ROS4HRI: Standardising an Interface for Human-Robot Interaction

Raquel Ros
raquel.ros@pal-robotics.com
PAL Robotics
Barcelona, Spain

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

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

Antonio Andriella
antonio.andriella@pal-robotics.com
PAL Robotics
Barcelona, Spain

Aina Irisarri
aina.irisarri@pal-robotics.com
PAL Robotics
Barcelona, Spain

ABSTRACT

Benchmarking and reproducibility in Human-Robot Interaction (HRI) is notoriously difficult to achieve, due in no small part to the inherent complexity and diversity of human behaviours. The lack of a standard and open-source software platform is however a significant additional hurdle, as it severely hinders the creation of readily available benchmarks, with comparable inputs and outputs.

The release in 2022 of the ROS4HRI standard aims at removing this obstacle, by defining both a standard and a reference implementation for a set of software modules dedicated to HRI, and build on top of the widely deployed ROS framework. We present in this article the main features of this open standard, visualisation tools and also include some recent developments towards the standardisation of *user intents*.

KEYWORDS

human-robot interaction, open-source, standards, social robotics

ACM Reference Format:
Raquel Ros, Séverin Lemaignan, Lorenzo Ferrini, Antonio Andriella, and Aina Irisarri. 2023. ROS4HRI: Standardising an Interface for Human-Robot Interaction. In *2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI) Workshop on Advancing HRI Research and Benchmarking Through Open-Source Ecosystems, March 13, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION: THE NEED FOR AN OPEN STANDARD FOR HRI SOFTWARE

Integrating real-time, complex social signal processing into robotic systems –especially in real-world, multi-party interaction situations– is a challenge faced by many in the Human-Robot Interaction (HRI) community. The difficulty is compounded by the lack of any standard model for human representation that would facilitate the development and interoperability of social perception components and pipelines. The situation has significant consequences for meta-research, where the lack of standard interfaces effectively prevents the design of generally accepted benchmarks or reusable data formats.

Arguably, the NAOqi framework (as found in the Aldebaran Nao and Pepper robots [6]) is relied on by a significant proportion of the HRI researchers, due to the large number of Aldebaran robots available across the world [4]. The NAOqi framework features advanced social perception capabilities like people recognition, gaze and expression monitoring, engagement estimation. However, NAOqi is not open-source, and as such, not available outside of the Aldebaran robots, with no possibility for the scientific community to influence its development. This severely limits its usefulness as a shared, open software platform for benchmarking.

ROS4HRI was introduced by Mohamed and Lemaignan [5] in 2021, and accepted in 2022 as the REP-155[1] *Open* ROS specification. ROS4HRI is a set of conventions and standard interfaces for HRI scenarios, designed to be used with the Robot Operating System (ROS), and not tied to any particular robot platform. ROS4HRI directly aims at promoting interoperability and re-usability of core functionality between the many HRI-related software tools, from skeleton tracking, to face recognition, to natural language processing. Importantly, these interfaces were designed to be relevant to a broad range of HRI applications, from high-level crowd simulation, to group-level social interaction modelling, to detailed modelling of human kinematics.

Some of the design choices of the ROS4HRI framework (like the use of both transient *features* ID and permanent *persons* ID to keep track of people even when they are temporarily not detected anymore) have been influenced by NAOqi, thus the two framework offer some similarities at the design level. ROS4HRI is however primarily designed around the ROS paradigms (like *topics*, *services*, *frames*, etc), and designed to fully integrate with the rest of the ROS ecosystem.

Specifically, we believe that ROS4HRI could significantly improve the current situation with regard to benchmarking and reproducibility by:

- establishing common interfaces (inputs/outputs) that allow direct performance comparison between different algorithms developed in the community;
- supporting sustainable software development, where evolving from one software platform to another should be transparent to the entire system architecture and shareable among different systems.

We next review the ROS4HRI representation of humans (Section 2) and outline the ROS specification underlying the ROS4HRI framework (Section 3). We introduce initial developments of *user intents* (Section 4) and visualisations tools lately implemented (Section 5).

[1] https://www.ros.org/reps/rep-0155.html

HRI 2023 Workshop, March 13, 2023, Stockholm, Sweden
2023.

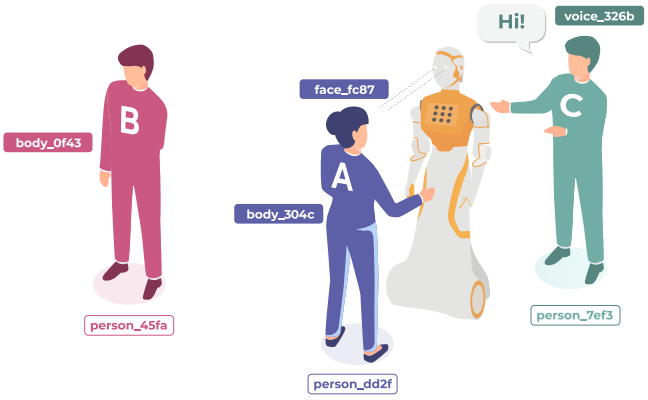

Figure 1: In this situation: A is facing the robot: A gets a unique `faceID`, a unique `bodyID`, and a unique `personID`; B's body is visible to the robot, but not the face: B only gets a `bodyID` and `personID`; C is not seen, but heard: C gets a `voiceID` and a `personID`.

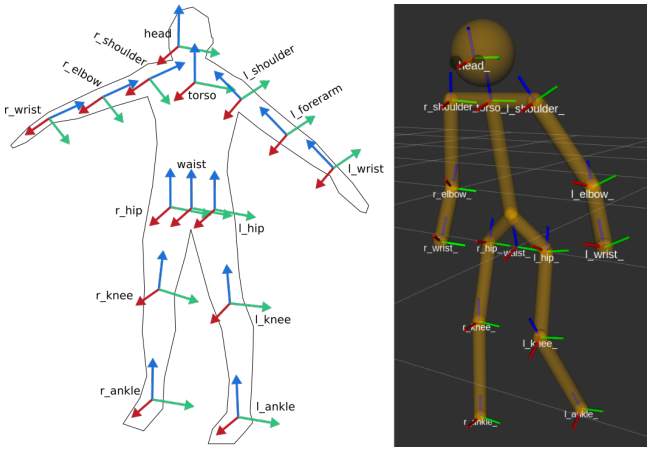

Figure 2: Left: the 15 links defined on the human body. Frames orientations and naming are based on REP-103 and REP-120. Right: the URDF kinematic model, viewed in RViz.

Finally, we provide an example of its use in the PAL ARI social robot and next steps (Section 6).

2 THE ROS4HRI HUMAN MODEL

2.1 The four human identifiers

To accommodate existing approaches used to detect and recognise humans, the representation of a person is built on a combination of 4 unique identifiers: a face, a body, a voice, and a person identifier. These four identifiers are not mutually exclusive, and depending on the requirements of the application, the available sensing capabilities, and the position/behaviour of the humans, only some might be available for a given person, at a given time (Figure 1).

2.1.1 Face, body and voice identifier. These identifiers are unique IDs that identify a detected face (`<faceID>`), a body skeleton (`<bodyID>`) or a voice `<voideID>`. These are typically generated by either a face detector, a body tracker or a voice separation module respectively.

Importantly, these IDs are **not persistent**: once a feature is lost (for instance, the person moves away from the robot view), its ID is not valid nor meaningful anymore. To cater for a broad range of applications (where re-identification might not be always necessary), there is no expectation that the different detectors will attempt to recognise the same feature and re-assign the same ID if the person re-appears.

Every face, body and voice location can be represented via ROS TF frames, where the frame IDs are `face_<faceID>`, `body_<bodyID>` and `voice_<voiceID>` respectively (Section 3.3 details the frame conventions).

At any given time, the list of tracked faces, bodies and voices are published under the `humans/faces/tracked`, `humans/bodies/tracked` and `humans/voices/tracked` topics.

2.1.2 Person identifier. The person identifier is a unique ID **permanently** associated with a unique person. This ID should be assigned by a module able to perform person identification (face recognition module, voice recognition module, sound source localisation + name, identification based on physical features like height/age/gender, clothes colour, etc.). This ID is meant to be persistent so that the robot can recognise people across encounters/sessions.

As soon as a face, a body or a voice is detected, the robot can infer the presence of a person, and therefore a person ID must be created and associated with that face/body/voice. As person IDs are permanent, that ID will permanently remain in the robot's knowledge.

When meaningful, a TF frame must be associated with the person ID, named `person_<personID>`. Due to the importance of the head in HRI, the `person_<personID>` frame is expected to be placed as close as possible to the head of the human. If neither the face nor the skeleton is tracked, the `person_<personID>` frame might be located to the last known position of the human or removed altogether if no meaningful estimate of the human location is available. We detail below the rules associated to the `person_<personID>` frame.

2.2 Human kinematic model

ROS4HRI adopts the URDF[2] standard to represent human kinematics. Humans anatomies, however, vary from one person to the other, reflecting individual height, weight, etc. Every time a body is detected, a custom URDF model should be generated to take into account for these differences, using the person's observed height as the main parameter, from which the other dimensions (e.g., shoulder-to-shoulder width) are derived, based on standard models of anthropometry.

The generated URDF model is then published on the ROS parameter server (using the `bodyID` unique identifier), making it available to the rest of the ROS network. The URDF model is used in combination with the computed joint state of each tracked body part

[2] http://wiki.ros.org/urdf

to then generate a kinematically-sound, real-time 3D model of the person (Figure 2).

3 ROS SPECIFICATION

3.1 Topics structure

ROS4HRI exposes social signals using a specific structure of ROS topics, and introduces a set of new ROS messages. The following rules apply to present human perceptions in a ROS system:

(1) all topics are grouped under the global namespace `/humans`
(2) five sub-namespaces are available:
   - `/humans/faces`
   - `/humans/bodies`
   - `/humans/voices`
   - `/humans/persons`
   - `/humans/interactions`
(3) the first four (`/faces`, `/bodies`, `/voices`, `/persons`) expose one sub-namespace per face, body, voice, person detected, named after the corresponding ID. For instance, `/humans/faces/<faceID>/`. In addition, they expose a topic `/tracked` where the list of currently tracked faces/ bodies/ voices/ persons is published.
(4) the `/humans/interactions` topic exposes group-level signals, including gazing patterns and social groups.

3.2 The `hri_msgs` ROS messages

Table 1 lists the most relevant ROS messages introduced for HRI. They are regrouped in the `hri_msgs`[3] ROS package, where the full list of messages is available.

3.3 Frame conventions

The ROS4HRI standard specifies several TF frames to spatially represent a human. Where meaningful, the frames follow the conventions set out in the ROS *REP-103 – Standard Units of Measure and Coordinate Conventions*[4] and *REP-120 – Coordinate Frames for Humanoid Robots*[5].

3.3.1 Body frames. Figure 2 shows the frames defined on the human skeleton. The `waist_<bodyID>` is collocated with the body's root frame, `body_<bodyID>`. The origin of this frame is located at the midpoint between the two hips, and the parent of this frame would typically be the sensor frame used to estimate the body pose. The x-axis of the frames points forward (i.e., out of the body), while the z-axis points toward the head. The 15 links are connected through 18 joints (head, shoulder, elbows, knees, hips and waist).

All skeleton points published as TF frames are suffixed with the same `<bodyID>`, thus enabling several unique skeletons to be tracked and visible in TF simultaneously (not visible on Fig. 2 for clarity).

3.3.2 Face frame. Head pose estimation modules are requested to publish the 6D head pose as a TF frame named `face_<faceID>`. The parent of this frame is the sensor frame used to estimate the face pose. The origin of the frame must be the sellion (defined as the deepest midline point of the angle formed between the nose and forehead). The x-axis is expected to point forward (i.e., out of the face), the z-axis is expected to point toward the scalp (i.e., up when the person is standing vertically).

Head vs face frames If the skeleton tracker provides an estimate of the head pose, it might publish a frame named `head_<bodyID>`, located at the sellion. It is the joint responsibility of the face tracker and skeleton tracker to ensure that `face_<faceID>` `head_<bodyID>` are consistent with each other, e.g. collocated.

Gaze In addition to the face, a head pose estimator might publish a TF frame representing the gaze direction, `gaze_<faceID>`. While collocated with the `face` frame, it follows the convention of cameras' optical frames: the z-axis points forward, the y-axis points down.

3.3.3 Person frame. The `person_<personID>` needs to be interpreted in conjunction with the value published on the topic `/humans/persons/<personID>/location_confidence`. We can distinguish three cases:

- The human is currently being tracked (i.e. `personID` is set, and at least one of `faceID`, `bodyID` or `voiceID` is set). In this case, `location_confidence` should be 1, and the `person_<personID>` frame must be collocated with one of the available IDs, i.e. the `face_<faceID>` frame, or the skeleton frame closest to the head, or the best available approximation of the person's position (e.g. based on sound source localization).
- The human is not currently seen/heard, but a prior localization is known. In this case, `location_confidence` must be set to a value < 1 and a person `person_<personID>` TF frame must be published while `location_confidence` > 0.
- The system knows about the person (for instance, from dialogue with another person), but has no location information. In this case, `location_confidence` must be set to 0, and no TF frame should be broadcast.

4 INTENTS

As we have seen until this point, the ROS REP-155 is currently (March 2023) only covering the *perception* side of human-robot interactions. Although the standardisation of *behaviours* and actions is more complex, as it tends to be very platform and application specific, we present hereafter an initial attempt to specify standard *intents*, as abstract descriptions of operations to be performed by the robot.

While inspired by the Android intents, ROS intents are primarily designed to capture either explicit or implicit user-initiated intents. For instance, a button click on a touchscreen (explicit), the result of a chatbot-based verbal interaction (implicit/explicit), or a user looking towards the robot (implicit).

Intents are represented as ROS messages of type `hri_actions_msgs/Intent`, and published on the `/intents` topic. They are emitted (*published*) by nodes that track the user's activities (e.g., the touchscreen, the dialogue manager, action recognition), and are consumed by the application controllers (e.g. behaviour scripts). This latter is then in charge of scheduling and running the different actions and behaviours based on received intents, and allocate the robot's resources (to ensure no two actions

[3] https://github.com/ros4hri/hri_msgs
[4] https://www.ros.org/reps/rep-0103.html
[5] https://www.ros.org/reps/rep-0120.html

Table 1: List of relevant ROS messages for HRI

Message name	Short description
`AudioFeatures`	Encodes 16 low-level audio features, based on the INTERSPEECH'09 Emotion recognition challenge [7].
`BodyPosture`	Recognised body posture (eg standing, sitting)
`EngagementLevel`	Engagement status of the person with the robot
`Expression`	Facial expression, in a categorical manner (Ekman's model [2]), or using the Valence/Arousal continuous plane.
`FacialAction Units`	Encodes the intensity and confidence level of detected Facial Action Units, following the coding scheme and nomenclature proposed in [3].
`FacialLandmarks`	Encodes the 2D coordinates in image space (and confidence) of 67 facial landmarks (including mouth, nose, eyes, and face silhouette).
`Gaze`	Represents the gaze direction from a particular person (sender ID) to the person that is being gazed to (receiver ID).
`Gestures`	Recognised symbolic gesture detected from a body (eg waving)
`Group`	List of person IDs being detected as forming a social group.
`LiveSpeech`	Encodes the live result of a speech recognition process. A series of incremental results might be provided, until a final recognition hypothesis is returned.
`Skeleton2D`	Encodes the 2D points of the the detected skeleton.
`SoftBiometrics`	Describes soft biometrics (age and gender), with associated levels of confidence.

are being used at the same time) based on whatever prioritisation policy the developer defines. By decoupling the intent generation from the execution, this approach would allow easy comparison of different supervision approaches and to benchmark the resulting robot behaviour.

Intents are designed to eventually generate robot actions: as such, interactions with no side-effect on the robot and short reactive responses, should probably not rely on them. For example, asking the robot the weather (which can be addressed by querying an online weather forecast engine and does not require any specific robot resources); or checking the battery level (has no impact on the robot state or resources).

An `intent` is defined by the following fields:

- the `intent`, one of the available predefined intents
- `data`, a JSON object containing the data required to fully instantiate the intent. The admissible data *key* are based on the thematic roles associated to each intent
- `source` of the intent (e.g. a user or the robot itself)
- `modality` by which the intent was conveyed to the robot
- `priority` of the intent (optional)
- `level of confidence`, a value between 0 and 1.

5 VISUALISATION TOOLS

Besides providing standard interfaces to homogenise communication and data exchange between modules in different system architectures, it is equally important to provide tools that support understanding and analysis of the system performance. As such, by sharing a common framework, developers can implement common debugging and visualisation tools, which is crucial for open-source ecosystems evolution and improvement.

The following ROS4HRI-based visualisation and debugging tools are available:

`Humans`: an `rviz` plugin to visualise human perception features over ROS image streams. Mainly: faces bounding box, faces landmarks, bodies bounding box and 2D skeletons joints. This tool is part of the `hri_rviz` package.

`Skeletons 3D`: inspired by RobotModel, an `rviz` plugin to visualise the kinematic model of multiple detected bodies represented through 3D skeletons according to their joint states, position and body orientation. This tool is part of the `hri_rviz` package.

`TF (HRI)`: re-implementation of the original TF tool, provides visualisation of human-related frames associated with currently detected faces, bodies, voices or persons. Whenever one of these is no more detected (therefore, its ID is no more present among those published in the `/humans/*/tracked` topic), all the related frames are immediately removed from the scene. This tool is part of the `hri_rviz` package.

`rqt_human_radar`: an `rqt` plugin offering a radar-like view of the robot's surrounding scene displaying humans according to their position w.r.t. to the robot's reference frame (e.g., the base link). It is planned to extend this tool to visualise additional information regarding a person (e.g., engagement status).

Any information regarding humans published under the ROS4HRI topics and messages format standard can be visualised using the described tools (Fig. 3). Developers can compare the output of other ROS4HRI based systems with those they are implementing, in order to debug or evaluate the different performances.

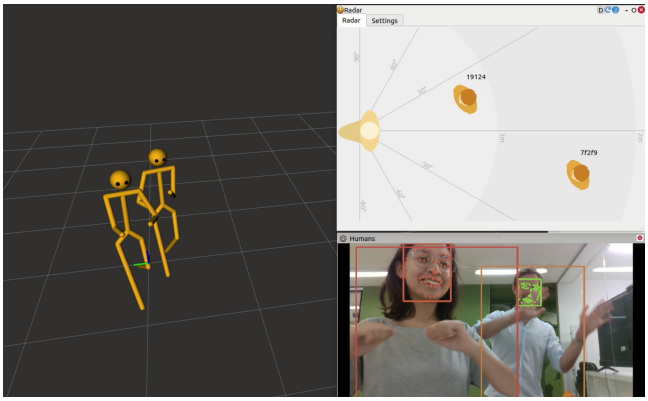

Figure 3: ROS4HRI visualisation tools: (left) `Skeletons 3D`**, (top right)** `rqt_human_radar` **and (bottom right)** `Humans`**.**

6 CONCLUSION AND NEXT STEPS

We have presented an overview of the ROS4HRI framework and its latest improvements with the aim of extending current specifications to better suit the community's needs. The advantages of sharing such open-source standard are not only limited to opening collaboration opportunities among researchers relying on a common framework, but also to provide tools to foster benchmarking, homogeneous data collection, and analysis of differentiated approaches, but built upon accepted standards, among others.

We are currently developing the cognitive architecture of ARI [1], a social robotic platform developed by PAL Robotics. It is designed as an anthropomorphic robot, with a focus on social interaction (Fig. 4). First launched in 2019, it has been deployed since in numerous research projects linked to social HRI, with a particular focus on elderly care and assistive robotics (both in EU projects: SHAPES, SPRING, TALBOT, PRO-CARED; and national ones: NHOA, RAADICAL, AMIBA).

We are building ARI's cognitive architecture on top of the ROS4HRI framework to implement its social interactive capabilities. The main pipelines are that of (1) human perception, aimed at effectively extending the awareness of the robot with respect to the users around it and thus become more alert and proactive towards the users' needs; and (2) the natural language processing pipeline, where evaluating diverse models for voice recognition, voice separation, sound source localisation on the one hand, and dialogue management on the other hand, are the key components to achieve satisfactory verbal interactions with users.

We plan to continue contributing to the ROS4HRI effort in the coming months, with a focus on:

(1) Additional tooling to easily visualise and record human-robot interactions;
(2) Additional social capabilities showcasing the advances of the ROS4HRI framework (e.g., emotion recognition, gesture recognition, voice identification);
(3) Further engagement with the community to make use of the framework to continue building open, shared and sustainable systems;

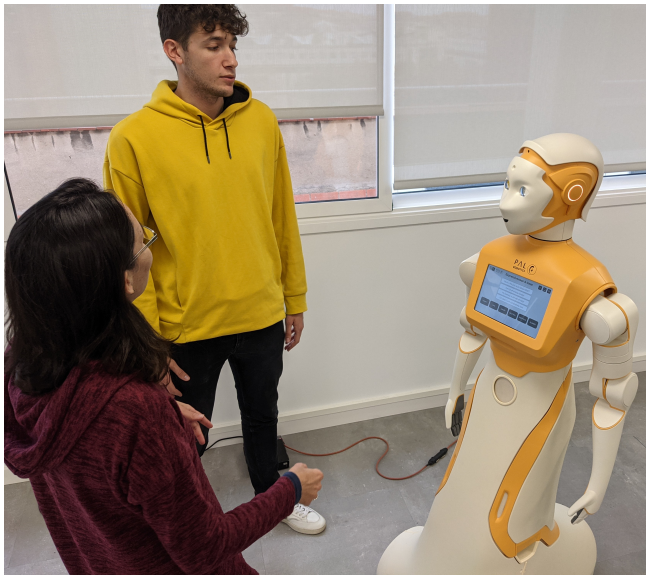

Figure 4: The ARI robot, interacting with two persons

(4) Porting of ROS4HRI to ROS2, with the aim of satisfying near future requirements from the ROS2 community.

7 ACKNOWLEDGMENTS

This work was partially supported by the H2020 SPRING project (grant agreement no. 871245), the H2020 ACCIO TecnioSpring INDUSTRY (grant agreement no. 801342, projects TALBOT and PRO-CARED), the H2020 PERSEO project (grant agreement no. 955778) and the NHOA and RAADICAL national project.

REFERENCES

[1] Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: The social assistive robot and companion. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 745–751.
[2] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
[3] P. Ekman and W. Friesen. 1978. Facial Action Coding System: A Technique for the Measurement of Facial Movement. (1978).
[4] IEEE. 2023. Robots Guide: NAO. *https://robots.ieee.org/robots/nao/* (2023).
[5] Y. Mohamed and S. Lemaignan. 2021. ROS for Human-Robot Interaction. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*. https://doi.org/10.1109/IROS51168.2021.9636816
[6] Amit Kumar Pandey and Rodolphe Gelin. 2018. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine* 25, 3 (2018), 40–48.
[7] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.