Romeo2 Project: Humanoid Robot Assistant and Companion for Everyday Life: I. Situation Assessment for Social Intelligence [1]

Amit Kumar Pandey[2], Rodolphe Gelin[3], Rachid Alami[4], Axel Buendia[5], Mohamed Chetouani[6], Laurence Devillers[7], David Filliat[8], Yves Grenier[9], Abderrahmane Kheddar[10], Frédéric Lerasle[11], Mounira Maazaoui[12], Roland Meertens[13], and Renaud Viry[14]

Abstract. Situation assessment is one of the basic abilities for robots to co-exist with us in our day-to-day live. For a socially intelligent robot, different levels of situation assessment are required, ranging from basic processing of sensor input to high-level analysis of semantics and intention. The combination of various perception abilities greatly increases the robot's socio-cognitive capabilities. However, this prompts new research challenges and the need of a coherent framework and architecture.

Romeo2 is a unique project, aiming to bring multi-modal and multi-layered perception of situation assessment on a single system and targeting for a unified theoretical and functional framework for robot companion for everyday life. This paper presents different aspects of situation assessment identified and perceived within the Romeo2 project. It aims towards a principled approach to develop different components in a collaborative manner when such basic blocks should be functioning together and discusses about some of the innovation potentials such approach brings for the companion robotics domain.

1 Introduction

As robots started to co-exist in a human-centered environment, human awareness capability is important to be considered. With safety being a basic requirement, such robots should be able to behave in a socially accepted and expected manner. This requires robots to reason about the situation, not only from the perspective of physical locations of objects, but also from that of 'mental' and 'physical' states of the human

[1] This work is funded by Romeo2 project, (http://www.projetromeo.com/), BPIFrance in the framework of the Structuring Projects of Competitiveness Clusters (PSPC)

[2] Aldebaran Robotics, A-Lab; akpandey@aldebaran-robotics.com
[3] Aldebaran Robotics; rgelin@aldebaran-robotics.com
[4] CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France; rachid.alami@laas.fr
[5] Spirops / CNAM (CEDRIC), Paris; axel.buendia@cnam.fr
[6] ISIR, UPMC; mohamed.chetouani@upmc.fr
[7] Limsi/CNRS UniversitÃĭ Paris-Sorbonne IV, Paris; devil@limsi.fr
[8] ENSTA ParisTech - INRIA FLOWERS; david.filliat@ensta-paristech.fr
[9] Inst. Mines-Télécom; Télécom ParisTech; CNRS LTCI; yves.grenier@telecom-paristech.fr
[10] CNRS-UM2 LIRMM IDH; kheddar@gmail.com
[11] LAAS-CNRS; frederic.lerasle@laas.fr
[12] Institut Mines-Télécom; Télécom ParisTech; CNRS LTCI; maazaoui@telecom-paristech.fr
[13] Spirops, Paris; rolandmeertens@gmail.com
[14] CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France; renaud.viry@laas.fr

partner. Further, such reasoning should result into knowledge building with the human understandable attributes, to facilitate natural human-robot interaction.

The Romeo2 project, the focus of this paper, is unique in that it brings together different perception components in a unified framework for real-life human-robot interaction scenarios. This paper outlines our multi-layer perception architecture, the categorization of basic requirements, the key elements to perceive, and the innovation advantages such a system provides.

1.1 The Romeo2 Project

Romeo2, as a project (project website [25]), aims at achieving a practical and real-life personal assistant and companion robot in an everyday scenario. There are 17 partners: *Aldebaran, SpirOps, INRIA, ALL4TEC, CNRS-LAAS, VOXLER, CNRS-LIMSI, CNRS-LIRMM, CEALIST, College de France, Armines-ENSTA, ISIR, Telecom ParisTech, University of Versailles, Strate* and *Approche*. Most of the partners are actively involved in one or the other aspect of perceiving the situation, ranging from sensor level signal processing to building a high-level multi-modal representation describing and predicting the situation. The robot platform is the *Romeo* robot, developed by *Aldebaran Robotics*, in collaboration with various academic and industrial partners. Fig. 1 shows the robot and placements of rich range of sensors. Romeo is a *40kg*, *1.4m* humanoid robot with *41 degrees-of-freedom*, *vertebral column*, *exoskeleton on legs*, *partially soft torso* and *mobile visual cameras*.

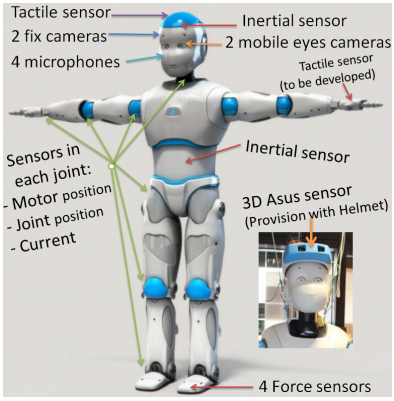


Fig. 1. Romeo robot's exteroceptive and proprioceptive sensors.

1.2 An Example Scenario

Mr. Smith lives alone (in fact not really, but with his Romeo robot companion). He is elderly and visually impaired. Romeo assists him in daily-life tasks. The robot understands his speech, emotion and gestures. Provides physical support by bringing the 'desired' items. Offers cognitive support by reminding about medicine, items to add in to-buy list, playing memory games, etc. As a social inhabitant, it plays with Mr. Smith's grandchildren who visit during the weekends. A playing child bumps into and the robot says, "Sorry", and moves away to provide playing space. It monitors Mr. Smith's activities and calls assistance when abnormalities are detected in Mr. Smith's behaviors.

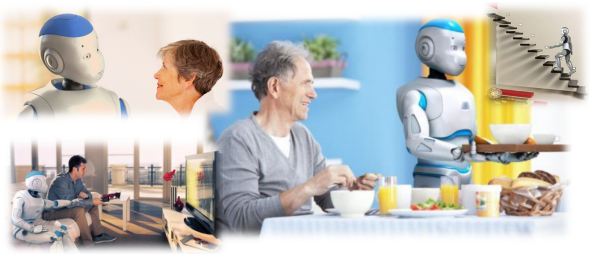


Fig. 2. Romeo2 Project scenario: A Humanoid Robot Assistant and Companion for Everyday Life.

This outlined target scenario of Romeo2 project, also partially illustrated in fig. 2, depicts that being aware about human, his/her activities, the environment and the situation are key aspects towards practical achievement of the project's objective.

1.3 Situation Awareness and Assessment

Situation awareness, or the ability to perceive and abstract information from the environment [2], is an important capability humans use to perform tasks effectively [10]. Three levels of situation awareness have been identified in Endsley *et al.* [11]:

- *Level 1 situation awareness*: To perceive the state of the *elements* composing the surrounding environment.
- *Level 2 situation awareness*: To build goal oriented understanding of the situation, by using level 1 situation awareness. Therefore, experience and comprehension of the meaning are important factors in level 2 situation awareness.
- *Level 3 situation awareness*: To project future. Uses the knowledge of status of the environmental elements (level 1) and understanding of the situation (level 2).

In general, *Situation Awareness* is viewed as the state of knowledge and *Situational Assessment* is seen as the processes used to achieve that knowledge. Moreover, the levels are not seen separated, instead the higher level incorporates the lower levels.

1.4 Related Works and Contributions

Already, researchers (including the partners of the Romeo2 project) are working on different aspects of situation assessments, ranging from geometric level of object and location identification to recognizing the emotion and intention of the human partner. Situation awareness is an important aspect of day-to-day interaction, decision-making, and planning, so as important is the identification of the *elements* and *attributes*, constituting the state of the environment, which in fact are domain dependent.

There have been efforts and work to integrate and utilize more than one component of perception and situation assessment. But most of them are specific for a particular task like navigating [23], intention detection [18], robot's self-perception [5], spatial and temporal situation assessment for robot passing through a narrow passage [1], laser data based human-robot-location situation assessment, e.g. human entering, coming closer, etc. [14]. Therefore, they are either limited by the variety of perception attributes, sensors or restricted to a particular perception-action scenario loop. On the other hand, various projects on Human Robot Interaction try to overcome perception limitations by different means and focus on high-level semantic and decision-making. Such as, detection of objects is simplified by putting tags/markers on the objects, detection of people without any audio information, [6], [16]. In [12], different layers of perception have been analyzed to extract information, useful to build representations of the 3D space, but focused on eye-hand coordination for active perception and not on high-level semantics and perception of the human.

In Romeo2 project, we are making effort to bring a range of multi-sensor perception components within a unified framework, at the same time making the entire multi-modal perception system independent from a very specific scenario or task, towards realizing effective and more natural multi-modal human robot interaction. In this regard, to the best of our knowledge, Romeo2 project is the first effort of its kind for a real world companion robot. In this paper, we do not provide the details of each component. Instead, we give an overview of the entire situation assessment system in Romeo2 project, the key elements and attributes to be perceived from robot companion domain, and how they are fitting in the different layers and components of perception in a unified framework. Interested readers could find the details in documentation of the system [20] and in dedicated publications for individual components, such as [4], [13], [21], [17], [3], [8], [19], (see the list of publications [25]). At the same time this global view of the project's perception system, helps us to identify some of the innovation potentials and develop them, as described in section 5.

***Outline*:** Next section first presents our layered architecture for sense-interaction loop of perception, sec. 2.1. Then we categorize different basic requirements and the

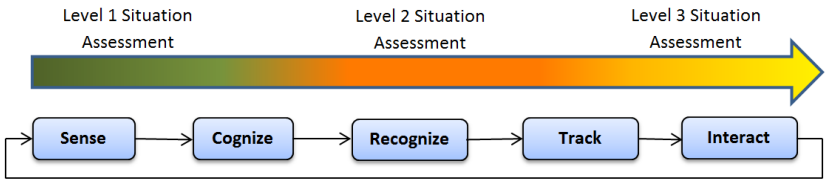


Fig. 3. A generalized perception system for sense-interact in Romeo2 project, with five layers functioning in a closed loop.

key attributes of a companion robot perception capabilities, followed by the developments in the Romeo2 project, sec. 2.2. In sec. 3, we discuss how our layered architecture and the carefully categorization of requirements help us in identifying dependencies and facilitate concurrent collaborative development. Sec. 4 presents the NAOqi framework facilitating to have a coherent functional framework. Then in sec. 5, we discuss some of the innovation potentials when such rich and diverse perception components are available on a single platform, followed by conclusion and future work.

2 Perceiving Situation in Romeo2 Project

2.1 A Generalized Perception Architecture for Sense-Interact for HRI

To better place the contributions of different partners in the Romeo2 project, we have adapted a simple yet meaningful, sensing-interaction oriented perception architecture, by carefully identifying various requirements and their interdependencies, as shown in fig. 3. The roles of the five identified layers are described next.

(i) Sense: Corresponds to receiving signals/data from various sensors. Depending upon the sensor type and their fusion, this layer can build 3D point cloud world; sense stimuli like touch, sound; know about the robot's internal states such as joint, heat; record speech signals; etc. Therefore, it belongs to *level 1* of situation assessment.

(ii) Cognize: Corresponds to the 'meaningful' information extraction from what all the robot is sensing. This can be learning shapes of objects, learning to extract semantics from 3D point cloud, meaningful words from speech, meaningful parameters in demonstration, etc. The definition of 'meaningful' in general refers to human-understandable level of meanings. However, it depends upon the domain, and the interaction requirement. Most of the times, this 'cognition' is provided a priori to the system e.g. the meaningful set of words, the types of objects, their shapes, associated semantics, etc. Therefore, in most of the perception-action systems, this *cognize* part is either absent or narrowed down to reacting to stimuli. However, in Romeo2 projects we are taking steps to make *cognize* layer more visible by bringing together different learning modules, such as to learn objects, learn faces, learn the meaning of instructions, learn to categorize emotions, etc. This layer lies across *level 1* and *level 2* of situation assessment, as it is building knowledge in terms of attributes and their values and also extracting some meaning, which are supposed to be used in future.

(iii) Recognize: Dedicated to recognize some previously perceived/cognized person, place, action, thing, etc. It is an important aspect for companion robot, e.g. recognizing an emotion of sadness, a place as kitchen, etc. Depending upon the context the recognition itself can be at different levels of abstraction. For example, detecting that there is a person or face in the environment and identifying a particular person, both are kinds of recognition. In one case it is recognizing a previously known category

(person) in which the perceived entity belongs to, in the other case it is to recognize a particular person known previously. In general, it corresponds to recognizing what has been 'cognized' by the system. This mostly belongs to *level 2* of situation assessment, as it is more on utilizing the knowledge either learned or provided a priori, hence 'experience' becomes the dominating factor.

(iv) Track: This layer corresponds to the requirement that the robot should not only be aware or recognize something, but also should be able to track it during the course of interaction. Tracking itself can be of various types and multi-modal, e.g. track a sound, an object, a person, etc. It can also be for various purposes, e.g. to track in a populated environment a particular person to interact, to track an object for visual servoing, etc. From this layer, *level 3* of situation assessment begins, as tracking allows to update in time the state of the beforehand entity (person, object, etc.) and need a 'projection'. Sometimes, for tracking requirements, the recognition can be overpassed, e.g. tracking something within a given bounding box. And sometimes for recognizing, a tracking system is required (mainly because of time taking processing of recognition system). Those situations are handled on case-by-case basis. Therefore, practically there is a kind of loop between tracking and recognition layers, which is not explicitly shown for the sake of making main idea of the architecture better visible.

(v) Interact: Corresponds to the high-level perception requirements for interaction (with human and environment). For example, activity prediction, perspective taking, social signal processing, gaze analysis, etc. to interact with the human. To interact with environment, semantic perception of objects and locations, e.g. objects which can be pushed/moved by the robot, objects on which someone can sit, etc. We put such perception reasoning at the top level, *interaction*, because it needs input from almost all the layers below and have specific additional analyses requirements for interaction. It mainly belongs to *level 3* of situation assessment, as involves 'predicting' side of perception, such as predicting the abilities and gaze of agents, affordances (to sit, push, etc.) in the environment, desire and intention of person, etc.

Note that the main novelty lies in the *closed loop* aspect of the architecture. As shown in some preliminary examples in section 5, such as Ex1, we are able to practically achieve this, which is important to facilitate natural human-robot interaction process, which can be viewed as: *Sense* → *Build knowledge for interaction* → *Interact* → *Decide what to sense* → *Sense* →...

2.2 Basic Requirements, Key Attributes and Developments in Progress

In Romeo2 project, we have identified the key attributes and elements of situation assessment, to be perceived from companion robotics domain perspective, and categorized along five basic requirements as summarized in table 1. As the functions of many of the listed modules are obvious, in this section we will briefly describe only those modules, which we think need some explicit descriptions. Interested reader can find more detailed descriptions and methods of most of these modules online [20].

I. Perception of Human

People presence: To perceive if there are people present in the robot vicinity or not and assign unique ID for each detected person.
Face characteristics: To predict age, gender and degree of smile on a detected face.
Face and person tracking: Tracks face, person's moving head, torso or whole body.

Table 1. Classification of key requirements, the different aspects and the perception oriented developments in Romeo2 project

(I) Perception of Human		(II) Perception of Robot Itself
(i) People Presence	(ix) Perspective Taking	(i) Battery Status
(ii) Face Detection	(x) Emotion Recognition	(ii) Body Temperature
(iii) Face Characteristics	(xi) Speaker Localization	(iii) Foot Status
(iv) Gaze Analysis	(xii) Speech Recognition	(iv) Robot Posture
(v) Face Recognition	(xiii) Speech Rhythm Analysis	(v) Fall Detection
(vi) Face and Person Tracking	(xiv) User Attention Detection	(vi) Self Collision Detection
(vii) Posture Characterization	(xv) User Profile Analysis	
(viii) Waving Detection	(xvi) Intention Analysis	

(III) Perception of Object	(IV) Perception of Environment	(V) Perception of Stimuli
(i) 3D Segmentation	(i) Landmark Detection	(i) Sound Detection
(ii) Barcode Reader	(ii) Darkness Detection	(ii) Chest Button Interpretation
(iii) Close Object Detection	(iii) Place Recognition	(iii) Movement Detection
(iv) Object Recognition	(iv) Location Tracker	(iv) Sound Localization
(v) Object Tracker	(v) Sound Tracker	(v) External Collision Detection
(vi) Semantic perception	(vi) Semantic Perception (place)	(vi) Contact Observer

Posture characterization (human): To find position and orientation of different body parts of the human, shoulder, hand, etc.
Waving detection: To detect if someone is waving his/her hand.
Perspective taking: To perceive reachable and visible places and objects from the human's perspective, with the level of effort required to see and reach.
Emotion recognition: To predict basic types of emotions anxiety, anger, sadness, joy, etc. based on multi-modal audio-video signal analysis.
Speaker localization: To localize spatially the person who is speaking.
Speech rhythm analysis: To analyzing the characterization of speech rhythm by using acoustic or prosodic anchoring, to extract social signals such as engagement, etc. Based on the concept of alternation in time of perceptual phenomena.
User attention detection: To detect the attention of the interacting person. Based on speech and head turning.
User profile: To generate emotional and interactional profile of the interacting user. Used to dynamically interpret the emotional behavior as well as to build behavioral model of the individual over a longer period of time.
Intention analysis: To interpret the intention and desire of the user through conversation. The module keeps a state of the dialog and can switch among different topic to talk. This enhances robot's ability to predict and help the user based on context. The context also helps other perception components about what to perceive and where to focus. Thus, facilitates closing the interaction-sense loop of perception of fig. 3.

II. Perception of Robot Itself

Fall detection: To detect (based on center of mass) if the robot is falling and to take some self-protection measures with its arms before touching the ground.

Other modules in this category are self-descriptive. However it is worth to mention that, such modules also provide symbolic level information, such as *battery nearly empty*, *getting charged*, *foot touching ground*, symbolic posture *sitting*, *standing*, *standing in init pose*, etc. All these help in achieving one of the aims of Romeo2 project: sensing for natural interaction with human.

III. Perception of Object

Object Tracker: It consists of different aspects of tracking, such as moving to track, tracking a moving object and tracking while the robot is moving.
Semantic perception (object): Extracts high-level meaningful information, such as object *type* (*chair*, *table*, etc.), *categories and affordances* (*sitable*, *pushable*, etc.)

IV. Perception of Environment

Darkness detection: Estimates based on the lighting conditions of the environment around the robot.
Sound tracker: To track a sound with the input distance (used to estimate sound position) and input confidence (used to filter sound location).
Semantic perception (place): Extracts meaningful information from the environment about places and landmarks (a kitchen, corridor, etc.), builds topological maps.

V. Perception of Stimuli

Contact observer: To be aware of desired or non-desired contacts when they occur, from interpreting information from various embedded sensors, such as accelerometers, gyro, inclinometers, joints, IMU and motor torques'. We use three levels of sensing: i) *Detect* whenever the robot enters in contacts with its surrounding (human or environment) on purpose (planed, e.g. take a support, grasp an object) or accidentally (not planed) or breaks a contact ii) *Locate* the contact spots, which limbs are eventually concerned, and where the contact is located on each limb, iii) *Measure/estimate* the contact forces if needed in the control (contact formation or contact avoidance).

3 Facilitating Collaborative Development

The two contributions of the paper as discussed above (our perception architecture and identification of the key perception elements and their categorization based on basic requirements) together facilitate different partners to identify the aspects to coordinate and complement each other.

Assume partner A has expertise in X. A identifies its contribution related to X from the list of requirements summarized in table 1. It finds the fit in requirement $I.(v)$ *Perception of Human: Face Recognition*. Then A identifies its fit within layers of architecture of fig. 3 and finds it between *Recognition* and *Track*. Hence, for partner A, *Cognize* will be the input for development along its expertise and *Track* the output.

Hence, each partner identified the precise input requirement and the output levels along different contribution aspects, in a unified manner. In this way, two partners can identify complementary roles even to work in same contribution aspects in parallel, with the agreement of the high-level interface (API). This also allows to identify multiple input and their status for a particular component, hence easy facilitates collaborative development of multi-modal perception system. For example, speaker localizing might use inputs from detection of face and sound direction. This approach also helps to identify the missing links and blocks and to take decision about who will supply those, including using some third party components, as the project progresses.

4 NAOqi, a unified functional framework

As mentioned, in Romeo2 project, we are bringing various perception components not only within a unified conceptual framework as described earlier, but also within

a unified functional framework. *NAOqi*, the operating system on the Romeo robot, is serving for this purpose.

NAOqi (online documentation [20]) framework allows homogeneous communication between different modules (motion, audio, video), homogeneous programming and homogeneous information sharing. It has essential features of a robot programming and control. Some of the features, which made it possible to be used for development by all the partners of Romeo2 project, having diverse system dependability, preferences and requirements are: **Cross-platform**, for development in *Windows*, *Linux* or *Mac*. **Cross-language**, identical API for different languages C++ and Python. **Introspection**, allowing to know, which functions are available in different modules and where. Very important for consistency in API. **Blocking and non-blocking method calls**, in non-blocking calls, a task is created in a parallel thread, enabling to instruct the robot to do multiple activities, e.g. walking while talking. **ALMemory**, the robot's memory, addressing the requirement of having a thread-safe centralized shared memory across modules. Different modules can read or write data, which should be available to monitor the state and shape the interaction. **Events**, the mechanism of event and subscription through callback methods, help in developing an efficient reactive multi-modal perception-interaction system. E.g. a *FaceReaction* module, having a method *onFaceDetected*, can subscribe to *FaceDetected* method of *FaceRecognition* module, with the *onFaceDetected* as callback. This will cause the face detection algorithm to run, and every time a face is detected, the appropriate method will be called back.

5 Discussion: Results of Work in Progress, Advantages and Gateway to Innovation

At the time of writing this paper, most of the modules of table 1 have been achieved within Naoqi framework. Even some modules, which are evolving, such as User Profile and Emotion analysis, are available for experimentation and collaborative development. We will not go in detail of these individual modules and the results, as those can be found online [20]. Instead, next we discuss some of the advantages and innovation potentials, which such modules functioning on a unified platform could bring.

Many times innovations are blocked or even not foreseen because of (i) unavailability of different basic components on a single system or (ii) the lack of capability to extract and ground symbolic information and the sensor signal. The Romeo2 project aims to achieve both these aspects. Thus, the project also opens doors of various innovation potentials and serves as base for various other projects, in a practical manner. Below we outline some of such pointers. Due to space limitation, we will not provide much technical details, but point to interesting aspects of the various ongoing experiments and the results obtained so far.

Ex1: One of the practical uses of Romeo2 project is in healthcare. The capability of multi-modal perception, combining input from the interacting user, the events triggered by other perception components, and the centralized memorization mechanism of robot, help to achieve the goal by dynamically shaping the interaction.

For example, a medication monitoring service component developed by one partner can influence the ongoing user-robot interaction by a dialog based module developed by other partner. The medication component fires the event "take_medication" at appropriate time. If the perception says that the user has not taken the medication, the

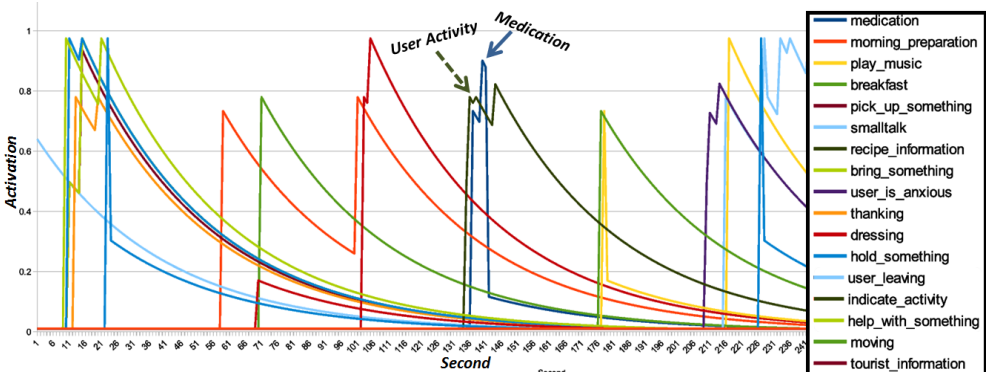


Fig. 4. Subset of topics for interaction (right), and their dynamic activation levels based on multi-modal perception and events.

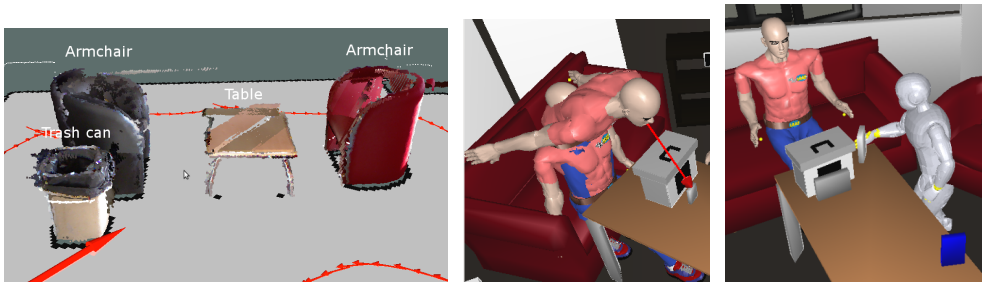
dialogue module reacts to these events by increasing the activation for the medication topic, and eventually deciding to talk about the medication of the user.

To demonstrate we programmed an extensive dialogue with 26 topics that shows the capabilities of the Romeo robot. During this dialogue the user often interrupts Romeo to quickly ask a question, this leads to several 'conflicting' topics in the dialogue manager. The activation of different topics during an interaction over a period is shown in fig. 4. The plot shows that around the 136th second the user has to take his medicine, but the memory about the assessed situation indicates that the user has ignored and not yet taken the medicine, resulting into the robot urging the user to take his medication (pointed by blue arrow), and surpassing the activity, which was indicated by the user during the conversation, to engage in reading a book (pointed by dotted arrow in dark green). Hence, a close loop between the perception and interaction is getting achieved in a real time, dynamic and generalized manner.

Ex2: Fig. 5(a) shows situation assessment of the environment and objects at the level of semantics and affordances, such as there is a 'table' recognized at position X, and this belongs to an affordance category on which something can be put. Fig. 5(b) show situation assessment by perspective taking, in terms of abilities of the human. This enables the robot to infer that the sitting human (as shown in fig. 5(c)) will be required to stand up and lean forward to see and take the object behind the box. Thanks to the combined reasoning of (a) and (b), the robot will be able to make the object accessible to the human by placing it on the table (knowing that something can be put on it), at a place reachable and visible by the human with least effort (through the perspective taking mechanism), as shown in fig. 5(c).

In Romeo2 we aim to go even further and combine the reasoning about abilities and efforts of agents, and affordances of environment, to perceive the situation and ground the interaction, for human-level understanding of task semantics through demonstration, for proactive behaviors developments in the robot, etc. Some complementary studies in those directions, [21], [22], well provide supporting evidences for such innovation potentials.

Ex3: Analyzing verbal and non-verbal behaviors such as head direction (e.g. on-view or off-view detection) [17], speech rhythm (e.g. on-talk or self-talk) [24], laugh detection [26] and their dynamics (e.g. synchrony [7]), combined with acoustic analysis



(a) (b) (c)

Fig. 5. High-level situation assessment. (a) The semantics map of the environment built by the robot with recognized objects (tables, chairs, trash can, etc.) along with the associated affordances information (to put on, to sit, etc.). (b) Effort and Perspective taking based situation assessment. The robot estimates that the person currently sitting on the sofa will be required to stand and lean forward to see the small object behind the big object. (c) Combining (a) and (b), the robot will be able to make the object accessible to the human, by placing it on appropriate affording support at appropriate place.

(e.g. spectrum) and prosodic analysis (e.g. mainly the fundamental frequency and energy) altogether greatly allows to improve social engagement characterization of the human during interaction, to better characterize the emotions and social signals, the fundamental for social intelligence. One of the aspects for situation assessment during interaction is discrimination between *On-talk* and *Self-Talk*. On-talk is system directed speech whereas Self-Talk is audible or visible talk people use to communicate, known to reflect the cognitive load of the user, especially for elderly.

To demonstrate the strength of multi-modality of Roemo2 in improving detection of such situations, in a context of assistance to elderly people with Mild Cognitive Impairments, we collected a database of human-robot interaction during sessions of cognitive stimulation. The preliminary result with 14 users shows that on a 7 level evaluation scheme, the average scores for questions, "*Did robot show any empathy?*", "*Was it nice to you?*", "*Was it polite?*" were *6.3*, *6.2*, *6.4* respectively. In addition, the multi-modality combination of the rhythmic, energy and pitch characteristics seems to be elevating the detection of self-talk as shown in table of fig. 6.

Features	SVM
Pitch-based	52.16 %
Energy-based	59.51 %
Rhythm-based	56.97 %
Pitch + Energy	64.31 %
Pitch + Energy + Rhythm	71.62 %

Fig. 6. Self-talk detection

Ex4: Inferring face gaze using embedded optic sensors (as illustrated in fig. 7), combined with sound localization using audio sensors and object detection, altogether provides enhanced knowledge about who might be speaking in a multi-people human-robot interaction, and further facilitates analyzing the attention and intention.

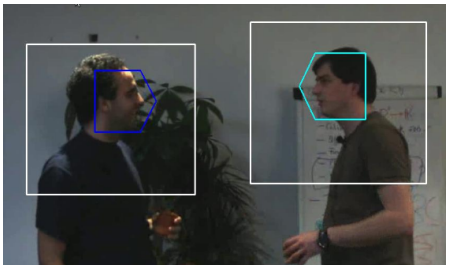


Fig. 7. Face, shoulder and face orientation detection of two interacting people.

To demonstrate this, we have experimented with two speakers initially speaking at the different sides of the robot. Then they slowly move towards each other and eventually separate away. The graph in fig. 8

shows the preliminary result for the sound source separation by the system based on beamforming. The left part (BF-SS) shows when only the audio signal is used. However, thanks to the rich multi-modal perception, when the system uses the visual information combined with the audio signals, the performance is better (AVBF-SS), which is evident in all the three types of analyses: signal-to-interference ratio (SIR), Signal-to-Distortion Ratio(SDR) and signal to artifact (SAR) ratio.

Ex5: The emotion analysis is significantly enriched, when the rich information about visual smile detection, audio speech rhythm analysis and the user profile along several dimensions (extraversion, emotionality, dominance, optimism, affinity, self-confident) are available. Further, the fusion between visual clue and the analysis of lexical content open doors for automated context extraction, and helps in not only for better interaction grounding but also for making the interaction interesting, like doing humor, [15].

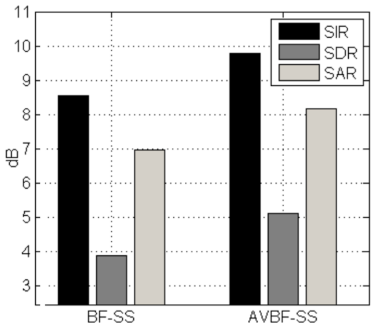


Fig. 8. Separation of two sound sources, only audio based (BF-SS) and audio-video based (AVBF-SS).

Basis for further exploration The project is not only identifying innovative aspects but also addressing new research challenges, which arise when such different blocks should be functioning together. All together, it is serving as a solid elevated perception system, as base for various other projects. E.g. in *EARS project* [9], which aims to develop the fundamentals for a natural dialogue between humans and robots in adverse acoustical environments, *JOKER project* [15], which aims to create a human-robot interaction system with social and affective communication skills including humor and other informal socially-oriented behaviors.

6 Conclusion and Future Work

In this paper, we have provided an overview of the rich multi-modal perception and situation assessment system within the scope of Romeo2 project. We have presented our sensing-interaction perception architecture and the key perception components requirements. Using both, we have outlined our collaborative development approach, which simplifies the identification of dependencies and facilitates parallel developments. Further, we have pointed towards some of the work in progress innovation potentials, achievable through such a system and how it is practically closing the sensing-interaction loop. In this way, the paper not only aims to present desired capabilities and key functional requirements for a companion robot, but also could serve as guideline in different context such as robot co-worker.

References

1. Beck, A., Risager, C., Andersen, N., Ravn, O.: Spacio-temporal situation assessment for mobile robots. In: Int. Conf. on Information Fusion (FUSION). pp. 1–8 (July 2011)
2. Bolstad, C.A.: Situation awareness: Does it change with age. vol. 45, pp. 272–276. Human Factors and Ergonomics Society (2001)
3. Buendia, A., Devillers, L.: From informative cooperative dialogues to long-term social relation with a robot. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 135–151 (2014)

4. Caron, L.C., Song, Y., Filliat, D., Gepperth, A.: Neural network based 2d/3d fusion for robotic object recognition. In: Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN) (2014)
5. Chella, A.: A robot architecture based on higher order perception loop. In: Brain Inspired Cognitive Systems 2008, pp. 267–283. Springer (2010)
6. CHRIS-Project: Cooperative human robot interaction systems. http://www.chrisfp7.eu/
7. Delaherche, E., Chetouani, M., Mahdhaoui, A., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony: A survey of evaluation methods across disciplines. Affective Computing, IEEE Transactions on 3(3), 349–365 (July 2012)
8. Devillers, L.: Automatic detection of emotion from real-life data. In Prosody and Iconicity pp. 219–231 (2013)
9. Ears-Project: Embodied audition for robots. http://robot-ears.eu/
10. Endsley, M.R.: Theoretical underpinnings of situation awareness: a critical review. In: Endsley, M.R., Garland, D.J. (eds.) Situation Awareness Analysis and Measurement. Lawrence Erlbaum Associates, Mahwah, NJ, USA (2000)
11. Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors: The Journal of the Human Factors and Ergonomics Society 37(1), 32–64 (1995)
12. EYESHOTS-Project: Heterogeneous 3-d perception across visual fragments. http://www.eyeshots.it/
13. Filliat, D., Battesti, E., Bazeille, S., Duceux, G., Gepperth, A., Harrath, L., Jebari, I., Pereira, R., Tapus, A., Meyer, C., Ieng, S., Benosman, R., Cizeron, E., Mamanna, J.C., Pothier, B.: Rgbd object recognition and visual texture classification for indoor semantic mapping. In: Int. Conf. on Technologies for Practical Robot Applications (2012)
14. Jensen, B., Philippsen, R., Siegwart, R.: Narrative situation assessment for human-robot interaction. In: IEEE ICRA. vol. 1, pp. 1503–1508 vol.1 (Sept 2003)
15. JOKER-Project: Joke and empathy of a robot/eca: Towards social and affective relations with a robot. http://www.chistera.eu/projects/joker
16. Lallee, S., Lemaignan, S., Lenz, A., Melhuish, C., Natale, L., Skachek, S., van Der Zant, T., Warneken, F., Dominey, P.F.: Towards a platform-independent cooperative human-robot interaction system: I. perception. In: IEEE/RSJ IROS. pp. 4444–4451 (Oct 2010)
17. Le Maitre, J., Chetouani, M.: Self-talk discrimination in human-robot interaction situations for supporting social awareness. Int. J. of Social Robotics 5(2), 277–289 (2013)
18. Lee, S., Baek, S.M., Lee, J.: Cognitive robotic engine: Behavioral perception architecture for human-robot interaction. In: Human Robot Interaction (2007)
19. Mekonnen, A.A., Lerasle, F., Herbulot, A., Briand, C.: People detection with heterogeneous features and explicit optimization on computation time. In: IEEE/RSJ IROS (2013)
20. NAOqi-Documentation: https://community.aldebaran-robotics.com/doc/2-00/naoqi/index.html/
21. Pandey, A.K., Alami, R.: Towards human-level semantics understanding of human-centered object manipulation tasks for hri: Reasoning about effect, ability, effort and perspective taking. Int. J. of Social Robotics pp. 1–28 (2014)
22. Pandey, A.K., Ali, M., Alami, R.: Towards a task-aware proactive sociable robot based on multi-state perspective-taking. Int. J. of Social Robotics 5(2), 215–236 (2013)
23. Pomerleau, D.A.: Neural network perception for mobile robot guidance. Tech. rep., DTIC Document (1992)
24. Ringeval, F., Chetouani, M., Schuller, B.: Novel metrics of speech rhythm for the assessment of emotion. Interspeech pp. 2763–2766 (2012)
25. Romeo2-Project: Humanoid robot assistant and companion for everyday life. http://www.projetromeo.com/
26. Soury, M., Devillers, L.: Nao makes me laugh: the impact of humor in human-robot interactions. In: IEEE/RSJ IROS (2013)