Socially-acceptable Extended Reality Models and Systems

D4.3 First interim release of the nonverbal communication software module

30 March 2024

Funded by the European Union

DELIVERABLE INFORMATION	
Deliverable leader	SUPSI
Document type	OTHER
Document code	D4.3
Document name	First interim release of the nonverbal communication software module
Work Package / Task	WP4 / T4.3, T4.4
Delivery Date (DoA)	30 March 2024
Actual Delivery Date	30 March 2024
Reviewers	T. Tran (TUDa) ...

DELIVERABLE HISTORY			
Date	**Version**	**Author**	**Summary of main changes**
30 Jan. 24	0.1	A. Paolillo (SUPSI) S. Arreghini (SUPSI)	Drafting the ToC and general concepts
8 Mar. 24	1.0	A. Paolillo (SUPSI) S. Arreghini (SUPSI) T. Thy (TUDa) M. Conci (SPXL)	Filling the content.

DISSEMINATION LEVEL		
PU	Public	**x**
SEN	Sensitive, limited under the conditions of the Grant Agreement	
RE	Restricted to a group specified by the consortium (including the EC services)	
CO	Confidential, only for the members of the consortium (including the EC)	

SERMAS partners

Disclaimer

This project has received funding from the Horizon Europe programme under the Grant Agreement 101070351.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the European Commission can be held responsible for them.

SERMAS • Grant Agreement: 101070351 • 2022 – 2025 | Duration: 36 months

Topic: HORIZON-CL4-2021-HUMAN-01-13

Public Executive Summary

This document presents the development and performances of the different modules related to non-verbal communication, their implementation within the SERMAS Toolkit, how they communicate with other modules, as well as some final ethical implications and privacy concerns.

The initial explanation starts from the non-verbal perception capabilities of the system and from the intention to interact detector.
This module can compute the future interaction probability of people (and potential SERMAS toolkit users) who are in front of an RGBD camera. This is achieved by using Machine Learning methods which take in input body motion information and facial features (such as the user gaze) and outputs the predicted probability that a specific user must interact in the future with the system on which the camera is mounted.
Some important aspects regard the data collection done for the algorithms training, architecture selection and performance evaluations.
One of the main performance points is the operating range. The module has been shown to retain good performances even at distances up to 5 m from the sensor. Indeed, this solution was designed for long range applications to provide meaningful proactive interactions with acceptable advance time.

The document then covers some research done on the non-verbal communication capabilities of a robotic platform. Following the intention to interact detector, we devised a reaction strategy to provide nice human-robot interactions with people, in this case offering a chocolate treat to people passing in the proximity of the system.

The software architecture behind these solutions is carefully described, detailing both the specific parts of the pipeline, how they communicate with each other, and the integration activities planned to use these packages as part of the SERMAS Toolkit.

Lastly, some ethical considerations are made regarding users' privacy during both the data gathering phase and system deployment.

Table of contents

List of Figures

List of Tables

Introduction

This deliverable aims to present the development done in the context of non-verbal communication between a potential user and the specific physical embodiment which makes use of the SERMAS Toolkit.

Nonverbal communication is a fundamental component of HRI for humans and robots [3], [4]. More specifically, we focus on the user's intention recognition and the reaction to the perceived user intentions.

The developments described in this deliverable are also reported in two conference paper

- "Predicting the Intention to Interact with a Service Robot: the Role of Gaze Cues", IEEE Int. Conference on Robotics and Automation 2024 [12]
- "A long-range mutual gaze detector for HRI," in ACM/IEEE Int. Conf. on Human-Robot Interaction, 2024 [2]

The overall situation can be described as in the block diagram in Fig. 1. The nonverbal communication between the user and the agent happens through two main channels: actuation and perception. For example, the agent can perceive the user's intention to interact, and actuate a proactive behavior in order to realize a successful interaction. Possibly, useful information exchanged in nonverbal communication can be shared externally, e.g. to be used in the dialogue system. For more details about this scheme, one could refer to D4.1.

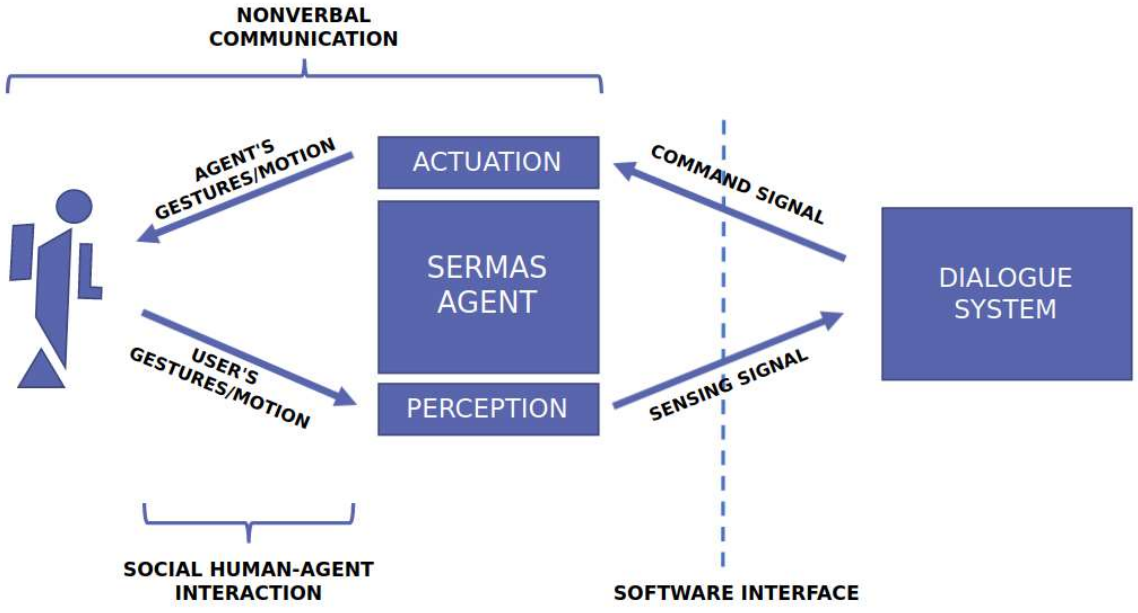

Fig. 1. High level system architecture and modules communication.

Perception capabilities

Different perception modules have been designed and combined to create a user data detection pipeline which can provide at the end information of different type such as user detection, body tracking, facial landmarks tracking and user intention recognition. These activities implied both the use of off-the-shelf algorithms, and the development of customized frameworks. The user detection and tracking are assumed to come from the chosen sensor software development kit and therefore will not be discussed in depth in this document. Examples of such sensors which offer a similar functionality are the Microsoft Azure Kinect DK and the Stereolabs ZED 2.

Indeed, the extraction of useful information about the users' behavior is not an easy task in HRI, especially when dealing with nonverbal communication [5]. A body of work aims at estimating the human intention, e.g. in the context of navigation [6], collaborative tasks [7], [8], or for social behavior interpretation [9]–[11].

Intention to interact classification

For a social entity, it is crucial to perceive people's intentions as early as possible, for example that an approaching person intends to interact. In such case, it can proactively enact friendly behaviors that lead to an improved user experience. Consider a robot stationery service in a public space, awaiting a possible user assistance request. The robot is assumed to be equipped with exteroceptive sensing capabilities, e.g. provided by an RGB-D sensor. The representative frame of the robot is chosen at its camera sensor frame and is denoted with Fs. People freely move in the environment, i.e., they can randomly enter or exit the scene, and possibly interact with the robot. The information about a person's behavior is described by properly chosen body frames and facial landmarks. The body frames of interest are one located in the middle of the person's chest (denoted with Fc) and on the person's head (Fh). We assume that such metric information, indicative of the proxemics of the subject, is measurable by the robot sensor. Furthermore, we also assume that the camera RGB images allow the detection of facial landmarks, which mainly consist of the projected locations, on the image plane, of

specific points of interest on the person's face. Multiple facial landmarks are detected at once with each landmark containing the 2D point coordinates on the image with an additional component corresponding to the depth of the landmark w.r.t. the face center of gravity. A scheme of the user data perception pipeline can be seen in Fig. 2.

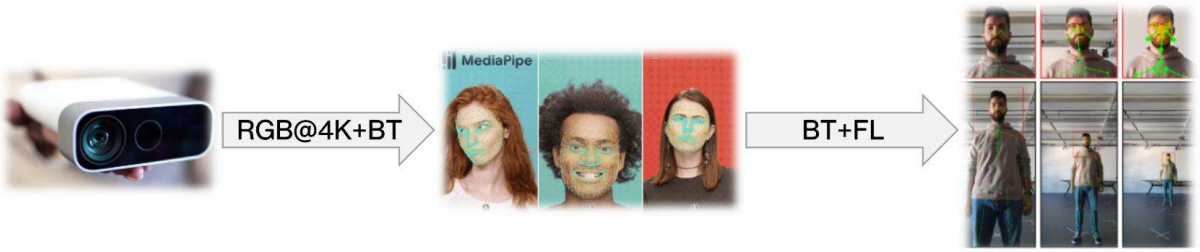

Fig. 2. System architecture of the user data perception pipeline. BT stands for Body Tracking, FL for Face Landmarks.

The facial landmarks, together with the body information of the subject's chest and head, are fed to a pre-trained mutual gaze classifier that outputs a score representing the probability that the subject is looking at the camera. Finally, the subject's intention to interact is indicated with y. The pipeline scheme can be seen in Fig. 3: our system is a classifier that, given the information about a potential user, provides an estimate y of its probability of interaction with the robot. The approach relies on (i) existing modules providing information on people motion (the sensor software development kit), and (ii) another specifically designed classifier, which computes the mutual gaze. Multiple users are supported.

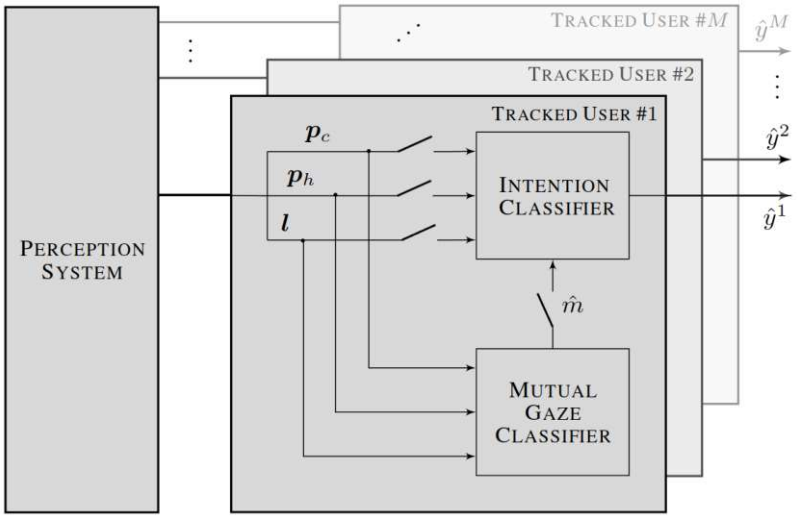

Fig. 3. Interaction intention detection pipeline architecture.

We expect that the motion dynamics of body frames, facial landmarks, and the temporal evolution of the mutual gaze are important cues to predict whether a given person intends to interact with the robot. To capture these dynamics, we use a recurrent Long Short-Term Memory (LSTM [1]) neural network as a stateful sequence-to-sequence classifier. We use the implementation available in the PyTorch5 library. To offer a more complete analysis, we compare the LSTM performance against a simpler, stateless model, i.e. a Random Forest (RF) classifier implemented using the scikit-learn package.

All the models are evaluated using a 5-fold stratified cross-validation strategy; folds are computed by splitting the set of sequences in training and testing sets, such that all frames for a given sequence stay in the same set.

In our setup, some feature sets include mutual gaze information through the variable “m”. Such information is extracted using a public mutual gaze detector implementation, see [2]. We first evaluate the algorithm performances frame-by-

frame using the *AUROC* metric, comparing the two different model architectures (RF and LSTM).

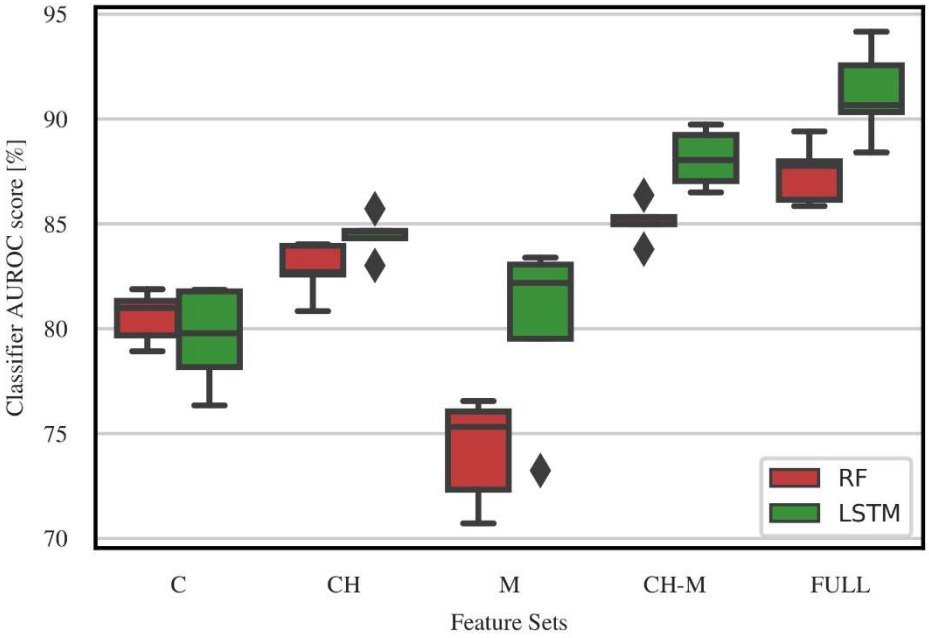

Fig. 4. AUROC of RF and LSTM classifiers with the different feature sets.

The plot in Fig. 4 shows that the stateful LSTM classifiers consistently outperform the simpler stateless RF counterparts, with the different test input sets.
Therefore, any further analysis is restricted to the LSTM architecture with two different input features:

- CH, which contains only information about the 3D position and orientation of the user's chest and head;
- FULL, which contains all the information available in the CH input features plus the user's facial landmarks and mutual gaze information.

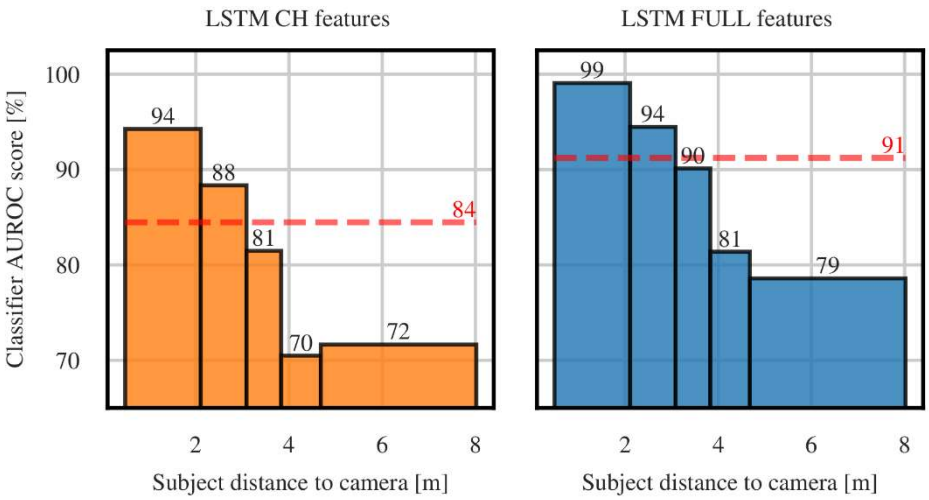

Fig. 5 AUROC for the LSTM using fCH (left) and fFULL (right) for different human-robot distance quantiles.

Fig. 5 reports an experiment in which we split the testing data into 5 distance quintiles and evaluate the classifier separately on each. All reported AUROC values are significantly greater than 0.5, which indicates that, even among subjects that are at approximately the same distance from the robot, the classifier is effective at differentiating those who are likely to interact and those who are not; i.e., even though subject distance from the robot is a powerful feature, it is not the only aspect that is considered by the models. The figure further shows that the improvement of the performance introduced by the contribution of the gaze is uniform across the whole range of subject-robot distances. In the bin of the closest distance (the one from 0.48 to 2.10 m), the gain in performance due to the additional gaze information and facial landmarks is about 5%. This improvement increases to 9% for the intermediate bin (containing distances from 3.08 to 3.83 m), and to 7% for the farthest distances (above 4.68 m).

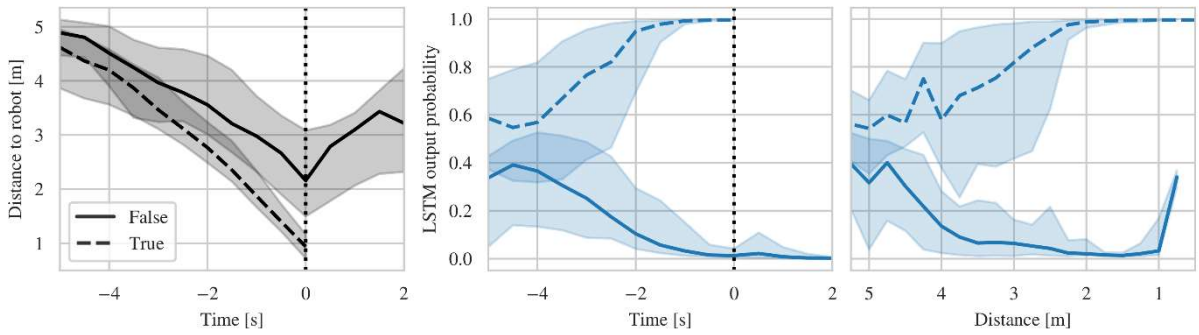

Fig. 6. Median distance to the robot (left) and median predicted probability of interaction (center) as a function of time. Time t = 0 is defined for each sequence as the moment when the subject either interacts, for positive

sequences (dashed line), or the moment in which the subject is closest to the robot, for negative sequences (continuous line). The rightmost plot reports the predicted probability of interaction as a function of distance to the camera, ignoring negative samples with $t > 0$. Shaded areas represent the interquartile range.

Fig. 6 shows the statistics from positive and negative sequences for the whole dataset. In the left and center plots, all sequences are temporally aligned in such a way that $t = 0$ denotes the time of interaction, in the case of positive sequences, and the time in which the subject is at the closest distance from the robot, in case of negative sequences. We observe from the left plot that negative sequences reach, on average, a distance from the robot of 1.6 m before moving further. The center plot shows that at $t = 0$ (vertical dotted line), the model yields very sharp predictions. The distribution of the output probabilities for positive and negative sequences starts diverging at $t = -4$ s, and clearly separate at $t = -3$ s. The rightmost plot reports the same data but with distance to the robot on the horizontal axis. Negative sequences that reach distances below 1 m are few, so the rightmost data is noisy.

We now report the performance of our models when evaluating them at the level of entire sequences. For each sequence, we simulate that the model is applied to each frame, and when exceeding a threshold τ, the robot takes an irreversible decision to enact a given behavior (e.g. facing, approaching or greeting the user). Negative sequences in which the output probability never exceeds τ are true negatives (i.e., the robot correctly ignored a non-interacting subject); positive sequences in which the output probability exceeds τ for at least a single frame are true positives; only for true positives, from the earliest frame whose classifier output exceeds τ , we compute the advance detection time (i.e. the amount of anticipation) and advance detection distance (i.e. the distance of the user when the robot reacted). False negatives denote sequences for which the robot did not react to an interacting user; false positives denote sequences in which the robot incorrectly reacted to a non-interacting subject. Given these definitions, we

compute sequence-level metrics: False Positive Rate (FPR), True Positive Rate (TPR), precision, recall and accuracy.

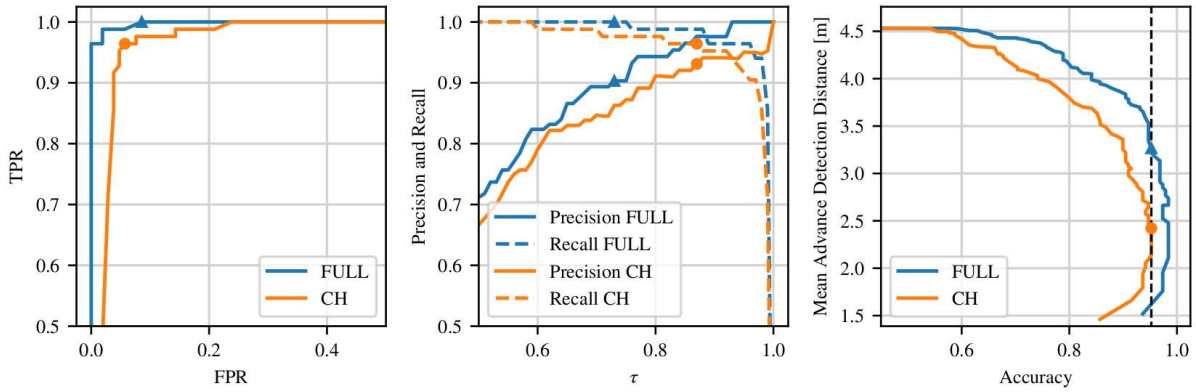

Fig. 7. Sequence-level performance metrics for the LSTM approach with (blue) and without (orange) gaze and face landmark features. Left: ROC curve. Center: Precision and Recall as a function of threshold τ. Right: Mean Advance Detection Distance (vertical axis) vs. achieved accuracy (horizontal axis) for different values of τ. The orange circles denote a threshold value of $\tau = 0.87$ for the baseline model set to achieve maximum accuracy. Conversely, the blue triangles denote a threshold value of $\tau = 0.73$ needed by our model to display the same level of accuracy.

Fig. 7 reports these metrics for both the baseline model (which uses fCH), in orange, and our model (that relies on the more complete information contained in fFULL), in blue. We observe that the latter outperforms the former in all metrics, regardless of the threshold value τ. In particular, the center plot highlights that high threshold values are key to obtaining very high precision and recall performance, with our model consistently outperforming the baseline in both metrics. Nevertheless, the plot does not show the complete picture: high threshold values yield high performance because, in this case, the model does not commit to a decision until very late in the sequence, when most positive sequences yield very high probabilities; this behavior is not useful in practice. The right plot in Fig. 7 studies the trade-off between sequence classification accuracy, on the horizontal axis, and mean advance detection distance, on the vertical axis, controlled by τ. Low values of τ yield early, distant but inaccurate detections (top left). Increasing τ decreases the mean Advance Detection Distance but improves accuracy up to a maximum value; further increases of τ lead to a marked increase in false negatives, and negatively impact both Advance Detection Distance and Accuracy, as can be also seen from the drops in the recall value. The orange dots denote a threshold ($\tau = 0.87$) yielding maximum accuracy (95.2%) for the baseline

classifier, and the blue triangles denote the threshold ($\tau = 0.73$) needed to get to the same accuracy for our model. At this threshold, our model yields a significantly better advance detection distance (3.27 m) w.r.t. the baseline (2.42 m): an improvement of 0.85 m.

Actuation capabilities

As of now, the reaction to perceived user intentions is purely based on the non-verbal intention detection described in the previous section. Further integration and development will investigate the possibility of integrating also verbal information.

Reaction to user intentions

Following the user interaction intention detection, the robot can react to this perceived need to offer a more custom and acceptable interaction.
An example of such interaction could be offering chocolate treats to people passing by if they are perceived as willing to interact.

Fig. 8. Interaction setup.

In this application, portrayed in Fig. 8, the robot stays still on the table in a resting position, with the arm aligned with the sensor's forward direction, the arm retracted, and LEDs turned off. Upon triggering, an offering motion can be initiated and consists in the robot turning on the spot toward the selected person, lighting up its LEDs to signal its activation, and reaching out with the arm.

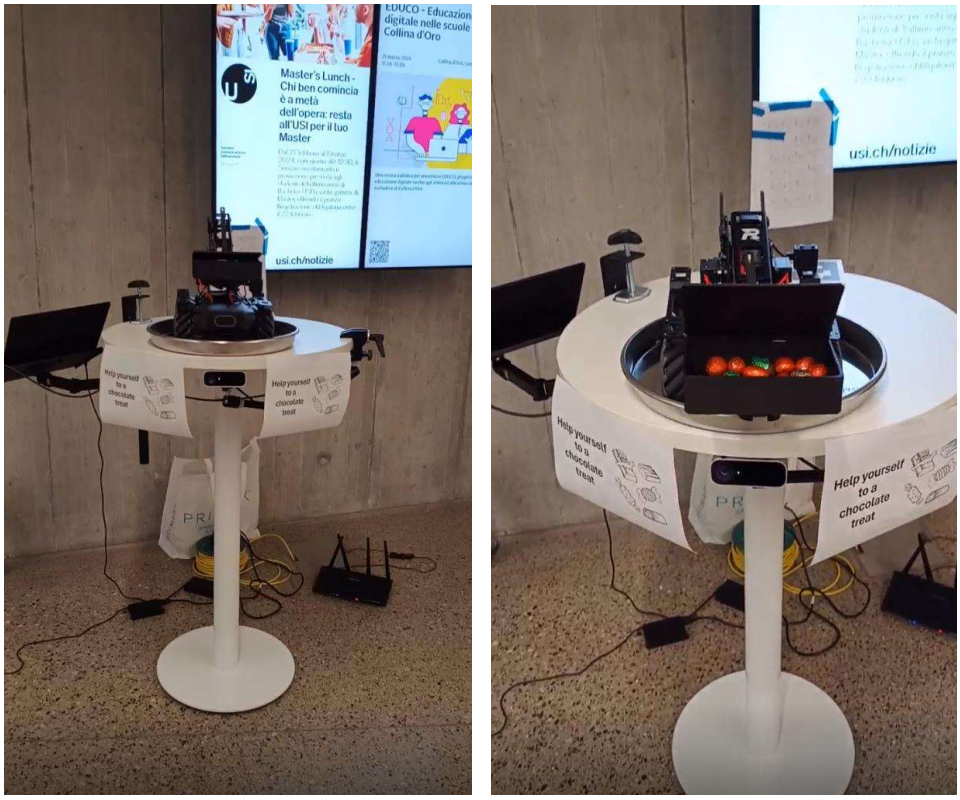

Fig. 9. Updated version of the waiter robot. On the left the rest position, on the right the offering position with the tray open and chocolate treats visible.

The evolution of this setup can be seen in Fig. 9. This updated version has a better offering setup with a custom tray with a lid that opens when the robot extends the arm.

This setup will be used in future experiments to validate the user perception module in real-world scenarios as well as quantify the impact of a proactive agent behavior on people.

Command and sensing signals

The textual dialogue module developed in WP5 takes non-verbal signals as textual descriptive input prepended by a special token indicating the type of non-verbal input. An example of the non-verbal input from the user emotion recognition module could be demonstrated as follows: `<emotion> happy`. The non-verbal communication signals considered in the dialogue module includes user emotions and gestures, robot actions in response to users (see D5.1 for details).

Software release

Perception modules

In the next section, there will be an explanation of the different nodes which are integral parts of the perception modules.

User Data Perception Pipeline

A representation of the user perception pipeline can be seen in Fig. 10 with a more high-level conceptual map available in Fig. 2.

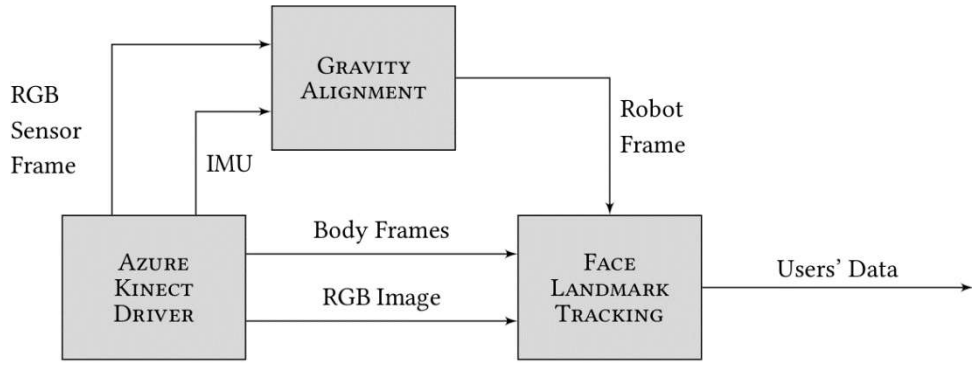

Fig. 10 User data perception pipeline nodes architecture.

1.1.1.1. Azure Kinect ROS Driver

This node runs the driver of the RGB-D sensor and the body tracking offered by its SDK. It publishes the following information: (i) the raw signals of the Kinect's onboard Inertial measurement unit (IMU); (ii) the RGB image streams; (iii) the body frames from the sensor SDK; and (iv) the RGB camera frame. The frame information is standardized in ROS2 as tf messages.

1.1.1.2. Gravity Alignment

This node calculates the gravity-aligned robot frame. It takes the IMU data to calculate the difference of the RGB sensor frame orientation w.r.t. the inertial vertical direction and broadcasts the aligned robot frame. Ultimately, the robot frame is defined as the one centered in the origin of the RGB sensor frame constrained vertically to be aligned with gravity and horizontally with the camera heading.

1.1.1.3. Face Landmarks Tracking

This node is the core of the perception pipeline and implements the face landmarks extraction. It takes as input the user's body and robot frames information and the current RGB image. Firstly, it performs projection of the users' 3D head positions onto related 2D points on the image plane. This information is used to crop the regions of interest of the RGB image corresponding to the faces of the detected users. Such cropped images are fed to multiple instances of the MediaPipe Face Mesher, which detects the face landmarks for each user. This crucial step allows us to overcome the distance range limits of the MediaPipe implementation. Exploiting the robust detection of the head frames provided by the Azure Kinect SDK, we can allow the face landmarks detection by MediaPipe at distances further than 2 m. To be independent of the camera orientation, the body frame poses, originally expressed in the RGB sensor frame, are transformed into the gravity-aligned robot frame. The detected face landmarks and the transformed body frame poses of the detected users are finally time synchronized and published as a custom ROS2 topic. Such message is called Users' Data in the scheme of Fig. 10.

Mutual Gaze Classifier

This node is taken as from the public implementation available online here:

https://github.com/idsia-robotics/mutual_gaze_detector/tree/hri

This node takes the users' data message in input and outputs the probability of mutual gaze for each person currently tracked by the camera.

Interaction Intention Classifier

This node is a ROS2 wrapper for the actual Pytorch implementation of our classifier. It takes as input the user data custom topic provided by the Face Landmark Tracking node. As output, it publishes a simple custom topic that contains the IDs of the detected users and the corresponding probability of interaction as computed by the classifier. This node can also run a GUI showing the real-time evolution of the predicted probability related to the user who has been tracked for the longest time, for visualization purposes.

Integration activities

All the perception modules and different nodes explained above can be run within a Docker container for easy and quick deployment on multiple platforms.
Afterwards, the integration of the perception modules concerns the capability to exchange data between the ROS2 environment and the SERMAS toolkit, following the API specification as reported in D6.2. To achieve this, a ROS2 proxy node has been developed as a reusable tool for connecting ROS2-based systems to the SERMAS toolkit.
It allows the communication between ROS2 nodes and the SERMAS toolkit by converting and forwarding the data exchanged between the two entities.
More specifically, the proxy node implements an HTTP client and a MQTT client that automatically connect and establish a secure communication with the SERMAS toolkit by authenticating using a client ID and secret credentials, which allow the node to publish and/or subscribe to the topics associated to a SERMAS application.
For each topic the communication happens in two possible directions, the proxy:

1. subscribes to the ROS topic, convert the payload and publish the data to the SERMAS topic
2. subscribes to the SERMAS topic, convert the payload and publish the data to the ROS topic

The proxy node is built within a Docker container and started alongside the perception container(s) with Docker Compose.
Please refer to D6.1 for more details about human-agent interaction and how the modules developed in WP4 are integrated in the agent architecture.

Ethical implications and privacy aspects

The ethical and privacy concerns should be evaluated in the two different phases: training and deployment.

During the first phase, there has been a procedure to gather and save data to be used for the training of the artificial intelligence components. All the participants in the training data acquisition have therefore signed a consent form (or their tutors), with all gathered data kept for private use only. The experiments were also approved beforehand by the local institution's ethical committee.

The people who participated in the data gathering come from several users which include adults and also a group of kids.

Nonetheless the data saved are in the form of anonymized body joint information and anonymized facial landmarks on which the models are trained.

In the deployment phase, the different components are used "as is" with no online changes. Therefore, even though user-specific private data (such as body joint tracking or facial landmarks) is extracted from the user, it is not saved nor shared with anyone.

Conclusion

We have presented the results of our non-verbal communication module. Firstly, in the perception field, understanding when someone wants to interact with the system, and in a successive stage in the robot movements field, devising some task dependent robot reactions triggered the perceived user need.
The modules have been developed to allow long-range operation and therefore the most advanced time possible. Having a higher advance time for the detection of a user in need is crucial. This allows early reaction from the system useful in providing targeted and possibly more socially acceptable interactions.
In the next stage, the developed modules will be tested and validated in the wild studying how different reactions or triggering threshold might affect human behaviors towards an autonomous system.

References

1. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
2. S. Arreghini, G. Abbate, A. Giusti, and A. Paolillo, “A long-range mutual gaze detector for HRI,” in ACM/IEEE Int. Conf. on Human-Robot Interaction, 2024, pp. –.
3. N. Gasteiger, M. Hellou, and H. S. Ahn, “Factors for personalization and localization to optimize human–robot interaction: A literature review,” International Journal of Social Robotics, pp. 1–13, 2021
4. S. Saunderson and G. Nejat, “How robots influence humans: A survey of nonverbal communication in social human–robot interaction,” International Journal of Social Robotics, vol. 11, pp. 575–608, 2019.
5. J. Rios-Martinez, A. Spalanzani, and C. Laugier, “From proxemics theory to socially-aware navigation: A survey,” International Journal of Social Robotics, vol. 7, pp. 137–153, 2015
6. P. Agand, M. Taherahmadi, A. Lim, and M. Chen, “Human Navigational Intent Inference with Probabilistic and Optimal Approaches,” in IEEE Int. Conf. on Robotics and Automation, 2022, pp. 8562–8568
7. A. Belardinelli, A. R. Kondapally, D. Ruiken, D. Tanneberg, and T. Watabe, “Intention estimation from gaze and motion features for human-robot shared-control object manipulation,” in IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2022, pp. 9806–9813
8. S. Vinanzi, C. Goerick, and A. Cangelosi, “Mindreading for Robots: Predicting Intentions via Dynamical Clustering of Human Postures,” in Int. Conf. on Development and Learning and Epigenetic Robotics, 2019, pp. 272–277
9. A. Zaraki, M. Giuliani, M. B. Dehkordi, D. Mazzei, A. D’ursi, and D. De Rossi, “An RGB-D based social behavior interpretation system for a humanoid social robot,” in RSI/ISM International Conference on Robotics and Mechatronics, 2014, pp. 185–190
10. A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, “Social behavior recognition using body posture and head pose for human-robot interaction,” in IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2012, pp. 2128–2133

11. F. Del Duchetto, P. Baxter, and M. Hanheide, “Are you still with me? continuous engagement assessment from a robot’s point of view,” Frontiers in Rob. and AI, vol. 7, p. 116, 2020
12. S. Arreghini, G. Abbate, A. Giusti, and A. Paolillo, “Predicting the intention to interact with a service robot: the role of gaze cues,” in IEEE Int. Conf. Robot. and Autom., 2024, pp. –.