Preliminary Study of Brain-Inspired Model for Multimodal Human Behavior Detection in Social Context

Alessandra Sorrentino, Laura Fiorini, Gianmaria Mancioppi, Olivia Nocentini, and Filippo Cavallo

Abstract In a near future, robots will permeate our daily life; indeed, they have the potential to proactively support the senior citizen in tedious and different daily tasks (i.e. cleaning, gaming, walking activity, promote socialization). However, to efficiently cooperate with human-beings, robots should have enhanced human–robot interaction capabilities. This work addresses the challenge of designing a robotic model by simulating the modality human beings interact with each other. The first objective of this work concerns the identification of the social cues which correctly describe the user's emotional and engagement state during the interaction. Based on selected descriptors, a perceptual system has been proposed to detect elderly's behavior in a social context. The proposed architecture is inspired to the human brain structure as concern the functionalities modules and analogies in the modules' location. The proposed perceptual system aims at transforming raw data coming from three kinds of sensors (camera, microphone, and laser) into behavioral patterns by mimicking the abstraction evolution which characterizes the consciousness process of the human brain.

A. Sorrentino (✉) · L. Fiorini · G. Mancioppi · O. Nocentini · F. Cavallo
The BioRobotics Institute, Scuola Superiore Sant'Anna, Viale Rinaldo Piaggio, 34, 56025 Pontedera (PI), Italy
e-mail: alessandra.sorrentino@santannapisa.it

L. Fiorini
e-mail: laura.fiorini@santannapisa.it

G. Mancioppi
e-mail: gianmaria.mancioppi@santannapisa.it

O. Nocentini
e-mail: olivia.nocentini@santannapisa.it

F. Cavallo
Department of Industrial Engineering, University of Florence, Via Santa Marta 3, 50139 Florence (FI), Italy
e-mail: filippo.cavallo@unifi.it

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

A. Monteriù et al. (eds.), *Ambient Assisted Living*, Lecture Notes in Electrical Engineering 725, https://doi.org/10.1007/978-3-030-63107-9_14

1 Introduction

As fertility declines and life expectancy rises, the proportion of the population above a certain age rises as well. This phenomenon, known as population aging, is occurring throughout the world. The population aged 60 or above is growing at a rate of about 3 percent per year. Currently, Europe has the greatest percentage of the population aged 60 or over (25%) [1]. For this reason, several global initiatives focus on defining solutions for satisfying the needs of the aging population in order to prolong independent living. Indeed, most elderly people prefer to live independently in their homes as long as possible, as this leads to a richer social life and it is paramount to maintaining established habits [2].

In this context, Social Assistive Robots (SAR) can play an important role in the promotion of quality of life by integrating activities with independent-living older adults [3]. Using robots as home-health aids is one promising solution to support older adults' needs. If we want to integrate a social robot for a long-term period in a house, one of the main goals of the robot is to create stimulating and engaging interactions in which a user actively participates for an extended period of time.

One strategy is to provide the robot with social intelligence. Thanks to this ability, the robot recognizes the current social behavior of the user and it is able to adapt its behavior so that to enhance the interaction in a polite and pleasant manner [4]. It highlights the importance of developing an automated assessing of user's social behaviors.

The attitude of a person towards social interaction is expressed by a set of social signals which conveys information about mental state, feelings and other personal traits (i.e. eye gazing, postures, voice quality) [5]. In human–human interaction, the listener automatically assesses the emotional and engagement state of the speaker. This human inherent ability in a social context is the effect of specific processes happening at the level of the brain, as described by Theory of Mind [6].

The challenge is how to develop a computational model that can simulate the modality human beings interact with each other. The first objective of this work is the identification of the social cues that are meant to describe the user's behavior during interaction. Among all, the ones listed in this work have been chosen as the most representative descriptors of the human emotional and engagement state. Secondly, this work aims at modelling a perceptual system able to recognize human social behavior. The importance of this system is twofold. Firstly, by recognizing human behavior online, the robot can act accordingly performing the most appropriate behavior (i.e. the perceptual system represents the sense block in the sense-plan-act loop which characterizes any human–robot interaction system). On the other side, the perceptual system can be used to assess human feedback on the sequence of actions performed by the robot.

The proposed system has been modeled in a brain-inspired way. It implies that our perceptual system aims at transforming raw data coming from the sensory equipment into behavioral patterns by mimicking the abstraction process occurring in the human

brain. The proposed perceptual model re-creates the flow of information exchanged by the thalamus and brain cortex areas to infer the social behavior of the user.

The paper is structured as follows. Section 2 reviews the current state of the art of brain-inspired systems. Section 3 summarized the neuroscientific findings that guided the development of the proposed architecture. Section 4 details the social cues of interest and the description of the system. Discussion and conclusions are listed in Sect. 5.

2 Related Works

Biological systems provide a new source of inspiration for developing intelligent and autonomous robots. Biological principles coming from plants or animals models are becoming widely used in the design of robots that can sense, think, walk, swim, crawl, or fly [7]. The innovative structures developed with this emerging field allow achieving intelligence, flexibility, stability, and adaptation for emergent robotic applications, such as manipulation, learning, and control [7]. In this context, human brain activity is also used as an inspiration model for robotic control architectures.

Neuro-robotic models are conceptualized as networks that produce and combine information [8]. The information captured by the sensory channel from the environment or other neurons is integrated and distributed along with the artificial neural network [9]. Neurophysiological insights were used to develop artificial architectures. According to the task the robot needs to perform, specific functionalities of the brain areas are replicated by neuro-robotic models. Several examples are reported in Table 1.

Computation models of the thalamus and limbic system are implemented to replicate a range of cognitive processes such as emotional feedback, selection of appropriate actions and memory in robotic platforms. In this context, the functionality of the thalamus consists in pre-processing the stimuli collected from the environment. The raw representation of the collected data is then elaborated at a higher level by the sensory cortex. Both areas communicate with other brain-inspired modules (i.e. amygdala) to infer the relative emotion, given the condition generating it, as shown in [10] and [11]. The flow of information from the thalamus to the limbic system is frequently mimicked not only to recreate some emotions, as in [10] and [11], but also to assess user's engagement state in human–robot interaction. The work described in [12] shows a complete architecture that allows a robot to recognize the level of attention of children and to react accordingly. The robotic model has been developed by mimicking thalamus, limbic system, and prefrontal cortex's functionalities. The only limitation of the system described by [12] is that the child behavior assessment is performed by detecting one social cue, the postural gesture.

To detect human behavior, multiple social cues need to be identified [5]. Multiple social signals imply multiple sensors mounted over the robot. Due to the complexity of information flow during the perception phase, our work proposes a brain-inspired perceptual system able to identify the behavior of the user from the multimodal data

Table 1 Overview of brain-inspired applications

References	Robot task	Robot	Input signals	Brain areas
[10]	Emotional learning	None	Visual and audio data. Exchange of information between brain areas	Thalamus, sensory cortex, orbitofrontal cortex and amygdala
[11]	Artificial fear generating	None	Environmental data	Sensory system, amygdala system, hippocampal system and working memory
[16]	Goals generation	NAO	Visual data (colors and shape of objects)	Amygdala-thalamus-cortical circuit at its functional level
[12]	Joint attention detection	Robotis Bioloid humanoid robot	Visual data (child's posture)	Amygdala, hypothalamus, hippocampus and basal ganglia
[13]	Person recognition	Pioneer 3DX (P3-DX)	Environmental data, visual data (body and face of the user) and biometric features	Hierarchical structure of the sensory cortex and replication of the spatial–temporal binding criteria

collected during the interaction. Concerning the work described in [13], our model does not replicate the brain structures at the level of neuron configurations, but the modules composing our system mimic functionalities of the brain areas involved. In details, computational models of the thalamus and brain cortex are implemented.

We propose a brain-inspired model which includes insights of the Theory of Mind to improve the human–robot interaction. Theory of Mind is commonly referred as the inference mechanism of identifying the mental state (i.e. emotions, intentions, beliefs) of a person, which is automatically performed by the human brain in social situations [14, 14]. Theory of Mind can be replicated artificially by combining the identification of multiple social cues defined in the Social Signal Processing (SSP) field [5], as proposed by the authors of [15]. This approach opens two challenges. The first issue is related to the identification of social cues which are more representative of the human mental state. The sources of information of our perceptual system are images, audio and laser data. Consequently, specific social cues have been selected according to each modality of perception. The second challenge is relative to the way the social signals should be combined so that to express the user's behavior. Our work focuses on two specific aspects of social behavior: emotional and engagement state. Each aspect is encoded by specific patterns, expressed by a sequence of multimodal data. Thanks to this structure, the proposed brain-inspired system can assess the current social behavior of the user in real-time.

3 Neuroscientific Background

In the following, we delineate the most significant functionalities of brain areas mimicked by our perceptual system.

3.1 Thalamus

The thalamus is the biggest group of neurons, which is encompassed in the diencephalon and it represents the antechamber of the brain. It receives inputs from several systems, such as the sensory, the motor, and the limbic systems, but also from the reticular formation. The fibers that carry this information outreach different nuclei in the thalamus, which, in turn, send these fibers to a very discrete portion of the ipsilateral brain cortex, except for the thalamic-reticular nucleus. This nucleus is the only thalamic nucleus with an inhibitory activity, and it does not project directly to the brain cortex. The thalamic networks represent the first preprocessing station of information and the first structure related with the integration of this information, and their neural activity is related with the consciousness state of the subjects, indeed thalamic neurons own two different fashion to fire; tonic firing, related with awake state, and burst firing, related with drowsiness and sleep state. The thalamic networks represent an important stop station also for descending fibers, which travel in the opposite direction from the Central Nervous System (CNS) to the Peripheral Nervous System (PNS) [17].

3.2 Brain Cortex

Human brain cortices are several and they accomplish for different goals. Particularly, concerning the aim of this work, sensory and associative cortices will be taken into account. Sensory cortices are the brain areas responsible to the processing of sensory information, such as hearing, touch, vision, proprioception, and others. Information gathered from human sensory organs travel from the PNS to the CNS passing throughout the thalamus nuclei, since they reach primary sensory cortices. Each sensorial modality refers to a specific neural substrate, which elaborates the sensorial inputs at a low level of complexity and integration. After that, the information travels towards others sensory cortices, belonging to the same sensorial modality, which cooperates among them analyzing different aspects of the sensorial inputs, assembling them, to reach a higher level of integration and complexity. Subsequently, the information is sent towards associative areas, in which it is merged with other types of information. At this step, the information is not strictly related to the unimodality nature of the stimuli. It is shaped by an increasing level of abstraction and complexity. These areas are not related to a specific sensorial modality, quite the opposite, they

handle the integration of several types of information together, to accomplish a higher level of processing. That represents one of the prerequisites to create consciousness [18].

4 Brain-Inspired Perceptual System

This work proposes reverse engineering of the human brain functionality to improve human–robot interactions.

In the designing phase, a set of social cues representative of the human behavior has been selected based on the findings highlighted by the SSP research field [5]. The chosen social cues belong to three input modalities: image, audio, and laser. It is the explanation behind the installation of the camera, microphones, and laser scan over the robotic platform.

In the implementation phase, the ability of the brain to process stimuli from the environment is mimicked by the interconnection of three modules: thalamus, sensory cortex and associative cortex. The functionality of each module shows analogies with the abilities of the corresponding human-beings' neural structure. Besides, the flow of information across the modules is characterized by an increasing level of abstraction, which extends structure complexity and recalls the "convergence-zone" described by Damasio [19].

The overall architecture of the system is shown in Fig. 1. The first module coincides with a lower level of abstraction (thalamus). In this layer, raw data collected during the interaction are preprocessed to reshape them. The medium level of abstraction corresponds with the second module (sensory cortex), where the system processes the reshaped data to extract social cues defined in the design phase. At the higher level of abstraction (associative cortex), the social signals are merged together to be descriptors of a particular social behavior. Descriptors are conveniently combined to perform the automatic assessment of human behavior online.

4.1 Social Cues

Data recorded by the sensors mounted over the robot are used to detect social cues. Social cues are descriptors of the behavioral state of the user during the interaction. Among the social cues described in [5], this work focuses on the following ones.

1. Posture and body movements: since it is assumed unconsciously, body posture represents honest information to assess the engagement of the user during the interaction [20]. In this work, the identification of inclusive and non-inclusive postures is performed. In details, the features of interest consist of body orientation and interpersonal distance between the user and the robot. The quantity of motion performed by the user during the interaction expresses the current

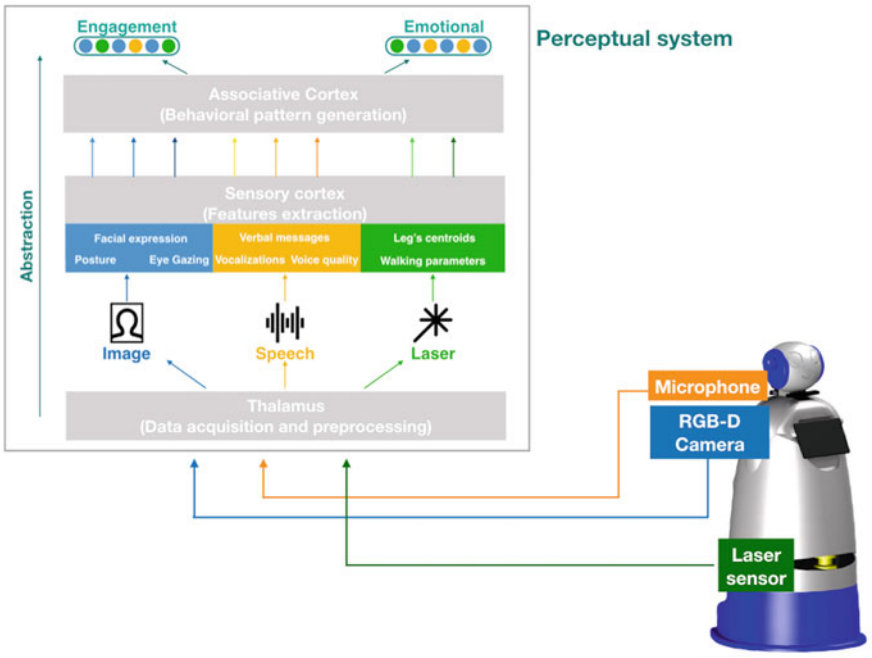

Fig. 1 Architecture of the proposed brain-inspired perceptual system

emotional state. Based on Darwin's discoveries [21], specific changes in arm movements and walking parameters (i.e. gait) express social affective states. For example, frequent changes of posture and high quantity of motion are usually related to embarrassment [22]. The recognition of the body postures and counting of their occurrences aims at assessing the engagement level of the user in the social context.

2. Facial expressions: they directly communicate the affective state of a person [23]. In details, facial expressions allow the identification of six emotions: fear, sadness, happiness, anger, disgust, and surprise.
3. Head orientation: it is described by the eye-gazing. It is the key component of the joint attention mechanism in the human–human interaction. By focusing on a certain object, one person keeps the eyes in the direction where the object is located and automatically bring the other person to look at the same direction (joint attention) [24]. Besides, when two people are engaged in eye contact, the interaction is perceived more likable [25]. Eye gazing provides an insight on the engagement level of the user and a feedback on the attention captured by the robot during the interaction [12].
4. Verbal messages: this category of social cues includes repetition, incomplete words, amount of silence during the interaction. The repetition of words in a small amount of time can be connected to a particular cognitive state of the user

(i.e. confusion) or it can be used as an input to modify the current behavior of the robot. The presence of incomplete words during the dialog can be an evidence of cognitive decline. In this context, the absence of speech is an important indicator to understand not only the role of the user in the interaction (speaker or listener) but also to evaluate the current emotional and engagement state [5].

5. Voice quality: it is described by a set of prosodic features of the speech signal to detect the emotional state of the user [26]. Tempo, energy and pitch are the prosodic features of interest. Tempo is defined as the speaking rate detected in the signal, i.e. the number of phonetically relevant units. Together with the information carried by energy feature, tempo is an emotional descriptor [27]. Pitch provides insights relative to the personality traits, such as lack of hesitations [27].
6. Vocalizations: they encode the emotion of the person speaking. Identifying two groups of vocalizations is possible. Linguistic vocalizations include small words (i.e. "ehm-ehm", "ah-ah") which occur more often when a person is feeling difficulty for the situation [28]. Non-linguistic vocalizations belong to the non-verbal sounds like sobbing, crying, laughing and they are more representative of the overall emotional state of the user.

4.2 Data Acquisition and Pre-processing

The model has been developed to enrich the interaction of the elderly user with ASTRO robot. ASTRO is a service robot designed to assist an elderly person with mobility needs. The first version of this robot was developed under ASTRO mobile project [29] and it was refined under ACCRA project [30]. ASTRO is based on the Scitos G5 robotic platform (Metralabs GmbH, Germany). Concerning safety, SCITOS G5 is equipped with a bumper and a couple of emergency-stop buttons to the stop motor. About Human–Machine-Interface (HMI) two touch screens for direct access to the services were mounted on the front and on the back side of the robot. External sensors were integrated into the system to extend the sensing abilities of ASTRO.

On the front, the platform is improved with an additional laser (Laser Sick 300) for the perception of the surrounding environment and the detection of walking parameters. SICK 30B-2011BA is a 2-dimensional laser sensor for area scanning. The light source of the sensor is a pulsed light laser diode (infrared) of wavelength 785 nm with laser class 1 safety. The scan area is 270º semicircle with an angular resolution of 0.5º. Scans are performed at a frequency of 12,5 Hz. The detection distance range is 30 m. These sensors can acquire a different type of information.

The ORBBEC Astra Pro[1] camera is mounted on the front of ASTRO to collect visual data from the environment and from the user. It is an RGB-D camera which provides, at the frame rate of 30 fps, RGB images with a resolution of 1280 × 720 pixels and depth images with a resolution of 640 × 480 pixels. One of the main

[1] https://orbbec3d.com/.

Table 2 Summary of social cues

Behavioral aspect	Social cue	Features	Sensor
Engagement state	Body posture	Body orientation	Camera
		Interpersonal distance	Laser
	Head orientation	Eye gazing	Camera
Emotional state	Expression	Facial expression	Camera
	Voice quality	Tempo	Microphone
		Energy	Microphone
		Pitch	Microphone
Engagement and emotional state	Quantity of motion	Arm movements	Camera
		Walking parameters	Laser
	Verbal messages	Repetitions	Microphone
		Incomplete words	Microphone
		Amount of silence	Microphone

features of the Astra Pro camera is the range it can cover. The range of the camera covers up to 8 m of distance. The camera integrates two built-in microphones, which can be used to record the surrounding sound and user's speech.

Multidata recording is handled by the thalamus module which collects the incoming signals and pre-processes them in terms of synchronization and filtering. Furthermore, the thalamus module reshapes the collected data accordingly to the requirements of the feature extraction's algorithms of the sensory cortex module. The detection of the aforementioned social cues is performed by integrating appropriate sensors and implementing dedicated machine learning algorithms over the robotic platforms. Table 2 summarizes the relation between social cues, the sensor used to collect corresponding raw data and the aspect of the human behavior they are more representative of.

4.3 Features Extraction

The second module is composed of the set of feature extraction techniques, each of them aims at detecting a certain cue from the incoming signals. This module consists of three blocks. Mimicking the functionality of the sensory cortex, which processes

unimodal signals in the human brain, each block composing the second module takes as input one interaction modality. In details, an image for the visual modality, a speech signal for the audio modality and a laser data for the motion modality.

On the image data, the following feature extraction techniques are applied.

- Pose estimator: to detect the posture of the user during the interaction. It consists of a skeleton tracker which extracts the 3d position of joints and compare the current joint configuration with a set of given meaningful poses. The methods developed by [31] allows the identification of joint poses in real-time with high accuracy without taking as input depth information. It can accelerate the perception process.
- Facial expression recognizer: to recognize the emotional state of the user when the face is visible from the robot perspective. Among all the available algorithm, a face detector based on Viola and Jones algorithm ([32]) is used to detect the face in the image. The face profile is then sent to a cloud face expression recognition API (i.e. Microsoft Face API) to select the emotion with the higher confidence level.
- Eye gazing estimator: to evaluate the engagement state of the user in interaction and to guide the attention of the robot towards the direction the user is looking at. Eye gazing estimator relies on the 2D data coming from the camera by using the more reliable technique among the ones described in [33].

On the speech signal, the following feature extraction algorithms are applied.

- Verbal messages detector: to identify the repetition of specific words and count the number of incomplete words, automatic transcription of dialog can be used to identify the repetition of specific words of interest. Given the speech-to-text output, the assessment of the number of times sequence of words are repeated can be performed. On the other hand, the recognition of incomplete words is out of the current development due to the challenges related to it. The presence of silence in the speech can be recognized by applying a Voice Detection Activity (VAD) algorithms. They allow the machine to distinguish from an audio frame containing silence from another with data.
- Voice quality estimator: to calculate the prosodic features (tempo, energy, and pitch), the tools listed in [5] can be combined;
- Vocalizations recognizer: it can be built over an automatic speech recognizer to identify specific linguistic elements (“uhm”, “ehm”, “ah-ah”) and certain sounds expressing emotional states of the user. Automatic speech recognition algorithm (i.e. Google Speech API, Microsoft Speech API) can be used to detect specific words of interest.

On the laser data, the following feature extraction techniques are applied.

- Leg’s centroid detector: this technique permits the computation of the interpersonal distance between the user and the robot. The distance is an indicator of the established relationship between the user and the robot (proximity feature) [5].
- Walking parameter recognizer: to evaluate the quantity of motion of the user. The quantity of motion expresses the feelings towards the social situation. The walking

parameter detected in this work is the gait parameter, which is computed terms of velocity and acceleration of the walking performance in front of the robot.

Most of the techniques described above rely on pre-trained models available on the Cloud. The novelty of this work is to combine them successfully by parallelizing the procedures. Due to the high workload, the strategy adopted to perform the feature extraction online is to process the data on Cloud so that to reduce the amount of running processes on the robotic platform. Furthermore, the pre-trained models strictly depend on the data they are trained on. Since in most of the cases the models are trained on datasets composed by young people's features, the models will be re-trained on a larger dataset including elderly data.

4.4 Behavioral Pattern Generation

The term behavior refers to (human) actions associated with emotions, personality, and psychological state [34]. In this work, the behavior of the user is analyzed in two different aspects: engagement in the conversation and emotional perception of the situation, which are representative of the current mental state of the user. The two aspects characterize the prototypical nature of human behavior in a social situation. The strategy adopted in this work is to combine the unimodal attributes extracted by the sensory cortex based on the behavioral aspects they represent the most. The idea is to represent the behavior as a tensor of multimodal data.

To automatically exploits the complementarity and redundancy of multiple modalities, an artificial neural network model is implemented to construct a joint mutual representation [35]. By using a joint representation, it is possible to compute similarities between the features in the representational space which reflects the similarities in the corresponding semantic concepts [36], that in this work are the behaviors. The neural network for multimodal representation is fed by the features of each modality. The architecture is composed of individual neural layers (one for each modality) alternated by hidden layers, which project the features into the joint space [35].

As shown in Table 2, it is possible to identify features that are more likely to be a meaningful descriptor of a particular behavior (i.e. the interpersonal distance and body posture of the user are meaningful descriptors of the engagement aspect). Tensor composed by the features belonging to a certain behavioral aspect can be used as training data of the neural network. Even if neural networks require a large amount of training data, they provide superior performances in joint representations tasks [35].

5 Discussion and Conclusion

Our work represents a new strategy for human behavior recognition in human–robot interaction. Since human being is a social animal and it can handle a social situation in a fast and appropriate way, the idea is to develop a multimodal perceptual module inspired to the human brain. Concerning previous works, the proposed architecture is composed by a confined number of modules organized in a hierarchical structure, influenced by findings of Theory of Mind. Furthermore, the hierarchical structure underlines the abstraction process developed to transform raw data into meaningful behavioral descriptors. This work provides a solution to the proper features selection and descriptors to capture human social signals by combining insights coming from social signal processing. By automatic assessing of human behavior, the system contributes to the implementation of a robotic social intelligence.

A limitation of the proposed system is relative to the absence of decision-making, planning, and acting modules. A possible approach is to re-create them in a brain-inspired way and to provide a complete interaction system. For example, the decision-making process can be included by adding an orbital prefrontal module, able to select the most appropriate robot's response based on the detected human behavior.

Another limitation regards the lack of availability of datasets to train the features extraction models. To make the system working for elderly users, the learning models should be trained on them. One possibility to overcome this issue is to organize an ad-hoc experimental session to collect raw data of interest. By increasing the features datasets, it is possible to increase the training dataset also fed into the behavioral pattern generation model. It will lead to a more reliable recognition system.

As specified above, the aim of this work is to describe a novel perceptual model for human behavior recognition based on selected social cues, expressed by the user involved in the interaction. This work details the theorical background and the insights behind some designing choices. As future works, we will organize specific experimental sessions to test the proposed architecture, developed by using the approaches described in this paper. Besides, we will like to improve the system by enhancing the abstraction process. Future development involves the abstraction of social signals into social actions so that to have a system able to detect the social role [37] and the intentions of a person. It is possible to reach this aim by integrating the perception module with a memory structure continually updated with robot's experiences.

Acknowledgements This work was supported by the ACCRA Project, founded by the European Commission—Horizon 2020 Founding Programme (H2020-SCI-PM14-2016) and National Institute of Information and Communications Technology (NICT) of Japan under grant agreement No. 738251.

References

1. United Nations Department of Economic and Social Affairs Population Division: E02 World Population Prospects The 2017 Revision: Key Findings and Advance Tables. World Popul. Prospect. 2017 (2017) https://doi.org/10.1017/CBO9781107415324.004
2. Cortellessa G, Fracasso F, Sorrentino A, Orlandini A, Bernardi G, Coraci L, De Benedictis R, Cesta A (2017) Enhancing the interactive services of a telepresence robot for AAL: Developments and a psycho-physiological assessment. In: Lecture Notes in Electrical Engineering. https://doi.org/10.1007/978-3-319-54283-6_25
3. Tapus A, Mataric MJ, Scassellati B (2007) Socially assistive robotics [Grand challenges of robotics]. IEEE Robot Autom Mag 14:35–42. https://doi.org/10.1109/MRA.2007.339605
4. Kaiser FG, Glatte K, Lauckner M (2019) How to make nonhumanoid mobile robots more likable: employing kinesic courtesy cues to promote appreciation. Appl Ergon 78:70–75. https://doi.org/10.1016/j.apergo.2019.02.004
5. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vis Comput 27:1743–1759. https://doi.org/10.1016/j.imavis.2008.11.007
6. Frith U, Frith CD (2003) Development and neurophysiology of mentalizing.https://doi.org/10.1098/rstb.2002.1218
7. Fukuda T, Chen F, Shi Q (2018) Special feature on bio-inspired robotics.https://doi.org/10.3390/app8050817
8. Tononi G, Sporns O (2003) Measuring information integration. BMC Neurosci 4. https://doi.org/10.1186/1471-2202-4-31
9. Sporns O (2007) What neuro-robotic models can teach us about neural and cognitive development. Neuroconstructivism Perspect Prospect 2:179–204. https://doi.org/10.1093/acprof:oso/9780198529934.003.0008
10. Balkenius C, Morén J (2001) Emotional learning: a computational model of the amygdala. Cybern Syst 32:611–636
11. Raymundo CR, Johnson CG, Vargas PA (2015) An architecture for emotional and context-aware associative learning for robot companions.https://doi.org/10.1109/ROMAN.2015.7333699
12. Dağlarlı E, Dağlarlı SF, Günel GÖ, Köse H (2017) Improving human-robot interaction based on joint attention. Appl Intell 47:62–82. https://doi.org/10.1007/s10489-016-0876-x
13. Al-Qaderi MK, Rad AB (2018) A brain-inspired multi-modal perceptual system for social robots: an experimental realization. IEEE Access 6:35402–35424. https://doi.org/10.1109/ACCESS.2018.2851841
14. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 1:515–526. https://doi.org/10.1017/S0140525X00076512
15. Wiltshire TJ, Warta SF, Barber D, Fiore SM (2017) Enabling robotic social intelligence by engineering human social-cognitive mechanisms. Cogn Syst Res 43:190–207. https://doi.org/10.1016/j.cogsys.2016.09.005
16. Franchi AM, Mutti F, Gini G (2016) From learning to new goal generation in a bioinspired robotic setup. Adv Robot 30:795–805. https://doi.org/10.1080/01691864.2016.1172732
17. Jones EG (2012) The thalamus. Springer Science & Business Media, Berlin
18. Cacioppo JT, Tassinary LG, Berntson G (2007) Handbook of psychophysiology
19. Damasio AR (1989) The brain binds entities and events by multiregional activation from convergence zones. Neural Comput 1:123–132. https://doi.org/10.1162/neco.1989.1.1.123
20. Scheflen AE (1964) The significance of posture in communication systems. Psychiatry 27:316–331. https://doi.org/10.1080/00332747.1964.11023403
21. Darwin C, Prodger P (1998) The expression of the emotions in man and animals.https://doi.org/10.1017/CBO9780511694110
22. Costa M, Dinsbach W, Manstead ASR, Ricci Bitti PE (2001) Social presence, embarrassment, and nonverbal behavior. J Nonverbal Behav 25, 225–240. https://doi.org/10.1023/A:1012544204986

23. Ekman P (1993) Facial expression and emotion. Am Psychol. https://doi.org/10.1037/0003-066X.48.4.384
24. Emery NJ (2000) The eyes have it: the neuroethology, function and evolution of social gaze. Neurosci Biobehav Rev 24:581–604. https://doi.org/10.1016/S0149-7634(00)00025-7
25. Mason MF, Tatkow EP, Macrae CN (2005) The look of love: gaze shifts and person perception. Psychol Sci 16:236–239. https://doi.org/10.1111/j.0956-7976.2005.00809.x
26. Scherer KR (2003) Vocal communication of emotion: a review of research paradigms. Speech Commun 40:227–256. https://doi.org/10.1016/S0167-6393(02)00084-5
27. Scherer KR (1979) Personality markers in speech. In: Social markers in speech. Cambridge University Press
28. Glass CR, Merluzzi TV, Biever JL, Larsen KH (1982) Cognitive assessment of social anxiety: development and validation of a self-statement questionnaire. Cognit Ther Res 6:37–55. https://doi.org/10.1007/BF01185725
29. Cavallo F, Limosani R, Fiorini L, Esposito R, Furferi R, Governi L, Carfagni M (2018) Design impact of acceptability and dependability in assisted living robotic applications. Int J Interact Des Manuf 12. https://doi.org/10.1007/s12008-018-0467-7
30. D'Onofrio G, Fiorini L, de Mul M, Fabbricotti I, Okabe Y, Hoshino H, Limosani R, Vitanza A, Greco F, Giuliani F, Guiot D, Senges E, Kung A, Cavallo F, Sancarlo D, Greco A (2018) Agile Co-creation for robots and aging (ACCRA) project: new technological solutions for older people. Eur Geriatr Med 1–6. https://doi.org/10.1007/s41999-018-0106-7
31. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp 7291–7299. https://doi.org/10.1109/CVPR.2017.143
32. Viola P, Jones M (2001) Robust real-time object detection. Int J Comput Vis 4, 34–47
33. Admoni H, Scassellati B (2017) Social eye gaze in human-robot interaction: a review. J Human-Robot Interact 6:25–63. https://doi.org/10.5898/jhri.6.1.admoni
34. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2. https://doi.org/10.3389/frobt.2015.00028
35. Baltrušaitis T, Ahuja C, Morency LP (2019) Multimodal machine learning: a survey and taxonomy.https://doi.org/10.1109/TPAMI.2018.2798607
36. Srivastava N, Salakhutdinov RR (2012) Multimodal learning with deep boltzmann machines. Adv Neural Inf Process Syst 2222–2230. https://doi.org/10.1109/CVPR.2013.49
37. Lan T, Sigal L, Mori G (2012) Social roles in hierarchical models for human activity recognition. In: IEEE conference on computer vision and pattern recognition. pp 1354–1361. https://doi.org/10.1109/CVPR.2012.6247821