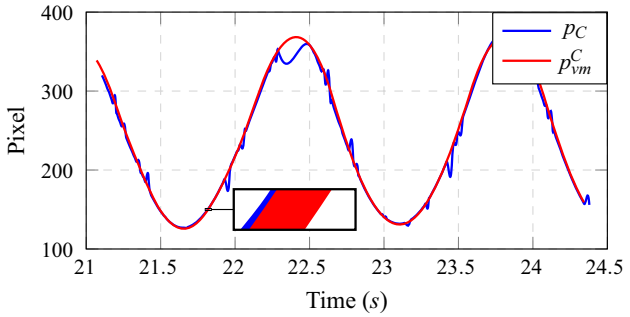


Fig. 3 Projection of the Vicon's marker in the ToF x-axis (red) and the marker's x-axis coordinate in the image frame (blue), giving the delay between systems $t(0)^{\text{ToF}}_{\text{Vicon}}$

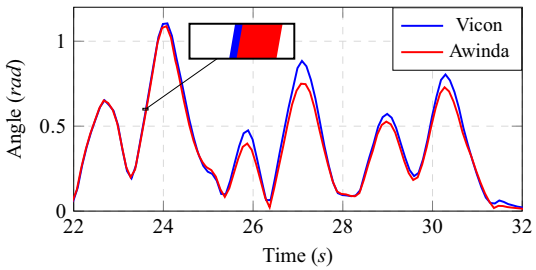


Fig. 4 Axis-angle temporal representation of both Vicon and Awinda head rotation, giving the delay between systems $t(0)^{\text{Vicon}}_{\text{Awinda}}$

were converted and represented in axis-angle θ_t (Eq. 5), while removing the offset rotation of the first frame through the quaternion conjugate $(q_{t(0)})^c$. This provides the head orientation as a relative rotation for each system, as shown in Fig. 4. Through the maximum of the cross-correlation of both signals, we get $t(0)^{\text{Vicon}}_{\text{Awinda}}$, which allows to synchronize the Awinda suit with the Vicon system [34].

$$\begin{aligned} \boldsymbol{q_t} &= q_{t,r} + q_{t,i}\boldsymbol{i} + q_{t,j}\boldsymbol{j} + q_{t,k}\boldsymbol{k} \\ \boldsymbol{qr_t} &= \boldsymbol{q_t} \cdot (\boldsymbol{q_{t(0)}})^c \\ \theta_t &= 2 \times \arccos(\boldsymbol{qr_{t,r}}) \end{aligned} \tag{5}$$

3.2.3 System temporal alignment

With each previous temporal calibration, all recorded systems are temporally aligned with each other (Eq. 6).

$$\begin{aligned} t(0)^{\text{ToF}}_{\text{Awinda}} &= t(0)^{\text{ToF}}_{\text{Vicon}} + t(0)^{\text{Vicon}}_{\text{Awinda}} \\ t_O &= t_O + t(0)^{\text{ToF}}_{\text{Vicon}} \\ t_A &= t_A + t(0)^{\text{ToF}}_{\text{Awinda}} \\ t_C &= t_C \end{aligned} \tag{6}$$

Each calibrated temporal mismatch, $t(0)^{\text{ToF}}_{\text{Vicon}}$ and $t(0)^{\text{ToF}}_{\text{Awinda}}$, serves as a time offset to the timestamp of each system's sample: t_C being the ToF, t_O the Vicon and t_A the Awinda.

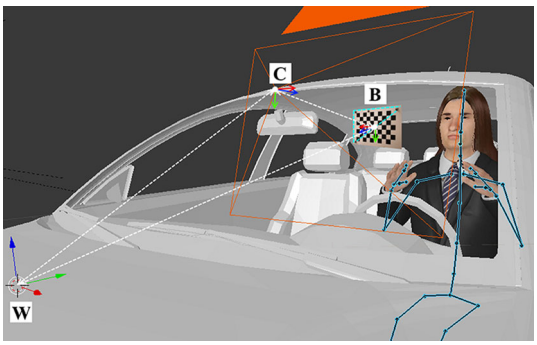


Fig. 5 3D representation of the coordinate systems when calibrating T^C_W. W: Vicon global coordinate system; C: ToF optical center; B: checkerboard object plane surface

3.2.4 Vicon-to-ToF spatial alignment

Vicon head to Vicon world (T^W_O) The T^W_O transformation is automatically recorded through the Vicon system and requires the setup of a head object with a fixed pattern of markers.

Vicon world to ToF camera (T^C_W) As Fig. 5 illustrates, for the calibration of T^C_W, a checkerboard pattern (B) was used together with four Vicon markers placed in its surface. Since the camera has been previously calibrated (Eq. 4 [39]), we are able to determine the checkerboard's plane and therefore the position of the four markers in the camera's world coordinates (T^C_B). Together with their known position in the Vicon system, T^W_B, one is able to calculate the transformation between systems through a least square minimization problem (see Algorithm S1 in Appendix and equation therein).

3.2.5 Awinda-to-Vicon spatial alignment

Awinda joints to Awinda head (T^A_J) The MVN Awinda body pose is tracked inside the MVN global positioning system. This is possible through specific proprietary algorithms that estimate the body gait movement. To minimize sources of error from the MVN Awinda system, we focused in using the relative body pose, removing the global positioning information and its associated errors. To convert the global body pose into a relative one, a joint of reference (root joint) needs to be defined; in this case, the head joint was chosen given its direct relation with the Vicon's head object. One is thus able to obtain the relative position/orientation of each joint wrt. the head, T^A_J.

Awinda head to Vicon head (T^O_A) Considering the body pose as being relative to the head (root joint), the alignment between the MVN Awinda head and the Vicon head needs

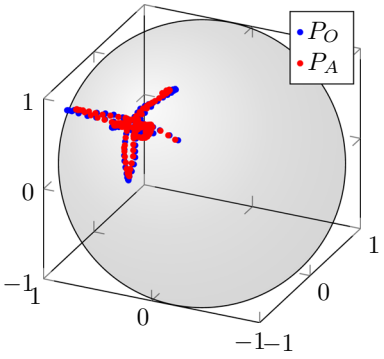


Fig. 6 Awinda head to Vicon head spatial alignment, T_A^O. Axes represent mesh points P_O, P_A translation limits related to vector P

to be determined (orientation alignment only). To determine the T_A^O transformation, each recording procedure is initiated by a small recording of head rotations in each axis (i.e., flexion/extension, lateral flexion and rotation). This information is then used to generate a mesh of points for the Vicon head object, P_O, and the Awinda head joint, P_A. Each mesh point P_O, P_A corresponds to a vector, $P = [0, 0, 1]^\intercal$, being rotated in each timestamp t_O, t_A with the head rotation R_O^W, R_A at the specific timestamp. Through the finite iterative closest point (FICP) algorithm [17], it is possible to find the transformation, T_A^O, that best aligns both point-clouds that represent the head joint from both systems, as shown in Fig. 6.

3.2.6 System spatial alignment

With each previous spatial calibration, the human body pose wrt. the ToF camera's perspective, T_J^C, can be calculated using Eq. 7.

$$T_J^C = T_W^C \cdot T_O^W \cdot T_A^O \cdot T_J^A \tag{7}$$

T_J^C is the combined transformation matrix (i.e., "$\cdot$" being matrix multiplication) that spatially aligns the inertial suit with the camera world, with T_J^A being the transformation that maps each joints' position/orientation wrt. to the head in the inertial suit system, T_A^O the transformation that maps the head joint from the inertial suit system to the Vicon one, T_O^W the position/orientation of the head object within the Vicon's world coordinate system, T_W^C the transformation that maps the Vicon's system to the ToF sensor coordinate system.

3.3 Rendering

After data alignment, the system starts rendering the dataset information, illustrated in Fig. 7:

- features: (a) an amplitude frame, (b) a depth frame and (c) a point-cloud (i.e., exported in BIN and PLY format);
- labels: 2D and 3D body pose of each human model (i.e., generated and exported in JSON format according to the CMU format [8]).

3.3.1 Amplitude frame

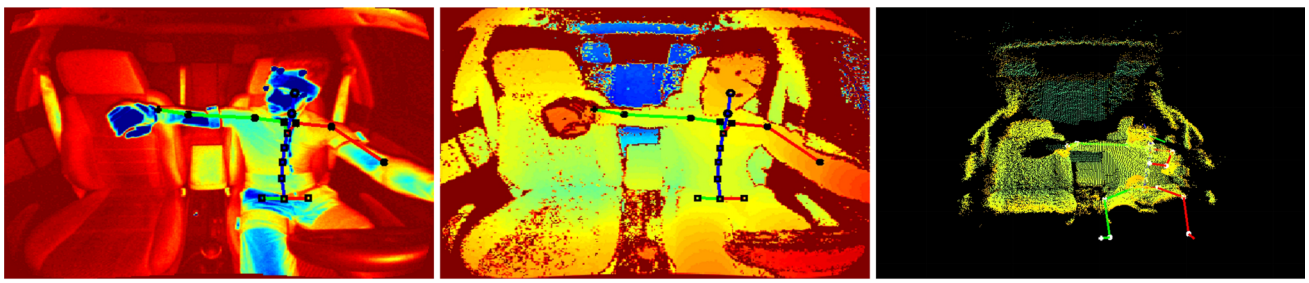
The amplitude frame $amp_{x,y}$, $x = 1, \ldots, X$, $y = 1, \ldots, Y$, where X is the camera's horizontal resolution and Y its vertical resolution, is rendered with the information recorded from the ToF camera. Each pixel information reflects the intensity from the object projected on the pixel in the camera's XY plane.

3.3.2 Depth frame

The depth frame $depth_{x,y}$, $x = 1, \ldots, X$, $y = 1, \ldots, Y$, is rendered with the information recorded from the ToF camera, with each pixel representing the distance from the object to the projected pixel in the camera's XY plane.

3.3.3 Point-cloud

The point-cloud has the Cartesian coordinates $[X, Y, Z]$ of the voxel that was projected in each pixel, pcx_i, pcy_i, pcz_i, where $_i$ indexes all point-cloud voxels.



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

Fig. 7 Rendered frames: **a** amplitude, **b** depth, and **c** point-cloud. Images **a**, **b**, **c** are represented in color for better visualization. Ground-truth: (black dots) 2D pose, (white dots) 3D pose

3.3.4 Ground-truth

The ground-truth information consists in exporting the 2D and 3D pose information for the human (Eq. 3). Both types of ground-truth consist in the same pose information for the human, that is, a structure comprised of all joints' pixel, p_J^C (2D), or voxel, P_J^C (3D), positions.

4 Experiments and results

4.1 System

4.1.1 Dataset generation

To evaluate the generation of real datasets, two experiments were established: (1) tightly controlled evaluation for the same sequence of driving actions, i.e., touching steering wheel, middle and passenger compartments, back seat and waving, for both outside and inside the car (TE1 and TE2, respectively), while avoiding any type of collision between sensors/straps and external materials (example videos of both TE1 and TE2 are provided in supplementary material); (2) free movement evaluation for three in-car seated positions (FE1, FE2 and FE3 corresponding to front passenger, driver and back passenger, respectively), while allowing the subject to interact freely with the scenery. For both evaluation procedures, we extracted the head quaternion behavior for both systems (Figs. 8 and 9), associated with a full output from specific timestamps, as shown in Figs. 10 and 11.

For the 1st evaluation procedure, the timestamps represent the same actions for inside and outside the car (Fig. 10).

In Fig. 10, it is possible to see that extreme head rotations increase the body pose error due to the head alignment. It is also possible to observe that magnetic distortions are not significant, where the same actions inside the car can project a body pose similar or better than outside of it.

For the 2nd evaluation procedure, the dataset timestamps represent the best and worst alignments (Fig. 11). From the analysis, it is possible to pinpoint the main sources of error that contribute to a sub-optimal body pose joint projection:

- Sources of error identified in the inertial suit evaluation (see Appendix B);
- Absence of projection, due to occlusion of the Vicon's head object;
- High projection error for joints further away from the head joint, due to bad alignment between head objects (indicated with red highlight in Fig. 9);
- High projection error for the lower body part, due to sensors and straps' movement as a result of collisions with seat, car door or steering wheel.

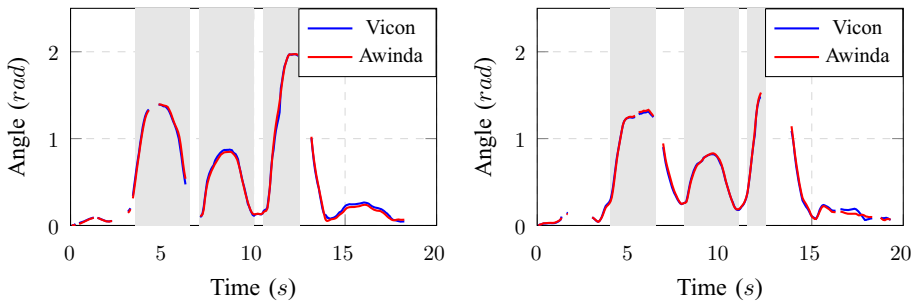


Fig. 8 Axis-angle representation between Vicon's head object and Awinda's head segment: **a** TE1; **b** TE2. Gray regions highlight the 1st, 2nd and 3rd head maximum rotations for each simulated action

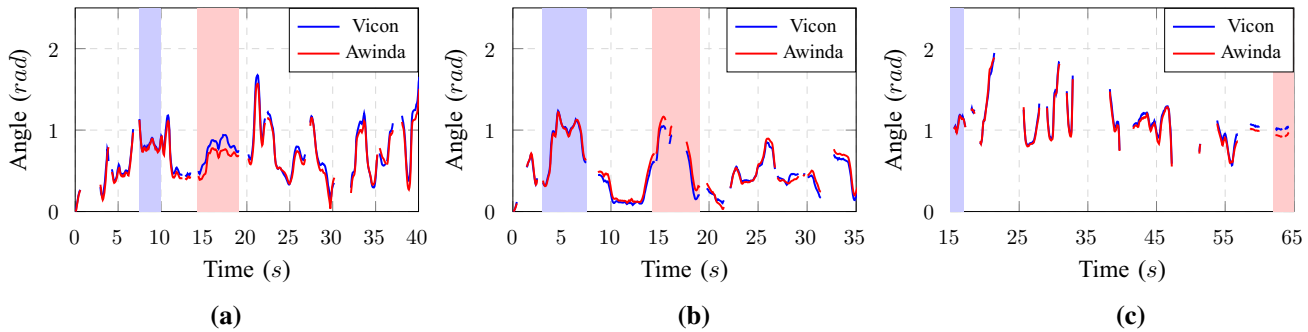


Fig. 9 Axis-angle representation between Vicon's head object and Awinda's head segment: **a** FE1; **b** FE2; **c** FE3. Blue region highlights the best alignment between both objects, while red region highlights worst alignment

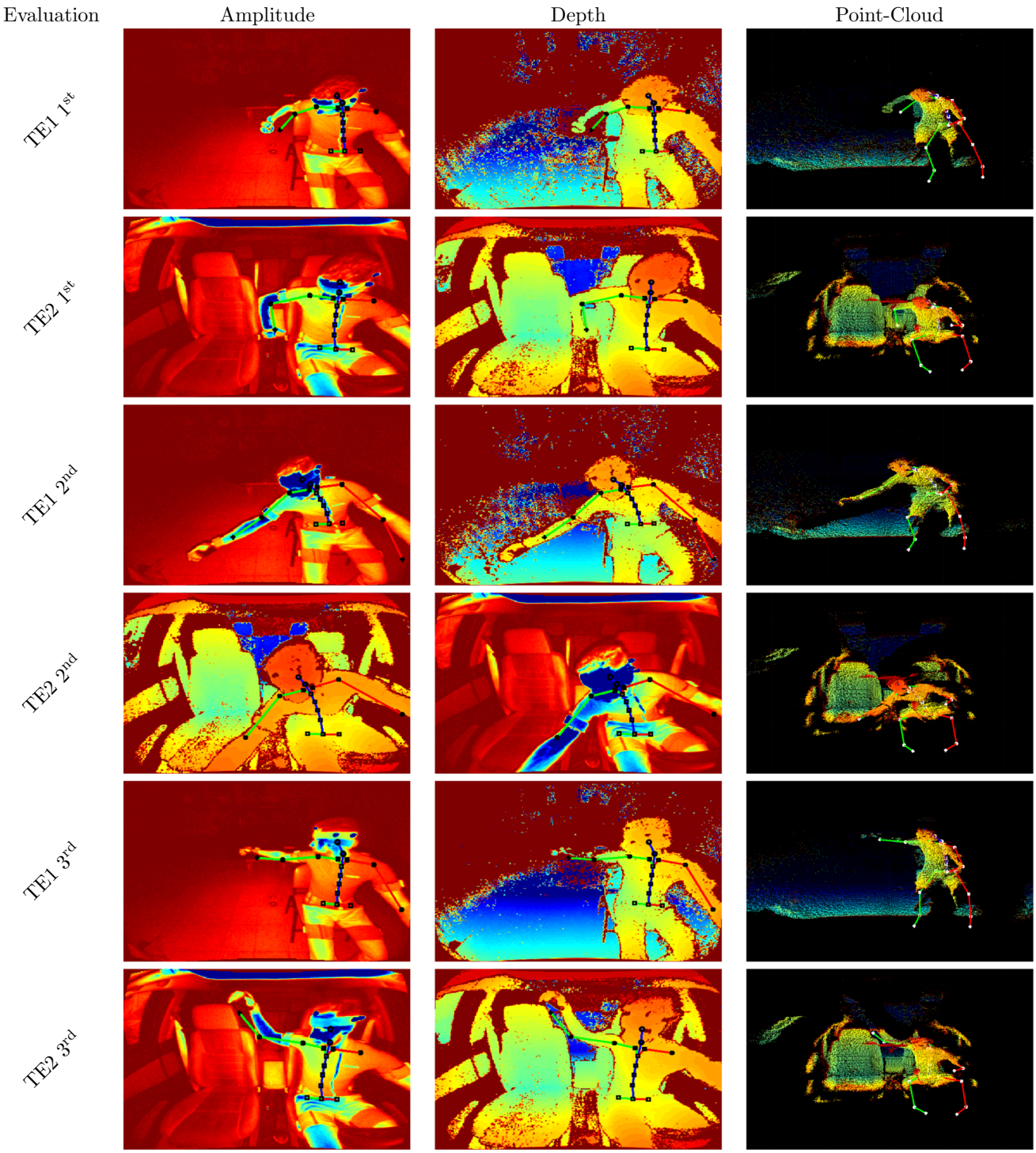


Fig. 10 Real dataset frames with associated body pose ground-truth. Representation is done considering the head maximum rotations for each simulated action in each evaluation (Fig. 8)

4.2 Application to pose estimation problems

To understand the validity of the data being generated with our system, as well as its ability to increase ML algorithmic accuracy, we defined three distinct experimental scenarios: 2D pose estimation from depth images; 2D pose estimation from point-cloud; and 3D pose estimation from 2D pose.

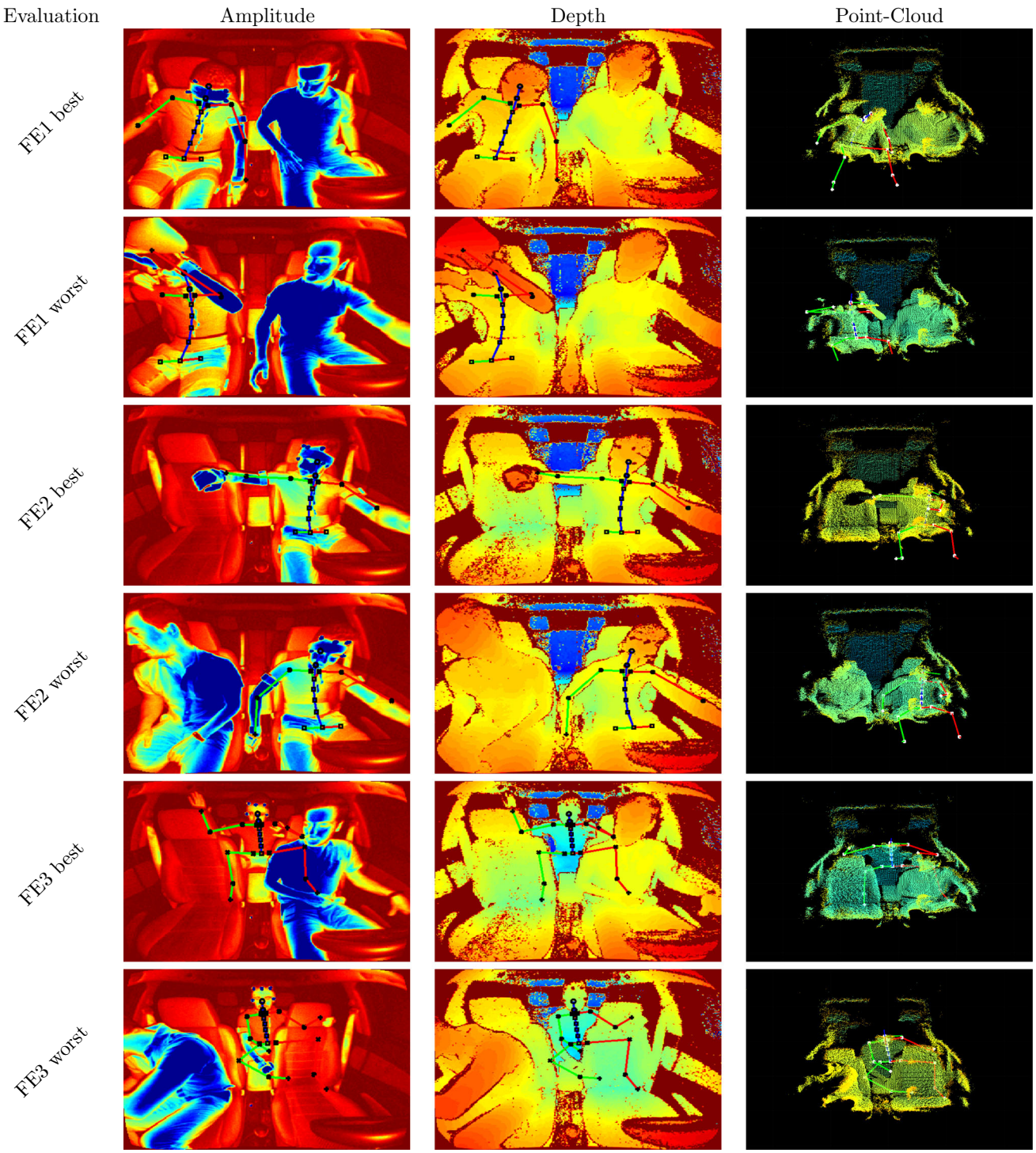


Fig. 11 Real dataset frames with associated body pose ground-truth. Representation is done considering the best and worst head alignment in each evaluation (Fig. 9)

4.2.1 Evaluation data

Since we aim to evaluate the potential advantage of using an in-car focused dataset over a generic one, we require samples from our system and from publicly available datasets. In this sense, we used two publically available datasets: The first was ITOP [12], containing 17,991 real images and corresponding ground-truth from a single subject (S_1), and the second was the NTU RGB+D [29], where the first subject (Z_1) and all of its planes (P), cameras (C), rotations (R)

Table 1 Evaluations related to system and public data quantities. (M1) system dataset; (M2) NTU RGB+D dataset; (M3–M4) system and NTU RGB+D datasets; (M5) ITOP; and (M6–M7) system and ITOP datasets

Evaluation	MoLa R8.7k InCar	ITOP	NTU RGB+D
$M1$[a]	6322	0	0
$M2$[a]	0	0	94,321
$M3$[b]	6322	0	94,321
$M4$[a]	6322	0	94,321
$M5$[a]	0	17,991	0
$M6$[b]	6322	17,991	0
$M7$[a]	6322	17,991	0

[a]200 epochs for the full dataset.
[b]180 epochs for the public dataset plus 20 for fine-tuning with our system dataset

and 49 actions ($A_{1:49}$) were used, resulting in 94,321 real images and corresponding ground-truth. Our system generated dataset consists in five recorded subjects ($H_{1:5}$), two actions ($B_{1:2}$) each, totaling 8754 samples. All datasets are identical in terms of sample data types for depth frame and 3D/2D body pose, giving us the opportunity to evaluate the first and third experimental scenarios comparatively to each other. However, the second experimental scenario is able to also evaluate our proposed dataset to estimation problems based on point-clouds. The available samples were divided into 3 groups: (1) a training set, with all public datasets plus 6322 system samples (from subjects $H_{1:4}$); (2) a validation set with 702 system samples (from subjects $H_{1:4}$); and (3) a test set with 1730 system samples (from subject H_5 performing distinct actions). To assess the influence of mixing public datasets with the system generated ones, we established seven sub-evaluations (M) for the first and third experiment, as shown in Table 1. For the second experiment, we used $M1$ sub-evaluation only. The proposed system generated dataset, plus the tools needed to reproduce all experiments, was made publically available [3].

2D Pose estimation from depth images (EV1) To evaluate the system generated depth frames and corresponding 2D ground-truth, the Part Affinity Fields [6] method was used. From it, a custom CNN was implemented consisting only on the first stage of the original PAF CNN [33], following the same training procedures. In each sub-evaluation, $M\#$, the method used the depth frame as input features and the 2D body pose as output labels (Fig. 12). For all samples, the depth frame was normalized into a grayscale frame ([0; 1.8] $m \equiv$ [0; 255]), while each 2D joint position was converted into a 2D heatmap. For metric evaluation, the joint position is estimated through non-maximum suppression applied to the inferred heatmap. In this experiment, the PCKh measure (in pixels, using a matching threshold given by 50% of the head segment length) and the area under curve (AUC) were used as metrics [1]. Table 2 summarizes the average results for the full body, with the results for individual joints being presented in Table S1. Figure 13a presents the PCKh@0.5 values for the full body for each sub-evaluation.

2D Pose estimation from point-cloud (EV2) To evaluate the system generated point-cloud and corresponding 2D ground-truth, the Part Affinity Fields [33] method was used. In this experiment, the point-cloud was used as input features. To this end, each point-cloud was normalized (pcx and pcy with $[-1.5; 1.5]$ $m \equiv [0; 255]$, and pcz with $[0; 1.8]$ $m \equiv [0; 255]$) and converted into a 3-channel matrix. As for EV1, the network's output was the 2D heatmaps generated from

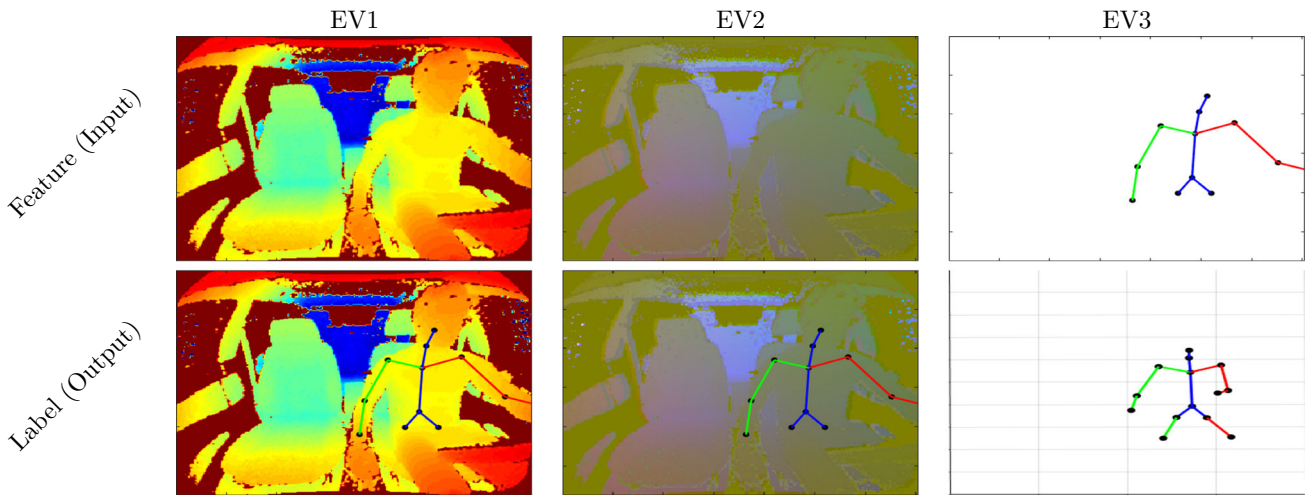


Fig. 12 Visual representation of input features and output used for each experimental scenario $EV\#$: (EV1) 2D pose estimation from depth images using normalized depth frame as input and 2D body pose as output; (EV2) 2D pose estimation from point-cloud using normalized point-cloud as input and 2D body pose as output; and (EV3) 3D pose estimation from 2D pose using 2D body pose as input and 3D body pose as output

Table 2 PCKh measure and AUC values averaged over all 14 joints, for the 3 experimental scenarios and all 7 sub-evaluations

	*M*1	*M*2	*M*3	*M*4	*M*5	*M*6	*M*7
EV1							
PCKh[a]	95.97	0.01	69.54	94.88	12.64	94.57	95.48
AUC	56.14	0.01	33.36	59.39	5.26	58.04	55.33
EV2							
PCKh[a]	96.47						
AUC	64.43						
EV3							
PCKh[b]	95.53	0.00	95.12	97.25	0.00	95.80	97.73
AUC	59.87	0.00	56.81	57.68	0.00	62.07	65.63

[a]EV1 and EV2 do matching threshold to 26 *pixels*.
[b]EV3 does matching threshold to 200 mm

each joint's position, with the inferred joint position being computed by non-maximum suppression. The same metrics from EV1 were employed. Results are shown in Tables 2, S1 and Fig. 13b.

3D pose estimation from 2D pose (EV3) To evaluate the system generated 3D ground-truth, a 3D pose estimation method [20] was used, following the same training procedures. The method uses a 2D body pose as input features (provided as joint pixel coordinates) and the 3D body pose as output (Fig. 12). Once again, similar metrics were employed, but in this case, PCKh matching threshold was normalized to a fixed head size of 200 mm. Results are shown in Tables 2, s1 and Fig. 13c.

5 Discussion

In this paper, we presented a system capable of generating human body pose and time-of-flight data for in-car scenario.

We evaluated a specific inertial suit (see Appendix B), highlighting its limitations through an extensive evaluation procedure that separates itself from other more specific methods [24] (where the evaluation is focused mainly in the calibration procedure and the full-body kinematics output). We believe that this evaluation of the inertial suit brings a better understanding on its behavior and limitations.

In terms of the system's output for human body pose detection, it falls behind other methods [8,32] when used in an open space. The big novelty and advantage are for the in-car scenario. Here, our system improves considerably on others that share the same in-car focus [4]. As previously mentioned, we fuse two state-of-the-art motion capture systems (optical and inertial). By doing it, we suppress their standalone limitations (marker occlusion, global positioning drift) and increase their added benefits by creating a motion capture system for highly occluded scenarios. We were able to record and project a human motion capture system into an image sensor in an heavily occluded scenario. The projection is possible through specific calibration procedures that allow for a temporal and spatial alignment of all recorded systems. Due to this complex pipeline, several sources of error exist, with the major ones being associated with the inertial motion capture suit (Figure S14) and the Awinda-to-Vicon head spatial alignment (Fig. 9). Despite this calibration sensitivity, we were able to record in-car datasets with proper human body pose motion capture (Fig. 11). Our system shows robustness to magnetic distortion scenarios, namely inside the vehicle, where it was possible to observe a performance similar to movements performed outside the vehicle (Fig. 10). We believe that our system can also be applied in other scenarios where ambient occlusion is a limitation factor for motion capture.

In terms of data validation, Fig. 13 and Table 2 demonstrate the interest in using system generated data in ML training for the in-car scenario, showing PCKh improvements in all experiments when adding system generated data.

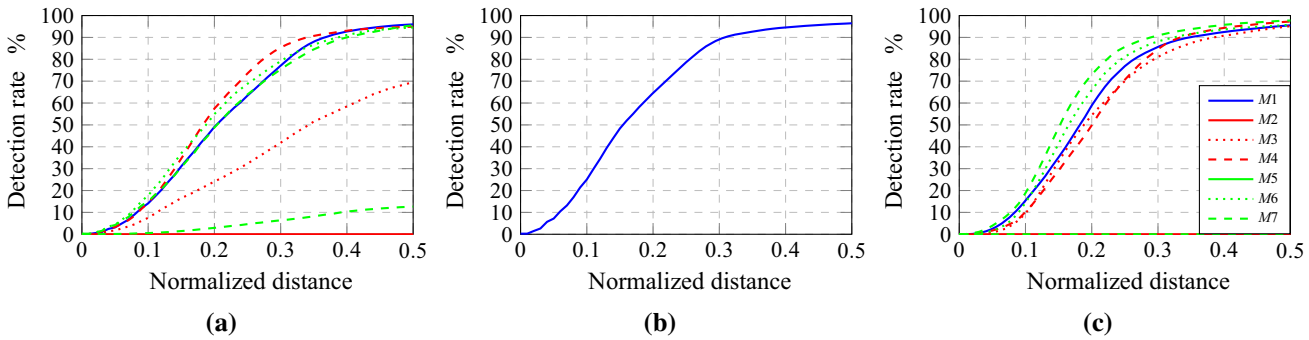


Fig. 13 PCKh total for all sub-evaluations, *M*#, and the three first experimental scenarios, *EV*#: **a** 2D pose estimation from depth images (EV1); **b** 2D pose estimation from point-cloud (EV2); and **c** 3D pose estimation from 2D pose (EV3). Color gradient represents different combinations of datasets. Continuous lines represent one dataset trained for 200 epochs, dotted lines represent two datasets trained in sequence (1st →180 epochs, 2nd →20 epochs), and dashed lines represent two mixed datasets trained for 200 epochs

$EV1{:}M1$ and $EV2{:}M1$ proved that training with specific use-case datasets can achieve best accuracy (in-car scenario). For evaluations $M5$ and $M2$, lack of domain representation hinders the performance. Factors such as scenery layout, human body behavior and sensor metrological characteristics can lead to considerable overfit. On the other end, mixed dataset combinations (i.e., ITOP or NTU RGB+D plus MoLa 8.7k InCar) also performed better then single generic dataset training (e.g., $M2 < M4$ and $M5 < M7$), demonstrating again the interest of using specific use-case datasets for training. Evaluations that rely on fine-tuning also presented similar results (e.g., $M2 < M3$ and $M5 < M6$), meaning that pre-trained models can be fine-tuned with our system's dataset to achieve good results. The influence of fine-tuning is also proportional to the ratio between added samples (from our dataset) and the public one. In this case, we see that for $M3$ (ratio of 1:15) specifically, we could use more in-car samples to achieve better performance (or a larger number of fine-tuning epochs). Interestingly, besides demonstrating the interest of system generated data for increased algorithmic accuracy, the present results also seem to suggest that the use of a 3D point-cloud as input may lead to a better pose inference when compared to networks using depth images (see, for example, the higher AUC values for $EV2{:}M1$ compared with $EV1{:}M1$). In the end, for the in-car 2D body pose estimation problem, we achieved higher PCKh@.5 total score in $EV2{:}M1$ (i.e., MoLa 8.7k InCar point-cloud-based training) of 96.47% and corresponding AUC of 64.43%. For the in-car 3D body pose estimation problem, the same conclusions can be made with regard to using specific use-case datasets vs. generic ones. However, we also achieved better results when mixing datasets instead of training with the system generated dataset alone (i.e., $EV3{:}M1 < M6 < M7$). In this case, considering both metrics (PCKh@0.5 total and AUC), we achieved 97.73% and 65.63%, respectively, for $EV3{:}M7$. Finally, Table S1 summarizes the results for individual joints, being possible to conclude that joints frequently present at the image's limits (wrists and elbows) are the most problematic. This may be related to the lower number of training samples with these joints visible (as they are more frequently outside of the camera's field of view or in the camera's deadzone). Hip joints also show similar problematic results, but in this case, it is related to two types of errors (i.e., Awinda-to-Vicon spatial alignment (Sect. 3.2.5) and forward kinematics error propagation from head the joint [Appendix B, Section C]), creating less stable ground-truth data for these specific joints.

6 Conclusions

In this work, a novel system for the generation of in-car human body pose datasets is presented. The system demonstrated to be able to generate datasets through a specific setup consisting in an inertial suit, a global positioning system and a ToF camera, coupled with a set of calibration procedures. The motion capture system was thoroughly evaluated, and the sources of error were presented. A system generated dataset is also made publicly available.

In terms of future work, and regarding the calibration procedure, an extra step could be added for the correction of the initial suit calibration, as previously presented in [24]. This would allow a reduced systematic error for the projected body pose. Notwithstanding, this step would have to be non-intrusive for the recording procedure. In terms of dataset quality, there are currently several limitations, mostly coming from the inertial suit (namely related to sensor fixation and soft tissue movement). This could be solved with future inertial suits or better initial calibration procedures from the supplier.

Acknowledgements This work is supported by: European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project no 002797; Funding Reference: POCI-01-0247-FEDER-002797].

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: new benchmark and state of the art analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2014). https://doi.org/10.1109/CVPR.2014.471
2. Baak, A., Müller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Consumer Depth Cameras for Computer Vision, pp. 71–98 (2013). https://doi.org/10.1007/978-1-4471-4640-7_5
3. Borges, J., Queirós, S., Oliveira, B., Torres, H., Rodrigues, N., Coelho, V., Pallauf, J., Henrique, Brito J., Mendes, J., C Fonseca J.: MoLa R8.7k InCar Dataset (2019). https://doi.org/10.17632/724C998H9C.1
4. Borghi, G., Venturelli, M., Vezzani, R., Cucchiara, R.: POSEidon: Face-from-depth for driver pose estimation. In: Proceedings 30th IEEE conference on computer vision and pattern recognition, CVPR 2017 2017-Janua, pp. 5494–5503 (2017). https://doi.org/10.1109/CVPR.2017.583, arXiv:1611.10195
5. Buys, K., Cagniart, C., Baksheev, A., De Laet, T., De Schutter, J., Pantofaru, C.: An adaptable system for RGB-D based human body detection and pose estimation. J. Vis. Commun. Image Represent. **25**(1), 39–52 (2014). https://doi.org/10.1016/j.jvcir.2013.03.011
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 2017-Janua, pp. 1302–1310 (2017). https://doi.org/10.1109/CVPR.2017.143, arXiv:1611.08050
7. Chen, W., Wang, H., Li, Y., Su, H., Wang, Z., Tu, C., Lischinski, D., Cohen-Or, D., Chen B.: Synthesizing training images for boosting human 3D pose estimation. In: Proceedings 2016 4th International Conference on 3D Vision, 3DV 2016, pp. 479–488 (2016). https://doi.org/10.1109/3DV.2016.58, http://irc.cs.sdu, arXiv:1604.02703
8. CMU (2016) CMU Dataset@mocap.cs.cmu.edu. http://mocap.cs.cmu.edu/

9. Demirdjian, D., Varri C.: Driver pose estimation with 3D Time-of-Flight sensor. In: 2009 IEEE Workshop on Computational Intelligence in Vehicles and Vehicular Systems, IEEE, pp. 16–22 (2009). https://doi.org/10.1109/CIVVS.2009.4938718, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4938718
10. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real time motion capture using a single time-of-flight camera. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR2010), pp. 755–762 (2010). https://doi.org/10.1109/CVPR.2010.5540141, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5540141
11. Ganapathi, V., Plagemann, C., Koller, D., Thrun, S.: Real-time human pose tracking from range data. In: European Conference on Computer Vision, pp. 738–751 (2012). https://doi.org/10.1007/978-3-642-33783-3_53, http://link.springer.com/10.1007/978-3-642-33783-3_53
12. Haque, A., Peng, B., Luo, Z., Alahi, A., Yeung, S., Fei-Fei, L.: Towards viewpoint invariant 3D human pose estimation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9905 LNCS, pp. 160–177 (2016). https://doi.org/10.1007/978-3-319-46448-0_10, arXiv:1603.07076
13. 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. **36**(7), 1325–1339 (2014). https://doi.org/10.1109/TPAMI.2013.248
14. Joo, H., Simon, T., Sheikh, Y.: Total capture: a 3D deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 8320–8329 (2018). https://doi.org/10.1109/CVPR.2018.00868, http://www.cs.cmu.edu/, arXiv:1801.01615
15. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: a massively multiview system for social interaction capture. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 190–204 (2019). https://doi.org/10.1109/TPAMI.2017.2782743
16. Jung, H.Y., Lee, S., Heo, Y.S., Yun, I.D.: Random tree walk toward instantaneous 3D human pose estimation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 07-12-June, pp. 2467–2474 (2015). https://doi.org/10.1109/CVPR.2015.7298861
17. Kroon, D.J.: Segmentation of the mandibular canal in cone-beam CT data. Ph.D. thesis, University of Twente, Enschede, The Netherlands (2011). https://doi.org/10.3990/1.9789036532808, http://purl.org/utwente/doi/10.3990/1.9789036532808
18. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2556–2563 (2011). https://doi.org/10.1109/ICCV.2011.6126543
19. Lee, S.J., Motai, Y., Choi, H.: Tracking human motion with multichannel interacting multiple model. IEEE Trans. Ind. Inform. **9**(3), 1751–1763 (2013)
20. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (2017). https://doi.org/10.1109/ICCV.2017.288
21. Martinez-Gonzalez, A., Villamizar, M., Canevet, O., Odobez, J.M.: Efficient convolutional neural networks for depth-based multi-person pose estimation. IEEE Trans. Circuits Syst. Video Technol. (2019). https://doi.org/10.1109/tcsvt.2019.2952779
22. Mcneal, J.R.D.P., Eastern, H.A.S., Education, P.: The united states olympic committee uses the polhemus LIBERTY ^{TM} to research the effects of acute static stretch on joint position sense in the shoulder. Computer **4777**, 800–802 (2003)
23. Mitobe, K., Kaiga, T., Yukawa, T., Miura, T., Tamamoto, H., Rodgers, A., Yoshimura, N.: Development of a motion capture system for a hand using a magnetic three dimensional position sensor. In: ACM SIGGRAPH 2006 research posters on: SIGGRAPH '06, p. 102 (2006). https://doi.org/10.1145/1179622.1179740, http://dl.acm.org/citation.cfm?id=1179622.1179740
24. Orozco, M.: Assessment of postural deviations associated errors in the analysis of kinematics using inertial and magnetic sensors and a correction technique proposal by assessment of postural deviations associated errors in the analysis of kinematics using inertial. Ph.D. thesis, University of Toronto (2015)
25. Pekelny, Y., Gotsman, C.: Articulated object reconstruction and markerless motion capture from depth video. Comput. Graph. Forum **27**(2), 399–408 (2008). https://doi.org/10.1111/j.1467-8659.2008.01137.x
26. Plagemann, C., Ganapathi, V., Koller, D., Thrun, S.: Real-time identification and localization of body parts from depth images. In: Proceedings—IEEE International Conference on Robotics and Automation, pp. 3108–3113 (2010). https://doi.org/10.1109/ROBOT.2010.5509559
27. Rahmatalla, S., Xia, T., Contratto, M., Kopp, G., Wilder, D., Frey Law, L., Ankrum, J.: Three-dimensional motion capture protocol for seated operator in whole body vibration. Int. J. Ind. Ergon. **38**(5–6), 425–433 (2008). https://doi.org/10.1016/j.ergon.2007.08.015
28. Roetenberg, D., Luinge, H., Slycke, P.: Xsens MVN: full 6DOF human motion tracking using inertial sensors. Technical report, Xsens Technologies (2013)
29. 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 Computer Society Conference on Computer Vision and Pattern Recognition 2016-Decem, pp. 1010–1019 (2016). https://doi.org/10.1109/CVPR.2016.115, arXiv:1604.02808
30. 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: Conference on Computer Vision and Pattern Recognition 2011, pp. 1297–1304 (2011). https://doi.org/10.1109/CVPR.2011.5995316, http://ieeexplore.ieee.org/document/5995316/, arXiv:1111.6189v1
31. 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. Stud. Comput. Intell. **411**, 119–135 (2013). https://doi.org/10.1007/978-3-642-28661-2-5
32. 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. **87**(1–2), 4–27 (2010). https://doi.org/10.1007/s11263-009-0273-6
33. Torres, H.R., Oliveira, B., Fonseca, J., Queirós, S., Borges, J., Rodrigues, N., Coelho, V., Pallauf, J., Brito, J., Mendes, J.: Real-time human body pose estimation for in-car depth images. In: IFIP Advances in Information and Communication Technology. Springer, New York LLC, vol. 553, pp. 169–182 (2019). https://doi.org/10.1007/978-3-030-17771-3_14
34. Whitehead, A., Laganiere, R., Bose, P.: Temporal synchronization of video sequences in theory and in practice. In: Proceedings—IEEE Workshop on Motion and Video Computing, MOTION 2005 (2007). https://doi.org/10.1109/ACVMOT.2005.114
35. Wu, G., Siegler, S., Allard, P., Kirtley, C., Leardini, A., Rosenbaum, D., Whittle, M., D'Lima, D.D., Cristofolini, L., Witte, H., Schmid, O., Stokes, I.: ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion-part I: ankle, hip, and spine. J. Biomech. **35**(4), 543–548 (2002). https://doi.org/10.1016/S0021-9290(01)00222-6
36. Xing, T., Yu, Y., Zhou, Y., Du, S.: Markerless motion capture of human body using PSO with single depth camera. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, pp. 192–197 (2012). https://

doi.org/10.1109/3DIMPVT.2012.21, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6374994
37. Yan, C., Shao, B., Zhao, H., Ning, R., Zhang, Y., Xu, F.: 3D Room layout estimation from a single RGB image. IEEE Trans. Multimed. **14**(8), 1–1 (2020). https://doi.org/10.1109/tmm.2020.2967645
38. Ye, M., Shen, Y., Du, C., Pan, Z., Yang, R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1517–1532 (2016). https://doi.org/10.1109/TPAMI.2016.2557783
39. Zhang, Z.: A flexible new technique for camera calibration. IEEE Trans. Pattern Anal. Mach. Intell. (2000). https://doi.org/10.1109/34.888718

João Borges received the I.M degree in Industrial Electronics and Computers Engineering from the University of Minho (Portugal), in 2008. In 2020, he completed his Ph.D. degree in Electronics and Computers Engineering from the University of Minho (Portugal). He has since been a postdoctoral researcher at the University of Minho. He has developed several industrial applications, such as online quality control systems for product validation, ECU assembling cells and CNC machines. He has worked also in R&D product development, such as intelligent furniture. His research interests include computer vision based solutions for quality control inspection through thermography, RGB-D and RGB-NIR sensors.

Sandro Queirós was born in Porto, Portugal, in 1991. He received his M.Sc. degree in Biomedical Engineering from the University of Minho (Portugal) in 2013. In 2018, he completed his joint Ph.D. degree in Biomedical Engineering and Biomedical Sciences from the University of Minho (Portugal) and KU Leuven (Belgium), respectively. He has since then been a postdoctoral researcher at the University of Minho. His current research interests include medical image analysis and computer vision. Specifically, his work has been mainly focused on the development of novel medical image processing solutions for the diagnosis and treatment of diseases, namely in the cardiovascular field.

Bruno Oliveira received his MS degree in Biomedical Engineering from the University of Minho in 2016. He is currently a Ph.D. student in the School of Engineering, University of Minho. His research interests include medical image processing, robotics, computer vision and machine learning applied in both medical and automotive areas.

Helena Torres received her MS degree in Biomedical Engineering from the University of Minho in 2016. She is currently a Ph.D. student in the School of Engineering, University of Minho. Her research interests include medical image processing, computer vision and machine learning applied in both medical and automotive areas.

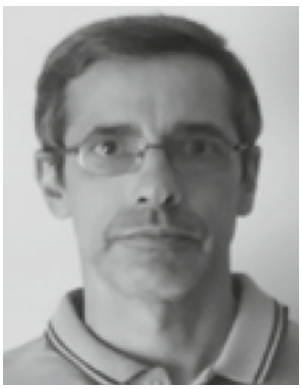
Nelson Rodrigues acquired the degree of Electrical and Computers Engineering in 2012 and the Master in Automation and Robotics in 2015, from the Institute Polytechnic of Cávado and Ave. He is currently a Ph.D. student in the AESI doctoral program. Since 2017 he worked in R&D for autonomous vehicles interior solutions. His research interests include electronics, robotics, software, computer vision and artificial intelligence.

Victor Coelho received his degree in Industrial Electronics in 1999, and his master thesis in Image Processing in 2005, both from the University of Minho, in Portugal. He has been working at Bosch Car Multimedia since 2005 as a member of the Center of Competence for Testing. Since 2018 he has become the Manufacturing Test Architect for Bosch Car Multimedia. His main interests are testing of PCBA, test strategies and test technologies.

Johannes Pallauf received the Dipl.-Ing. degree in electrical engineering from the Technical University of Munich, in 2011 and his Ph.D from Karlsruhe Institute of Technology in 2016. He is currently a computer vision developer at Robert Bosch Car Multimedia GmbH, Hildesheim, Germany, in the area of interior sensing.

José Henrique Brito is Assistant Professor at the School of Technology of the Polytechnic Institute of Cávado and Ave. He graduated in Electrical Engineering and Computers at Instituto Superior Técnico in Lisbon, Portugal in 1999, earned a Masters degree in Computer Graphics and Virtual Environments from the University of Minho, Portugal in 2009 and a Ph.D. also from the University do Minho in 2014. The research focus of his Ph.D. thesis was in computer vision, namely in multiple view geometry and camera calibration. He has since worked on research projects in machine learning and deep learning, object detection, object recognition, and semantic segmentation.

José Mendes holds a PhD in Electronics by the University of Hull, UK. Currently is an Assistant Professor of the Industrial Electronics Department, School of Engineering, University of Minho Portugal. José Mendes is a member of the Embedded Systems Research Group of the Industrial Electronics Department. As a result of his research he holds several patents (some under submission process), scientific papers and successfully completed the guidance of several postgraduate students (PhDs and MSc). He is currently involved in applied research projects with European Companies under "FP7 - Research for the benefit of SMEs".



Jaime C. Fonseca is an Associate Professor (with tenure) at the Department of Industrial Electronics, University of Minho. He is the coordinator of the Industrial Electronics line at the Algoritmi research centre, Member of the scientific committee of the doctoral program in Industrial Engineering Electronics and Computers and CEO of iSurgical3D company Spin Off of Minho University. His research focus on automation, robotics and medical devices. He participated in 39 research projects mostly in close cooperation with the industry. He has authored or co-authored in 77 indexed publications in international peer reviewed journals and conferences. He has a total of 317 scopus citations and 538 Google Scholar citations. He has co-authored in 3 Portuguese patents, 1 international patent and won 2 awards.