This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



# Design and Evaluation of a Unique Social Perception System for Human-Robot Interaction

Abolfazl Zaraki[1], Michael Pieroni[1], Danilo De Rossi[1,2], Daniele Mazzei[1], Roberto Garofalo[1], Lorenzo Cominelli[1] and Maryam Banitalebi Dehkordi[1]
[1]Research Center "E. Piaggio", University of Pisa, Italy
[2]Department of Information Engineering, University of Pisa, Italy

***Abstract*—Robot's perception is essential for performing high-level tasks such as understanding, learning, and in general, human-robot interaction (HRI). For this reason, different perception systems have been proposed for different robotic platforms in order to detect high-level features such as facial expressions and body gestures. However, due to the variety of robotics software architectures and hardware platforms, these highly customized solutions are hardly interchangeable and adaptable to different HRI contexts. In addition, most of the developed systems have one issue in common: they detect features without awareness of the real-world contexts (e.g., detection of environmental sound assuming that it belongs to a person who is speaking, or treating a face printed on a sheet of paper as belonging to a real subject). This paper presents a novel social perception system (SPS) that has been designed to address the previous issues. SPS is an out-of-the-box system that can be integrated into different robotic platforms irrespective of hardware and software specifications. SPS detects, tracks and delivers in real-time to robots, a wide range of human- and environment- relevant features with the awareness of their real-world contexts. We tested SPS in a typical scenario of HRI for the following purposes: to demonstrate the system capability in detecting several high-level perceptual features as well as to test the system capability to be integrated into different robotics platforms. Results show the promising capability of the system in perceiving real world in different social robotics platforms, as tested in two humanoid robots i.e., FACE and ZENO.**

***Index Terms*—Context-aware social perception, Human-robot interaction, Humanoid social robots, Meta-scene, Platform-independent system, Scene analysis.**

## I. Introduction

With the rapid advancement of robotics and related computing techniques, social robots are becoming more integrated into human empathic and emotional daily life [1]. They can be employed in different areas (e.g., research [2], autism therapy [3], [4], education [5], and domestic applications [6]), in which they are required to establish autonomous interactions with *human* and *environment*. For example, they should be able to properly interact with humans by sharing attention, tracking objects, looking at human face/body, and by responding to the conscious and unconscious human social signals such as facial expressions, voice tone, speech, and gesture. In addition, they should be able to interact with the environment by showing dynamic attention behavior gazing at salient regions (e.g., bright/flicker points, sudden motions, etc.) that occasionally appear, or by reacting to human touch or the environmental light and temperature changes. To achieve the above-mentioned goals, social robots have to be equipped with perception systems able to gather and interpret the information of social world, similarly to humans.

Modeling perception systems for social robotics has been a big challenge over the last decades [7]. The common aim of perception systems development is to enable robots to have a reliable and acceptable interaction in various human-centered scenarios. For this reason, the previous efforts in this area investigated the problem from two main perspectives: social psychology, and robotics implementation. The first group identified several important high-level perceptual features by analyzing human social behavior (e.g., verbal and nonverbal cues) that have communicative roles in human daily interactions and thus, the perception of these features can lead benefit to HRI. The second group designed perception systems in order to enable robots to detect and analyze the identified high-level features. In the latter case, different robotic perception systems have been proposed, which partially support the perception tasks such as face or body detection and analysis (see Table I). However, a comprehensive system able to detect and analyze simultaneously a wider range of features (e.g., human- and environment relevant features) and track all the features in real-time, is missing. Moreover, each of the perception systems reported in Table I has been designed for the robot with specific software/hardware specifications, and therefore they cannot be used interchangeably.

In addition to the platform incompatibility, most of the developed systems have their own limitations in the interpretation of the social features, which prevent an acceptable HRI. For example, a face detection system detects any face-like shape, as human face, and without awareness of the real-world content. Such a perception system may address to the robot, a face printed on a sheet of paper as the interaction partner or it may consider an environmental sound as a person who speaks. This limitation can lead to a failure of HRI.

Designing a unique perception system able to detect a wide range of features with the awareness of real-world contents, and compatible with different robotics platforms, has appeared as a complex problem to solve. Such a perception system should have its own technical requirements as follows:

- *Modularity:* allows developer to decompose the perception system into a number of components that can be mixed and matched in a variety of configurations, in order to adapt the system to a specific robotic framework.
- *Interconnectivity:* system components should have the capabilities to connect and exchange resources and in-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



formation in some ways, in order to infer and deliver the real-world content rather than only the detected features.
- *Extendability:* it allows extending perception system by easy adding/removing or replacing system components and the corresponding software libraries without affecting other modules.
- *Communication:* it allows perception system to transfer all the extracted perceptual information to other machines running in different operation system.

The community of HRI would benefit from having such a social perception system compatible with most of the common robotic platforms and easily adaptable to the several social HRI scenarios.

In this work, we present a social perception system (SPS) that is able to perceive and create a higher-level description of the world, based on the pre-defined templates that we have defined for human and environment. SPS has been designed to meet the important technical requirements of a standard perception system. It has four main layers: data acquisition, high-level feature extraction, meta-scene creation and data communication. Through these layers, it manages the data flow constructed from the social world and generates high-level signals and transfer all the data to other machines/robots. SPS collects sensory information through Kinect, RGB camera, and Zerynth Shield[1], extracts high-level human and environment relevant features through several perception modules and components, and provides high-level perception merging all the detected features. In summary, SPS can be connected to the input sensors and the robot will receive the social perception of the real world as a dynamic data holder called meta-scene, with the standard structured format (XML script), which is readable in any machines and can be processed by any framework and middleware used in social robotics.

## II. Background and Related Work

In the last decades, perception modeling for HRI applications has received substantial interest from the robotics community. Many research projects have demonstrated the capability for a robot to interact with humans, perceiving people through vision sensors, listening to people through microphones, understanding these inputs, and reacting to people in a humanlike manner. The common goal of the previous efforts in this field was to enable robots to detect and interpret *high-level features* (e.g., facial expressions, body gestures, voice direction) analyzing low-level sensory information such as video and audio streams. Motivated by this fact, numerous methods and techniques have been proposed, which partially supports perception modeling. The type of the high-level extracted features usually depends on the robotic applications. For example, in a social robotics context, robots are required to interact with humans, and thus, the features to detect are human face, facial expressions, body gestures, and speech; or in an industrial robotics context, robots have to recognize the object to grasp and the obstacle to avoid when navigating.

As shown in Figure 1, the general workflow for the perception of a high-level feature has three main stages: (A) data acquisition, (B) low-level feature analysis, and (C) high-level feature extraction [8]. Through these stages, a perception system creates a high-level description of the world that is provided to the robot for higher level control process, e.g., understanding, learning. *Data acquisition*, delivers the raw data of environment acquired by different sensors. The aim of *low-level feature analysis* is to reduce the dimensionality of the acquired raw data and provide a more efficient description of the information to the following stage (e.g., for video data, it results in detecting features such as shape, texture, edge, etc.) The low-level features have to fulfill the important requirements of the algorithm used for high-level interpretation. For example, if the perception task is detection of an object with a specific color, the low-level features cannot miss the chromatic information, or in detection of two different objects, the shape is usually determinant. Finally, in the last stage, the *high-level features* are inferred analyzing low-level features, using different computer vision/audition based algorithms.

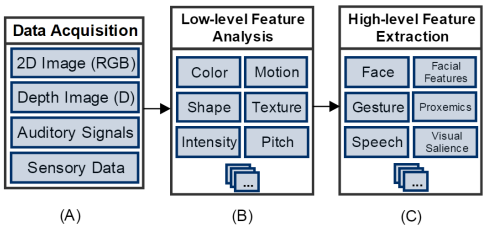



Figure 1. The general workflow of a perception system consists of three main stages: data acquisition, low-level feature analysis, high-level feature extraction

In the following, we shortly discuss the perception workflow presented in Figure 1 and reviewing the previous works, we enumerate the corresponding extracted high-level features. In particular, we are interested in real-time perception systems that have been developed for HRIs. Besides, we discuss from social psychology point of view, the importance of high-level features in human social interactions.

### A. Data Acquisition Systems

Visual signals have been widely used in HRI to achieve important perception tasks such as face detection and facial features analysis, face recognition, body tracking, gesture recognition, and salience identification.

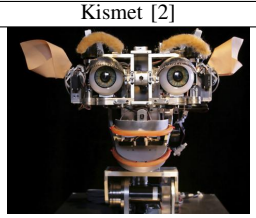
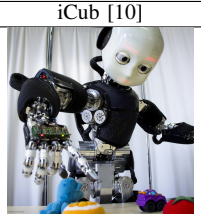
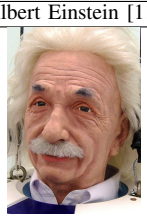
For visual scene perception, the RGB camera is one of the most conventional solution. This device provides a bidimensional and trichromatic representation of the light reflected or emitted by the surrounding. Cameras can be selected according to the requirements about chromatic sensibility, resolution, frame rate, optics (depth of focus, field of view), etc. However, a single RGB camera lacks depth information that help in perception tasks, such as proximity estimation. The common approaches to overcome such a limitation include image processing (e.g., depth from defocus [12]), and using a system of cameras exploiting disparities among images (e.g., stereo camera [13] or array of sensors). Both of the solutions need extra-processing to deliver the acquired depth data. This is reflected as an increase of computational time and data accuracy.

[1] *http://www.zerynth.com/*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



Table I
SOCIAL ROBOTS AND THEIR REPRESENTATIVE PERCEPTION CAPABILITIES

| Name | Kismet [2] | KASPAR [3], [4] | ASIMO [9] | iCub [10] | Albert Einstein [11] |
|---|---|---|---|---|---|
| **Social Robot** |  |  |  |  |  |
| **High-level Perception Features** | -object detection,<br>-face recognition,<br>-emotion recognition | -multiple points touch detection<br>-object recognition<br>-gesture recognition | -face recognition,<br>-human tracking,<br>-touch detection,<br>-sound discrimination | -human detection,<br>-object tracking,<br>-touch detection,<br>-sound localization | -face tracking,<br>-face recognition,<br>- facial exp. recognition,<br>-object tracking,<br>-speech recognition |

Recently, RGB-D sensors such as Kinect or ASUS Xtion have been widely used as convenient and successful solution. These devices simultaneously deliver RGB and depth images. However, the efficiency of the RGB-D sensors' cameras are relatively lower than the stereo cameras, but they have good performances in indoor perception systems. The main drawbacks are the limitation of field of view (smaller than 80 °) and the depth range (from a few cm up to 4 m), the low resolution of the depth image, and unusability in outdoor environment.

Auditory signals are often used as a secondary source of sensory information. For this reason, audio signal analysis has gradually achieved a key role in social robots' sensory systems. The auditory perception systems in HRI are focused on audio signal processing to extract high-level features such as speech, spoken utterance, affective state, etc. The auditory signal acquisition has been usually performed using either a single microphone or an array of microphones to improve perception tasks. The built-in microphone array of Kinect provides a convenient solution to integrate on a unique device, the acquisition of both RGB-D images and auditory data.

In addition to the visual and auditory perceptions, tactile-based perception is also important to extend the robot perception capability to perform physical HRI. Various tactile sensors are now available that can detect human touch [14]. The sensors can be placed either on the robots' body or in an object. However, tactile-based perception development is not as popular as visual-auditory perceptions and its application is limited for specific purpose applications.

### B. Low-level Feature (LLF) Analysis

The main aim of low-level analysis is to provide descriptors (e.g., color, texture) from the input data, which are functional for the high-level feature extraction stage. The type of the LLF that can be extracted is limited by the available raw data. For example, some low-level features such as color, motion, or texture can be extracted analyzing RGB image while some others require depth information. Since perception of HLF is the focus of our work, here we do not describe the details of the state-of-the art LLF analysis methods.

### C. High-level Feature (HLF) Extraction Methods

The high-level feature extraction is the task of inferring HLF analyzing LLF. For example, *shape* is effective for human body and face detection while *texture* is important for facial expression and feature analysis tasks. In this section, we review some of the representative work on HLF perception modeling discuss from the social psychology point of view, the importance of the representative HLF in human-human as well as human-robot interactions.

*1) Face Detection/Recognition and Facial Features Analysis:* Human *face* is a rich source of social information that people convey when interacting with each other. As stated by Argyle [15], facial expressions are the most important nonverbal channel for expressing attitudes. Through facial expression, we express our intentions and emotions, emphasize on a verbal message, and regulate a daily social interaction. Facial features e.g., facial expressions play an important role in social interaction [16]. Through facial expressions, we convey information about emotional states and we initiate or terminate a social interaction. In addition, other facial features such as human age and gender are fundamental since they might change the meaning of other nonverbal features. Therefore, the perception of the face and the capability of analyzing facial features are essential for social robots. While face detection and facial features analysis are trivial tasks for humans, their extraction is a complex issue in perception modeling.

To figure out how a perception system can detect a face in an image and track in a video analyzing low-level feature descriptors, several computer vision based approaches have been proposed. Yang et al. [17] have grouped the face detection methods into four categories: knowledge-based methods, feature invariant approaches, template matching methods, and appearance-based methods. In the *knowledge-based* methods [18], human knowledge and assumption of faces (e.g., each face has two symmetric eyes and a nose with a known distance and geometry) is being coded into a rule set in order to locate the face position in an image. *Feature invariant* approaches detect faces analyzing those facial features (e.g., skin color, texture, face size and shape), which are invariant in different head pose and lighting conditions. In the *template matching methods*, a standard face pattern is being parameterized to be used as template to detect faces in an image [19] while *appearance-based* methods use statistical analysis and machine learning methods to learn face pattern from a training data set. The face tracking task is possible by applying the described methods on a sequence of images in real-time.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



Face recognition and facial features analysis (i.e., facial expressions, age, gender) are relatively more complex than face detection. The difficulty of these tasks is associated to variances of head pose, perspective, and environmental conditions, presented within both the image and the training dataset. For that, several works have been proposed different pose-invariant face recognition techniques, which can be grouped into four categories: pose-robust feature extraction approaches, multiview subspace learning approaches, face synthesis approaches, and hybrid approaches. Ding et al. [20] presented a comprehensive survey on pose-invariant face recognition techniques and nicely reviewed and compared the strategies, performance and pros/cons of the mentioned categories. In spite of the recent progress and technologies, face recognition task is still complex and challenging task for system developer.

*2) Body Tracking and Gesture/Posture Recognition:* Body posture and gesture are important nonverbal signals that people use to intuitively communicate social information when interacting with others. People usually use these signals in conjunction to other cues such as speech and facial expressions, in order to attract other people's attention, to express their emotions and intentions, and in general, to manage the flow of social interaction [15]. The word posture refers to the way a person sits or stands while the word gesture is defined as continuous body movements that conveys a social message to others. Due to the importance of these signals, development of a perception system to interpret posture and gesture is a key priority, which improves the quality of social exchange between human and robots.

From the technical point of view, gesture recognition refers to the task of labeling spatio-temporal patterns in video sequences, based on a pre-trained observation model. The perception of body gesture can be done in two sequential steps: body detection, and body gesture classification. Motivated by this fact, several works have been proposed to detect the body region presented in the image through analyzing LLF by using image-understanding techniques such as Kalman filter [21], condensation algorithm [22], boosting method [23], nearest-neighbor classifier [24]. They have been implemented to infer the contour of human body analyzing LLF such as color and shape, motion, Haar-like features [25], and laser row data, respectively. The gesture classification requires LLF discriminative with respect to human gesture. A classifier that has been pre-trained by the extracted features should be employed to distinguish different gestures. Following this notion, several classifiers (e.g., hidden Markov model and neural network [26], [27]) have been previously proposed by different research groups.

*3) Proxemics:* Proxemics, the use of physical distance, is a subcategory of nonverbal communication study that was firstly introduced by anthropologist Edward Hall in 1963. In his famous book (*The hidden dimension*) [28], he emphasized that the humans Proxemics behavior (use of space) influences the implicit and explicit interactions with others. He defined four spaces called 'reaction bubbles,' in human peripheral surrounding environment, which are circles located around the human body at varying distances. These distances are the amount of space that people feel necessary to set between themselves and others for different social exchanges. Proxemics has been recently considered also in HRIs [29]. Holthaus et al. [30] defined a Proxemics-based behavior model that controls a social robot's behavior based on the human distance. Mazzei et al. [31] considered human Proxemics, in a rule-based engine, which controls the gaze shift and facial expression of a social humanlike robot.

*4) Speech Recognition:* Speech is the most natural and intuitive form of communication between humans and therefore has attracted great interest from researchers. Given the objectives of the perception modeling, development of accurate speech recognition is one of the key research priorities. The recognition of speech requires processes that are similar to those used for gesture recognition since the speech pattern should be extracted from low-level sensory information. For this reason, the researches in this field concentrated on low-level feature detection, and classification methods to detect specific speech patterns. For example, Betkowska et al. [32] proposed robust speech recognition using factorial hidden Markov model for home environments. Many efforts have been done in this area and several commercial speech recognition software are now available notwithstanding, the speech recognition problem is still an open challenge.

*5) Visual Salience Detection:* Visual salience as an environment-relevant feature refers to the distinct subjective perceptual quality, which makes some items of a visual image stand out from their neighbors [33]. The items can be for instance an object, a red dot on a wall, a bright/flicker point, a sudden motion, etc. As proven by previous works, the salience region can immediately grab human attention and thus, the perception of salience is important for HRI.

The detection of visual salience region plays an important role in simulating and implementing human attention system on social robots. The basic approach for identifying salience region is to detect locations whose attributes significantly differ from the surrounding regions in terms of color, edge orientation, luminance, and motion direction [34]. One of the seminal works in perception of salience in task-free condition (i.e. bottom-up attention) that influenced a large number of research, was proposed by Itti et al. [35]. In spite of its good prediction in detecting salience regions, it operates relatively slow and it is not suitable for real-time human-robot interaction applications. Zhang et al., [36] proposed a computational model aka SUN (Saliency Using Natural statistics), based on a Bayesian framework trained on a collection of natural images. On the basis of that, Butko et al. [37] implemented a simplified version of the original model that is designed to operate in real time with reduced computational cost, making the integration in a robot perception system more feasible.

As discussed, the common aim of developing perception system is enabling robots to perceive high-level social features using the computer-based techniques. As shown in Table I, different systems have been developed for different social robotic platforms. However, due to the complexity of the structures and implementation, they are only compatible to the specific robotic platform. Besides, as it has been shown in the Table, the perception systems of these robots are able to detected limited number of high-level features while the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



detection of a wider range of features is required. Above all of these, each of the perception capabilities shown in Table I has its own limitation in the interpretation of the social features that may prevent a natural and acceptable HRI. As far as we know from reviewing literature, nowadays there is no unique and efficient real-time social perception system with a standard structure, able to perceive the social world and deliver a wide range of perceptual features to robots. Motivated by these facts, we present a unique social perception system to fill the current existing gap and we think it would be beneficial for researchers in human-robot and human-machine interaction communities.

## III. Social Perception System (SPS) for Human-Robot Interaction

SPS[1] is an out-of-the-box perception system that has been designed to enable robots to perceive a *wide range* of social features with the *awareness* of their real-world contents, as humans do. Thanks to the SPS *compatibility* feature, it can be integrated to different robots irrespective of their working operation systems. Due to the SPS *modular* structure, it can be reconfigured and adapted to different robotics frameworks by adding/removing its perceptual modules.

SPS consists of four distinct layers (as shown in Figure 2): data acquisition, low-level and high-level features extraction, and meta-scene creation and communication. SPS collects visual-auditory information of environment, detects and tracks a wide range of high-level perceptual features, through different parallel perceptual modules. Then, it actively manages all the extracted information in a dynamic storage called meta-scene. The SPS data communication layer streams the created meta-scene out in real-time through YARP middleware [38]. In this section, we describe in detail, the four SPS layers and underlying perceptual modules.

### A. Data Acquisition

The data acquisition layer contains three main components that acquire sensory information constructed by Microsoft Kinect, RGB camera, and TOI shield [39]. The Kinect records 2-D image with a resolution of $1920 \times 1080$ pixels, and depth images with a resolution of $512 \times 424$ pixels at 30 fps with 70° horizontal and 60° vertical FOV wide-angle lens. For audio signal acquisition, it uses a four-element microphone array operating at 48 kHz [40]. TOI shield is a collection of multiple sensors that acquire environmental information required for human-robot interaction (e.g., illuminance and sound levels, and temperature). It provides the raw data in order to estimate environment features that could be relevant within the social interaction.

### B. Low-level and High-level Features Extraction

The feature extraction layer processes the acquired raw data in order to extract *human-relevant* features (i.e., face and facial properties, body skeleton and gesture, head pose, Proxemics and orientation) and *environment-relevant* features (i.e., visual salient point in pixel (x,y), illuminance and sound levels, temperature, touch and proximity information). It then manages and stores in real-time, all the extracted information in a dynamic storage with the standard structure. Each perceptual module of the feature extraction layer delivers the high-level features (HLF) analyzing corresponding low-level features (LLF), which are being extracted in each module. The algorithms/libraries and extracted HLF are summarized in the following section.

*1) Facial Features Analysis:* To detect and track human face as well as to analyze the facial features, we developed the facial feature analysis module using the sophisticated high-speed object recognition engine presented in [41], [42]. The module allows the quick detection of arbitrary number of faces, as well as the analysis of faces in image sequences, videos and single frames. Furthermore, the module estimates a variety of facial features such as gender, age, and facial expressions (see Figure 3). The module runs in real-time and it is able to detect faces down to a minimum size of $8 \times 8$ pixels. The module has a short-term memory for the recognition of faces that works fully anonymously: persons, who were out of the cameras FOV for a short term, can be recognized and recollected. Using the facial analysis module the feature extraction layer collects the HLFs summarized as follows:

- Positions of arbitrary number of faces' rectangle (in pixel).
- Positions of the eyes, nose, and mouth (in pixel).
- Gender classification (male/female).
- Age estimation (in years).
- Recognition of facial expressions ('Happy', 'Surprised', 'Angry', 'Sad') (in percentage).

The feature extraction layer detects and analyzes multiple faces with high accuracy (face detection with the rate of 91.5%, and gender classification with the rate of 94.3%) [41], assigns a unique ID to each of detected faces, and sends all the detected features in the meta-scene data storage.

*2) Identity Assignment:* In order to create a consistent description of the subjects involved in the long/short term interaction, it is necessary to recognize and associate a unique identity to each subject over the interaction time. This is mandatory especially for storing data in the meta-scene by reporting a list of the people with their associated extracted features. As discussed, the conventional identity assignment systems are usually based on the face recognition techniques, which can easily be affected by the performance of the processing machine, lighting conditions, head posture, and difficulties related to training stage. On the other hand, the identity assigned by other perceptual modules are not unique since they are changed when subjects leave the scene and re-enter the field of view. To avoid such issues that may affect all the perception and data storage creation processes, following [43], we implemented the identity assignment module in a smarter way.

The SPS Identity Assignment module is a smart Quick Response (QR) code reader, which allows a fast readability and performance with respect to the face recognition based systems. Once a subject enters in the robot's field of view, the system quickly scans the body region where the QR

[1]The relevant source codes are available online: *http://www.faceteam.it/*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



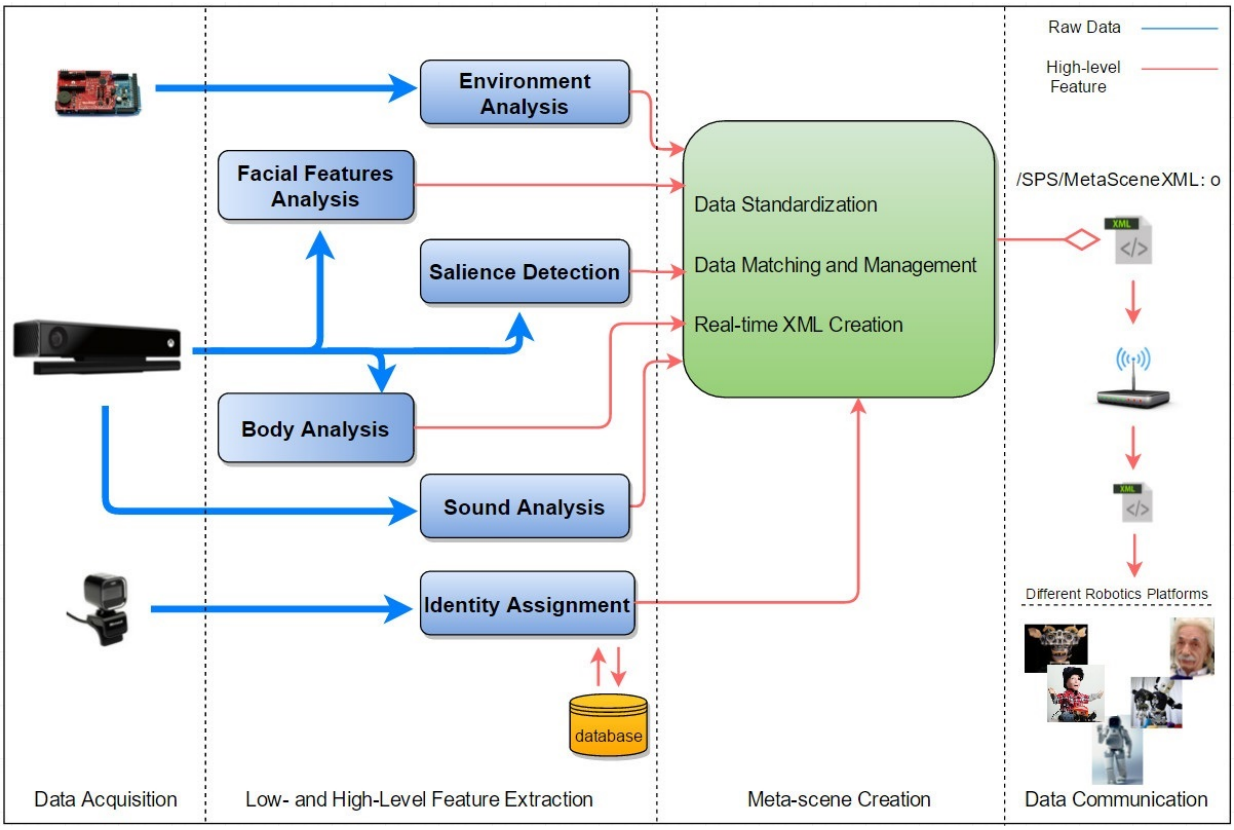



Figure 2. Modular structure of the social perception system (SPS): the system receives audiovisual information acquired by the sensors, and extracts human and environment relevant features. Based on these features, SPS creates meta-scene data and streams it out through a YARP port. The output of the system is an XML file that contains all the extracted high-level features

code is placed and assigns a unique identity to the subject according to the information stored in the code. In the current implementation, the system is able to easily track the subject in real time and match his/her identity information with the body analysis module. In addition to the subject recognition capability, the module is connected to a database, which enables SPS to remember subjects and the previous social interactions. The database can store any information about subjects and the overall social interaction in terms of its emotion, positive/negative markers, etc. In general, having the database is essential for social HRI, since it enables the robot to differently shape the long/short term interactions depending on different subjects and their experienced interactions with the robot.

*3) Body Analysis:* Using Kinect Software Development Kit (SDK), SPS receives RGB-D scene constructed by Kinect, and localizes and tracks up to six humans (see Figure 4). Thanks to the skeleton tracking library available in Kinect SDK, SPS extracts the following features of six humans and passes all the extracted features to the corresponding pre-defined template in the meta-scene. These features are then being used by body analysis module for the recognition of subject's body posture and gesture.

- 3D position of 25 body joints, in Cartesian space as $(X, Y, Z)$
- joint rotation such as wrist rotation (roll, pitch, yaw)
- head pose tracking simultaneously with body tracking
- finger tracking (hand tip and thumb)
- sound direction with voice filter
- voice recognition using speech-to-text converter engine

The main aim of developing body analysis module is to enable SPS to interpret human body gesture and posture analyzing bodily information captured by Kinect.

For this reason, we developed a dynamic posture/gesture recognition model that recognizes different per-defined gesture and postures, through algorithms of geometrical relations among subjects' body joints 3D positions. To test the posture recognition capability of SPS, we implemented some fundamental expressive gesture that SPS have to recognize, as follows:

- *seated/standing*: we distinguished between standing and seated person through computing the hip angle for the right and left joints of the body. Thanks to the SPS modular structure, this module can be easily updated in order to enable SPS to recognize customized gestures and postures required for a specific robotics application.
- *raiseHand*: when one of the hand is above the head. This gesture is usually associated with request of attention or wish to speak.
- *yeah*: when both the hands are above the head. This gesture is related to high arousal such as exultation.
- *headRubbing*: when only one hand touches the head and the elbow is directed outwards the body figure. This could have a lot of different meaning such as have forgotten

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



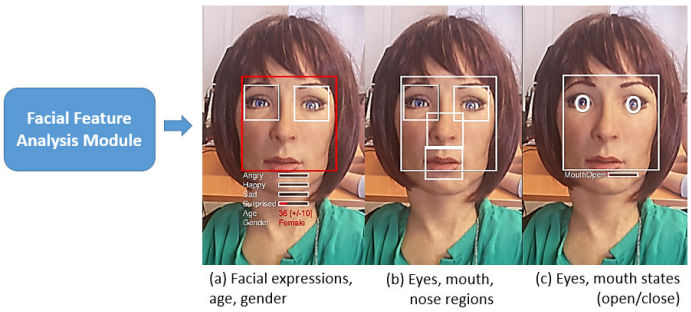



Figure 3. Face and facial features analysis module detects and tracks in real time, a range of human face's high-level features.

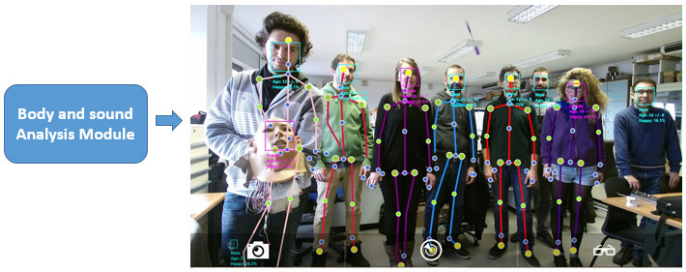



Figure 4. Multiple human tracking system: SPS detects and tracks in real time, human body and facial features. It detects also the direction of speaker comparing 3D position of human subjects and the detected sound angle (shown at the top of the figure).

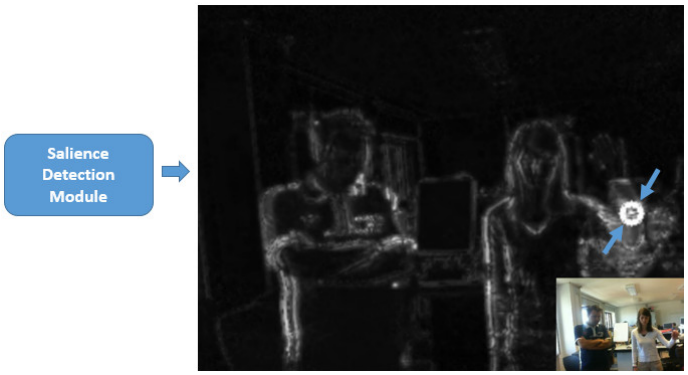



Figure 5. Salience detection module detects salient point/regions of a visual scene, analyzing low-level visual features (e.g., human body and objects regions, hand motion)

something, wondering about and headache.

- *armCrossed*: when each hand touches the opposite elbow simultaneously. This gesture signifies defensive or negative attitude about what is happening.

*4) Salience Detection:* The SPS's salience detection module is entrusted with the creation of visual salience map, which contains important regions of image (e.g., object), as well as the detection of important events, which dynamically occurs in the image stream (e.g., motions). Besides, the module predicts in real time the salient point of the image (see Figure 5) on the basis of a computational model of human attention system [37]. This module is important since it detects the visual salience, even outside the border of other modules' detection range (e.g., close and far distances). The information created by the salience detection module can be used by the robots to orient their vision sensors to the salience region/point, in order to analyze only that region in detail. This feature allows an important reduction of the SPS computational costs.

The module receives, as input, the Kinect's 2D image stream, and sends the salient point location (in pixel) to the meta-scene, as an environment relevant feature. The developed module can be run with a little computational cost and thus, it is well suited to real-time implementation perception system needed for social robotics. It runs for the 2D video stream converted from RGB value to grayscale, and down-sampled to 320x240 pixel in order to provide the most salience region in around 40-50 ms. The following steps summarize the mechanism:

- the input video stream is converted in grayscale
- Spatio-temporal features are estimated through center-surrounded operation in input data with difference-of-box and difference-of-exponential filters
- Salience map is computed by the natural statistic embedded in the model

### C. Meta-scene Creation and Data Communication

*1) Meta-scene Creation:* The aim of the layer development is twofold. Firstly, to process the extracted data in order to interpret and store HLFs with the awareness of their real-world contents, and then to create and deliver in the standard format, all the structured features of people and environment. As shown in Figure 2, the interpretation of features can be done in two steps: data standardization, and data matching and management. The standardization of data is done by scaling different HLFs provided by feature extraction layer, into the same numerical range. Once the information has been standardized, the interpretation of features will be possible based on the models that we defined for human and environment. Once the information satisfies the model requirement, the layer stores and delivers the information, otherwise it removes from meta-scene any conflicting data. For example, the system removes the information relevant to a subject, which only has face or skeleton, etc.

We defined human as an object with a face and body with underlying components. In our *human* model, the face has two eyes and a mouth (can closed or opened), a nose, a rectangle, four facial expressions, age, gender, and a unique ID (name). In addition, a body has 25 joints, a head with 3D rotation angles (r,p,y), a gesture, a posture, a proximity, and speaking probability. Moreover, our *environment* model has a salient point in every 40 ms, temperature, illuminance level, sound level, and an object which reports human touch and proximity information. As discussed, only information satisfying the human/environmental model are stored. For instance, as shown in Figure 4, SPS detected 10 faces, however 6 of them are accepted and stored in the meta-scene. Although, SPS detected a face-like shape at the left-bottom corner of the figure, the robot's face at the left side, and a face without skeleton at the right side, it does not store these data, since they are not well fitted in our defined models. Or as another example, SPS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



detects human as a speaker, only if the estimated sound angle and human direction become the same degree.

The meta-scene has a hierarchical structure through which an arbitrary number of people can be inserted. Each person includes a unique ID and the associated high-level features. The ID is assigned to the people merging the IDs assigned by other modules, based on the information of body and head positions. Once a new person has been identified by the identity assignment, a new human instance is being created in the meta-scene, which is populated with the features, extracted by the feature extraction layer. Through the NET object serialization, the created meta-scene data package is converted in real-time into a standard XML structure.

*2) Data Communication:* The aim of the communication layer development is to transfer the result of perception (meta-scene XML file) to robot controllers, which can be run either on the same perception machine, or in a different machine. This layer is important since allows the integration of SPS into different target robots, which are running in different operation systems. Motivated by this fact, we used YARP middleware to the communication layer in order to support data transferring in a peer-to-peer way.

## IV. Evaluation Process

Our aim is not only providing a platform-independent SPS but also to deliver useful information about its practical use. Indeed, when integrating different software modules is very hard to understand the simultaneous reliability of each output, we solve this issue by conducting an extensive analysis on a standard social scenario and by using SPS with a Kinect ONE sensor. This is expected to allow to easily develop applications and/or experimental protocols in social HRI contexts.

SPS includes several perceptual modules, which extract entire social-relevant features included in a complex visual-auditory scene. The output of SPS is a collection of features ($f_i : i = 1, N$ where $N$ denotes the number of total features that SPS managed to provide). Each feature can be extracted only if the data provided by the sensor allows the specific module to perform a correct analysis. The performance of the specific feature-detection capability of SPS is partially related to the test performed by developers of each library. For instance, face and facial analysis detection rates are as follows: face detection rate is 91.5% of frontal facial detection (CMU+MIT database), 94.3% for gender detection (Bio-ID database), and 6.85% mean absolute value error in age estimation (FG-NET database) [41]. Moreover, a face can be detected till 90° of out-of-plane rotation and up to $8 \times 8$ pixels in size. Here, we are not interested in replicating the results about the accuracy of each single module and instead, we work to evaluate the entire SPS framework based on Kinect ONE sensor in order: (i) to demonstrate the simultaneous delivering of features by SPS, as well as the working area of each feature called Feature Map (FM), (ii) to develop a toolkit that can be used to assess which areas will be useful to receive which kind of information, and by which module. Having this toolkit, anyone is able to insert some data as input (i.e., position of the Kinect, type of human [standing/seated, elder/children, etc.]), computing the virtual map of the scenario, in which the experiment has to be set. We performed several analysis and experiments to deliver each FM based on the specific sensor setting.

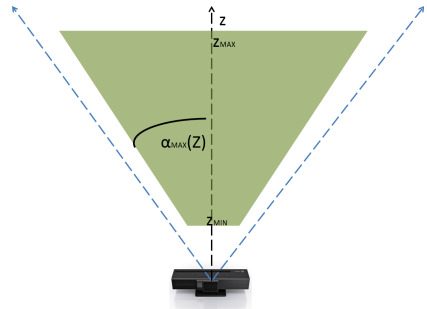



Figure 6. General Feature Map (FM) referred to the Kinect One perspective presented by $Z_{min}$, $Z_{max}$, and $\alpha(Z)$.

Table II
THE STANDARD HUMAN SIZE OBTAINED FROM [44]

| | | |
|---|---|---|
| Person Height | t | 180cm |
| Neck-shoulder (Shoulder Center- Shoulder Right/Left) | D | 30.5cm |
| Shoulder-Palm (Shoulder R/L,- Hand R/L) | F | 64cm |
| Face Width | S | 24cm |

As shown in Figure 6, we defined FM as a subsection area of the sensor FOV in which the data of a certain feature delivered by SPS are reliable. To identify FMs, the boundaries are reported in term of human head coordinate. A feature is available when the subject head is within $Z_{min}$ (bottom limit for depth), $Z_{max}$ (top limit for depth), and $\pm\ \alpha_{max}(z)$ (absolute top limit for angular eccentricity function of depth) considering the symmetric response of the sensor. In this section we provided a detailed description of the method used to estimate $Z_{min}$, $Z_{max}$ and $\alpha_{max}$ for each feature provided by SPS.

### A. Strategy for FMs Identification

In order to identify $Z_{min}$, $Z_{max}$, and $\alpha_{max}(z)$, we perform a mathematical analysis that considers both sensor limitation and the specifications of each software library shown in Figure 2. In addition, the analysis includes the parameters describing standard human body size (reported in Table II) and sensor's height ($T_{kinect}$). However, such parameters can be adjusted in order to fulfill the application requirements. The strategy for FM identification is related to the sensor position and orientation. For instance, a navigating robot that tracks elder people needs to analyze adult body size from standing point of view, or a story-telling robot dealing with children needs to analyze pupil body size from a seated point of view. Without loss of generality, we perform experiments with male students involved in a HRI with a seated humanlike robot where the Kinect and TOI board placed relative to that robot. Finally, a MATLAB source code is provided that allows simulation of SPS' FMs with the robot in different conditions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



### B. Experimental Setup and FMs Validation

To simulate a scenario in which a seated robot interacts with students, we set up an experiment shown in Figure 7. The proposed setup is similar to [45] for an assistive robot that coached elder people in performing exercise. The Kinect sensor was parallel to the ground at 1.55m in height ($T_{kinect}$) simulating a point of view of a seated person. The indoor environment was set up with distances and eccentricity angles marked on the floor, and the students were allowed to interact with the robot from any marker of this area. As shown in Figure 7 the sampling for distances were at $[0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 6, 7, 8]$ m and for the eccentricity angles were at $[0, 10, 20, 30, 35]^\circ$. In sampling the room, we considered the accuracy of the 3D coordinates provided by the Kinect SDK v2. In fact, we assume the tolerance of $\pm 10$ cm for depth and $\pm 1^\circ$ for angle eccentricity in the overall working space. These values are larger than the system resolution (for Kinect sensor v1, depth resolution: 5 cm (@z= 4m [46]), angular resolution: 0.086° (@z=2m [47])). The tolerances are limited by subject standing uncertainty. We calculated the average deviation standard for all the body joint coordinates when asking to the same subject, to stay again on the same place on the ground. Values for the tolerance are given at the 95% confidence interval. Therefore, in order to have data that can be considered reliable, distance and eccentricity angle are sampled with resolution not smaller than 5 times the tolerance limit.

In addition, only half of the horizontal field of views (HFOV) was sampled considering the symmetry of the sensor response. The boundaries at $\pm 35^\circ$ are given by the FOV of the depth maps. Distances closer than 0.5 m are not relevant for the purpose of the system. The distance 4.5 m is the upper limit to detect skeleton recommended by Microsoft for Kinect ONE. Linear interpolation of the result is assumed in the region within two sampled points. The experiments were performed in the main hall of University of Pisa, while two indoor light projectors controlled the light conditions.

In order to fulfill with the proposed scenario, we selected 4 male students whose average body sizes fulfil the parameters reported in Table II. The subjects were instructed to perform specific actions to allow SPS to detect a range of features. The actions are for example entering the experiment space walking in the FOV, showing different facial expressions and body postures, sitting and standing, and speaking.

## V. Results

This section reports the analysis and the experimental results conducted to estimate and validate each FM for the features delivered by SPS. Figure 8 shows the obtained result, the identified FMs for the described HRI scenario. Each FM reports the capability of SPS to detect and analyze specific kind of information delivered by humans from a given location within the sensor FOV. Figure 10 nicely illustrates the intersection of all FMs and corresponding high-level feature in a table. In the red region, SPS delivers all the features at 100% of the potentiality provided by the different perceptual modules. For that, we called this region as the *interaction space*, which is the best subsection of FOV for HRI. We will discuss it in more detail in Section VI.

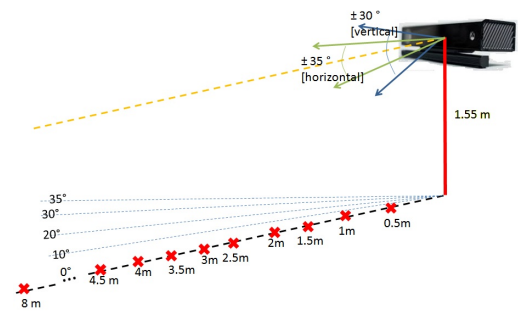



Figure 7. The experiment setup: several points have been marked on the floor, in different distances and eccentricities.

### A. Face Detection

Since facial feature analysis module requires the face rectangle in the image in order to provide all the facial properties, firstly, we aimed at obtaining the FM for face detection, where SPS is able to detect faces. Only one subject at the time was in the scene during the experiments, the system capability has been proven in multi-party interaction, as it is shown in Figure 4.

We define the close distance ($Z_{min}$) when face exceeds frame edges and far distance ($Z_{max}$) when a face image became too small or too noisy to be detected. These boundaries were measured with full-frontal position of the face at the different marker positions (see Figure 7). The face size (reported in Table II) is the main limitation of the detection process. Accordingly, the lower limit ($Z_{min}$) and the angular limit ($\alpha_{max}$) can be calculated as follows, where S parameter shows the subject face width

$$Z_{min} = \frac{0.5S}{\tan(hHFOV)} \tag{1}$$

$$\alpha_{max} = \arctan(\tan(hHFOV) - S/2z) \tag{2}$$

Since the upper limit ($Z_{max}$) of the face detection FM exceeds the experimental setup edge, we identified the detection range and set it to eight meters. In addition, the angular limit of the FM is limited also by non-fully-frontal faces. Therefore, we define the effective out-of-plane rotation of the face ($e\omega_{face}$) given by the following equation:

$$e\omega_{face} = \varphi_{face} - \omega_{head} \tag{3}$$

where $\omega_{head}$ is the head eccentricity and $\varphi_{face}$ is the angular rotation of the face.

We considered three different conditions and consequently three separated FMs: when the subject looks at the sensor called *inCam*; when the subject looks straight ahead called *ahead*; and when the subject turns the head around Yaw called *turned*. The perception module is able to perform face detection as follows [48].

$$|e\omega_{face}| \leq 60^\circ \tag{4}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



Table III
THE FEATURE MAP (FM) OF FACE IN DIFFERENT POSTURES: INCAM, AHEAD, AND TURNED

| | $Z_{min}(m)$ | $Z_{max}(m)$ | $\lvert\alpha_{max}(z)\rvert$ |
|---|---|---|---|
| inCam | 0.5 | 8 | 35° |
| Ahead | 0.5 | 8 | 35° |
| Turned | 0.5 | 8 | Min($\lvert 60-\varphi_{face}\rvert$,$\lvert -60-\varphi_{face}\rvert$) |

Analyzing these conditions, we observed that:

- For inCam the $e\omega_{face}$ is always equal to zero since by definition $\varphi_{face} = \omega_{head}$. For that, the face is always frontal from the sensor point of views, and thus it should not limit the detection (see Figure 8.a).
- For Ahead $\varphi_{face} = 0$ and then $e\omega_{face} = -\omega_{head}$. Therefore, the actual head position of the subject in the field of view of the sensor is the factor limiting the detection of the face.
- For Turned we consider the maximum deviation for a given $\omega_{head}$. To not exceed $|e\omega_{face}| \leq 60°$, larger is the subject head eccentricity smaller is the absolute value that he/she can turn the head outward.

Table III reports the boundaries of the face detection FMs for all the above-described conditions. Note that the $Z_{max}$ exceeds the depth sensor range (i.e., 4.5 m), and thus, a subject face can be detected even without a skeleton detection. It shows the capability of the module to work with small faces ($8 \times 8$ pixel).

### B. Unique ID Assignment

Once the subject skeleton has been detected by the feature extraction layer, the module uses the skeleton ID provided by the body analysis module, and updates this ID with the name of the subject, as soon as the subject's QR code has been detected. The detection capability of the system depends on the size of the QR tag. Obviously, with increasing the size, the detection range become wider. In the current implementation, SPS detects the QR tag with the size of $(5.5cmX5.5cm)$, in the range of $[30cm, 130cm]$ from the Kinect sensor.

Since some of HRI studies are focused on the long-term aspects of interaction, social robots need to remember subjects and some remarks from the previous interactions. For this reason, we designed an SQL database, which serves as robot's memory. It is a structure storage, where it can be filled with subjects' name as well as the positive/negative markers that robots previously assigned to the subjects.

### C. Skeleton Tracking

Gesture, posture and subject recognition need skeleton tracking (SkT), which is provided by body analysis module for up to six subjects. We distinguish among the horizontal plane (shoulder-shoulder) SkT, vertical plane (head-feet) SkT, and full SkT. Each of them has a different FM and allows to recognize only those poses and/or gestures outlined by the available skeleton joints. For instance, without a whole vertical SkT (head to feet) it is difficult to discriminate between standing or seated person. SkT both for vertical

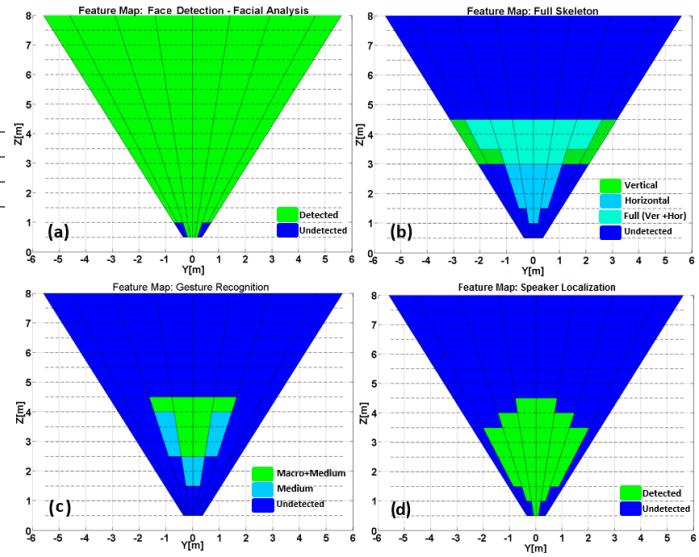



Figure 8. The graphical demonstration of the different feature maps (FMs) for the high-level features delivered by the Social Perception System (SPS)

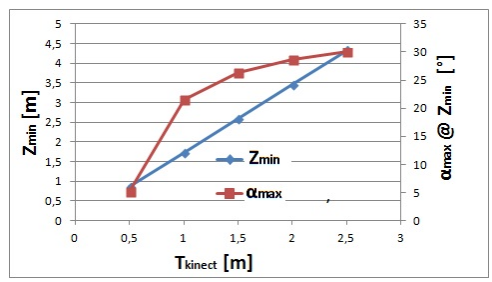



Figure 9. The effect of sensor height on depth and eccentricity for the bottom limit of the feature map of full body tracking

and horizontal plane depends also on the sensor point of view. Assuming the optical axis parallel to the ground, the boundaries of FM for full SkT can be expressed as function of $T_{kinect}$ (Kinect Sensor height), $D$ (shoulder-shoulder distance), Kinect horizontal HFOV (hHFOV), and Kinect vertical HFOV (vHFOV). Table IV reports the FM limits both for general condition and for experimental setup. The three different FMs are shown in Figure 8.b.

For the Kinect placed at 1.55m in height, a full body tracking (shoulder-shoulder + head feet) is possible only beyond 2.7 m for a FOV $\pm 30°$. $T_{kinect}$ can be adjusted to fulfill the requirements of the specific application. This would lead to vary directly $Z_{min}$ and indirectly $\alpha_{max}(Z_{min})$, as reported in Figure 9.

### D. Gesture and Posture Recognition

The tasks of gesture recognition (GestR) and posture recognition (PoR) require the tracking of all the joints involved in the specific gesture or related to the posture. For PoR, to discriminate between the implemented pose, seated or standing person, a full body tracking is required. Therefore, we can assume the FM for PoR coincides with Full body tracking FM. In the experiment, subjects were asked to stand near or seat on a chair for 20 times in one marker between [2.7 , 4.5]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



Table IV
THE FEATURE MAPS (FMS) FOR HUMAN BODY SKELETON

| | | $Z_{min}$(m) | $Z_{max}$(m) | $\lvert\alpha_{max(z)}(^\circ)\rvert$ |
|---|---|---|---|---|
| Sensor limits | | 0.5 | 4.5 | $35^\circ$ |
| Horizontal SkT | Gen. cond. | $\text{Max}\left(0.5, \frac{D}{\tan(hHFOV)}\right)$ | 4.5 | $\arctan\left(\tan\left(hHFOV\right) - D/z\right)$ |
| | Exp. Setup | 0.5 | 4.5 | $5^\circ @z = 0.5m, 32^\circ @z = 4.5m$ |
| Vertical SkT | Gen. cond. | $\text{Max}\left(0.5, \frac{D}{\tan(vHFOV)}\right)$ | 4.5 | $\arctan\left(\tan\left(hHFOV\right) - D/z\right)$ |
| | Exp. Setup | 2.7 | 4.5 | $30^\circ @z = 2.7m, 32^\circ @z = 4.5m$ |

Table V
THE FEATURE MAPS (FMS) FOR HUMAN BODY GESTURE IN THE GENERAL CONDITION AS WELL AS IN THE EXPERIMENT SETUP

| | $Z_{min}$(m) | $Z_{max}$(m) | $\lvert\alpha_{max(z)}\rvert(^\circ)$ |
|---|---|---|---|
| Gen.,cond. | $\text{Max}\left(0.5, \frac{Tkinect}{\tan(vHFOV)}, \frac{D}{\tan(hHFOV)}\right)$ | 4.5 | $\arctan\left(\tan\left(hHFOV\right) - D/z\right)$ |
| Exp. setup | 2.7 | 4.5 | 30° @z=2.7m,32° @z=4.5m |

Table VI
THE FEATURE MAPS FOR MACRO AND MEDIUM GESTURES

| | | $Z_{min}$(m) | $Z_{max}$(m) | $\lvert\alpha_{max(z)}\rvert(^\circ)$ |
|---|---|---|---|---|
| Macro movement | Gen. cond. | Max $(0.5, \frac{D+F}{\tan(hHFOV)})$ | 4.5 | $\arctan(\tan((hHFOV) - (D+F)/z))$ |
| | Exp. setup | 1.35 | 4.5 | 4° @z=1.5m, 29° @z=4.5m |
| Medium movement | Gen. cond. | Max $(0.5, \frac{D+F}{\tan(hHFOV)})$ | 4.5 | $\arctan(\tan(hHFOV) - (D+\epsilon F)/z)$ |
| | Exp. setup | 0.9 | 4.5 | 15° @z=1.5m, 29° @z=4.5m |

m in distance within 30° of $\alpha_{max}$. The system was able to discriminate the postures of the subjects in 100% of cases.

On the other hand, for GestR, all the body joints are required. We have to consider that humans performing gestures move arms freely, and thus they require a larger area than the one considered for SkT. For this reason, we distinguish the gestures between macro-movement (involving the maximum possible extension of the arms) and medium-movement (involving a reduced portion of the possible extension of the arms). In this case, we adopted the same approach as for Horizontal SkT assuming a horizontal encumbrance given by 2(D+F) for macro-movement and 2(D+F$\epsilon$) for medium-movement (D and F are reported in Table II). The value of D and F are reported in Table II. Instead, $\epsilon$ represents a fraction of lateral displacements of the arm used for medium-gesture and it is assumed 0.5.

Among the implemented gesture,"'Yeah"' and "'raiseHand"' are considered as macro-movements, and "'armCrossed"' and "'headRubbing"' are classified as medium-movements. Figure 8.c reports the FM for both medium and macro gestures were tested by asking the subjects to perform different gestures looking directly to the camera 20 times, within different regions. The system was able to discriminate perfectly as long as the subject's skeleton is available.

### E. Speaker Localization

The FM of speaker localization mainly depends on angular accuracy of sound detection. Since this data is not given by the Kinect sensor specification, we perform a specific experiment to estimate it. We compare the sound angle direction ($\omega_{sound}$) and the speaker head position ($\omega_{head}$). Each subject is associated to a speaking probability inversely proportional to the following equation

$$\Delta\omega = (\omega_{sound} - \omega_{head}) \tag{5}$$

The experiments have shown the factors that drastically compromise the localization of the speaker are the environmental sound noise, echo in indoor scenario, and multi-parties interaction with people speaking simultaneously. Here, we consider a scenario with no-echo, controlled environmental noise, and only one person speaking at the time. In order to estimate a FM for the speaker recognition, we asked subjects to move within the environment and stand on one of the ground marker. Alternatively, one of them started to talk for five second, while we recorded $\Delta\omega$. All the position is sampled for five times. Figure 8.d reports the FM for the speaker localization. This FM is obtained considering that two speakers could stand next to each other at the same distance and the system should be able to distinguish between them.

## VI. DISCUSSION

This section summarizes some important remarks regarding SPS and the evaluation results. Each FM reports the capability of SPS to detect and analyze specific kind of information delivered by humans from a given location within the sensor FOV. We define the *interaction space*, where the SPS detects 100% of human-relevant features required in the HRI scenario and used in building the meta-scene. For instance, a coaching robot assisting one or more elders in performing body exercises [45] needs all the features delivered by SPS. Face detection and skeleton tracking are needed for the purpose of subject recognition, gesture and pose recognition help to detect human activities, and speaker localization is required to discriminate the source of voice commands. In such a case, the interaction space is given by merely overlapping all FMs.

Despite of the specific application, FM of face detection/analysis and FM for skeleton tracking in horizontal plane have to be included in the interaction space, since SPS exploits such an information to discriminate a human subject when delivering the meta-scene.

The interaction space can be described using two main parameters: minimal distance of interaction (mDI) and total interaction area (TIA) (see Figure 11). The former enables a specific kind of HRI interaction according to the Proxemics theory. The latter parameter is linked to number of subjects and their activity within the interaction. For instance, a seating or standing adult in a rest position requires only 0.27 $m^2$, but this area increases up to 2.63 $m^2$ when performing activity with upper limbs [44]. The ultimate constraints of the perception system are given by the sensor. For Kinect ONE is possible to track up to 6 subjects in the distance range from 50 cm to 4.5 m. So the minimal distance of interaction is within personal space of Proxemics spaces ($mDI_{limit} = 0.5m$) and $TIA_{limit}$ is equal to 14 $m^2$.

We evaluated the SPS working space for a scenario simulating a seated robot interacting with adult subjects. The overall working space of the perception system is shown in Figure 10 where the red region shows the interaction space. According to Proxemics theory, a robot in this scenario could fully interact with humans in their social and public spaces (mDI equals to 3 m). In addition, a TIA of 2,72 $m^2$ allows the robot to interact with one adult subject who is free to perform body activity or alternatively, with 4-5 subjects resting and located in a way to avoid body occlusion. Examples of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



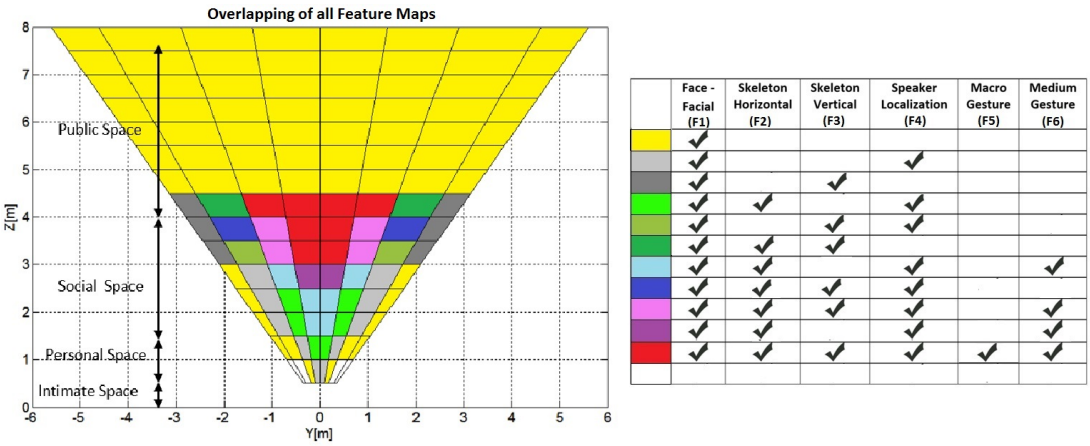



Figure 10. The graphical demonstration of the different FMs and corresponding high-level features. Left side of the figure shows the Proxemics spaces including intimate, personal, social, and public spaces

suitable HRI applications are an assistive robot for coaching elders in performing physical and psychological exercises or a synthetic tutor for children/young adults.

Within the analyzed scenario, a robot cannot physically assist humans or interact with them in personal and intimate space (e.g., bartender or nursery robot). Such a limitation could be faced by reducing the perception capability of the SPS (i.e., de-activating some of the modules) and/or modifying the Kinect height. The first solution is not always practical since it drastically reduces performance with the risk of making the robotic system unable to perform tasks within the interaction. Figure 11.A shows the change on mDI and TIA adopting this approach for the considered experimental setup. The second solution is feasible by adequate robot design. For a humanoid robot, the most natural solution for the sensor location is to integrate it within or above the head but if necessary, in any case, it could be integrated into other part of the body or placed remotely. Figure 11.B shows the effect of different Kinect heights for both mDI and TIA, when all the modules of SPS are active. Placing $T_{kinect}$ lower than 1.25 m allows a closer interaction with the larger interaction area. Instead, placing the sensor above 2.25 m the interaction area cannot be found since mDI exceeds sensor depth limit (4.5 m). In general, exploiting a hybrid solution is preferable and it can be easily found using proposed equations in this work and the MATLAB code provided as complementary material.

## VII. Integration of SPS in Two Real Robots: Practical Note

We describe our experience of integrating SPS with two real robots i.e., FACE and ZENO. In both cases, SPS entrusted with doing perception task to enable fluent HRIs. In these scenarios, SPS collected the perceptual information as instances of meta-scene and delivered them to the social agent control machines using YARP middleware. We ran SPS on a machine with Intel CPU Core (TM) i7 - 3.4 GHz, 16 GB RAM with Windows 8.1 Enterprise operating system. As shown in Figure 12, FACE robot [49], [50] is a humanlike robot with female appearance able to express realistic facial expressions

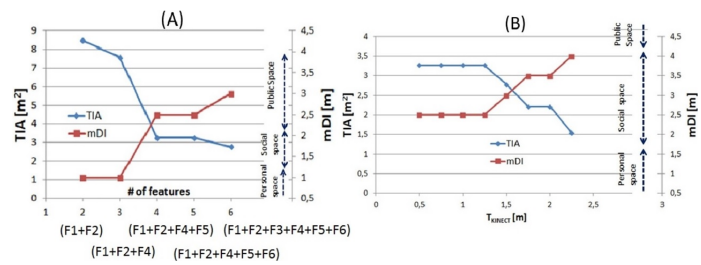



Figure 11. Total interaction area (TIA), and minimal distance of interaction (mDI) are reported for several conditions: (A) Shows the effect of selecting a subset of available features. The parameters F1-F6 refer to the name of features reported in Figure 10. (B) reports the effect of changing Kinect height. Boundaries of Proxemics spaces are shown for both the graphs

and head-eye movements. Its control system was running on Windows platform; ZENO robot is a humanoid robot with full head and body moving capabilities, which runs Ubuntu Linux on board, with an Intel Atom processor and 1 Gb RAM [51]. In latter case, due to the robot's hardware specification, processing was done on another machine and all commands were sent to the robot via the API. In both cases, the SPS successfully created and delivered meta-scene irrespective of the robots' OS, and allowed acceptable HRIs.

In the first scenario, the developer used meta-scene provided by SPS to design and control an "'I-CLIPS Brain: A Hybrid Cognitive System for Social Robots"' [31], [52], and thus the SPS has been employed as the perception unit of the robot's brain. In the second scenario, SPS was used as the perception unit to enable the robot to have a set of context-aware behaviors. As the result, ZENO behaved differently when it perceived positive and negative emotions of users. The following section summarizes some practical note about integrating SPS as perception unit of real robots used for real-time HRI as well as some issues, which can affect the performance of SPS.

- *Platform independence:* SPS composed of four layers that can be run in either one machine or two different machines. In the latter case, the perception machine that provides meta-scene has to be run on Windows platform

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



while the robot controller machine, which receives meta-scene can be run in different operation systems such as Linux, Mac, etc.

- *The performance of perception machine:* is the most important factor, which affects the SPS performance. Running the perception system on a high-performance machine results in a fluent perception and data communication. However, with nowadays machines we do not expect any incompatibility.
- *The overall computational time:* SPS aims at enabling real-time HRI and thus, its computational time in delivering meta-scene is important. As shown, SPS composed of four layers, in which feature extraction layer has six parallel perceptual modules: body analysis, facial feature analysis, salience detection, sound analysis, identity assignment, and environment analysis. The body analysis module and the meta-scene layer are always needed to be activated in SPS. Instead, the other two modules (i.e., facial analysis and salience detection) can be deactivated according to scenario requirements. The overall computational time depends on the number of running modules. As we tested, body analysis module that process the data stream of Kinect, works at 33 ms (30 fps) of computational time. It is not affected by the number of people in the Kinect's field of view. Facial feature analysis module can analyze a whole image in less than 22 ms although, for technical and methodological reasons, we decided to slow down the reading process through YARP to 200 ms. Salience detection module can deliver data in less than 100 ms according to software setting although, for the same methodological and technological reasons, we read this data every 200 ms. Since the aforementioned modules have different computational time, employing all the modules together requires the longest computational time that is 200 ms. In summary, SPS can deliver a meta-scene with full list of features in every 200 ms.
- *Sending and receiving threads:* in addition to the machine performance and SPS's inherent computational time on delivering data, due to the processing time, SPS has two internal adjustable timers for sending and receiving meta-scene.
- *The quality of network connection:* since the SPS deliver meta-scene wirelessly through YARP to a specific Internet Protocol address, the quality of network connection is an important factor for data communication performance.
- *Visualization:* SPS can be run with the user interface that shows on display all the extracted features which are important for system initialization and calibration. Delivering meta-scene in the visualization mode is a little bit slower than running it without visualization however, the delay is negligible.

The last important feature of SPS is the capability of doing partial perception. For example, if an application requires only the face position of users, we can easily adjust SPS in partial detection mode that delivers meta-scene with the same structure while only it assigns the values to the users face positions.

Figure 12. Two humanoid robots FACE and ZENO.

## VIII. Conclusion

We performed an analysis of the state-of-the art technology of perception systems, which often are used in the context of social robotics. We noticed that many of these perception systems are not able to gather simultaneously a wide range of both human and environment relevant features guiding a human-robot social interaction, and all of these systems have been built for specific robotics platforms making hard to be used in different robots. Moreover, usually the data delivered by these perception systems are a mere list of features and they are not fully exploited for a real interpretation the surrounding, moving from data to meta-data. To overcome such limitations, we proposed and tested a novel social perception system that has been designed to meet the important technical requirements of a standard perception system. The proposed system has the features that have been fully described in Section I: modularity, interconnectivity, extendability, communication capability.

We performed an extensive analysis and experiments on human-relevant features that SPS can detect from environment. Such results allow to detect the interaction space that is subsection of the visual sensor FOV, where humans can stay to fully exploit the perception capability of SPS. We tested SPS communication capability by integrating it into two different humanoid robots i.e., FACE and ZENO. As demonstrated, SPS was able to detect and deliver to different robots, a wide range of human-relevant social features, which shows the promising capability of the system to be used as a unique social perception system for social HRI.

## Acknowledgments

This work was partially funded by the European Commission under the 7th Framework Program projects EASEL, "'Expressive Agents for Symbiotic Education and Learning'", grant agreement No. 611971 - FP7-ICT-2013-10.

## References

[1] M. Beetz, B. Johnston, and M.-A. Williams, *Social Robotics*. Springer, 2014.

[2] C. Breazeal, "Toward sociable robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 167–175, 2003.

[3] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 3, pp. 183–199, 2014.

[4] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "Kaspar–a minimally expressive humanoid robot for human–robot interaction research," *Applied Bionics and Biomechanics*, vol. 6, no. 3-4, pp. 369–397, 2009.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems




[5] A. Causo, G. T. Vo, I.-M. Chen, and S. H. Yeo, "Design of robots used as education companion and tutor," in *Robotics and Mechatronics*. Springer, 2016, pp. 75–84.

[6] H.-M. Gross, S. Mueller, C. Schroeter, M. Volkhardt, A. Scheidig, K. Debes, K. Richter, and N. Doering, "Robot companion for domestic health assistance: Implementation, test and case study under everyday conditions in private apartments," in *Inte. Conf. on Intel. Robots and Sys.(IROS), 2015*. IEEE, 2015, pp. 5992–5999.

[7] H. Wechsler, *Neural Networks for Perception: Human and machine perception*. Academic Press, 2014.

[8] H. Yan, M. H. Ang Jr, and A. N. Poo, "A survey on perception methods for human-robot interaction in social robots," *International Journal of Social Robotics*, vol. 6, no. 1, pp. 85–119, 2014.

[9] M. Hirose and K. Ogawa, "Honda humanoid robots development," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1850, pp. 11–19, 2007.

[10] G. Sandini, G. Metta, and D. Vernon, *The icub cognitive humanoid robot: An open-system research platform for enactive cognition*. Springer, 2007.

[11] K. Okada, T. Ogura, A. Haneda, D. Kousaka, H. Nakai, M. Inaba, and H. Inoue, "Integrated system software for hrp2 humanoid," in *Robotics and Automation, 2004. Proceedings. ICRA'04. 2004 IEEE International Conference on*, vol. 4. IEEE, 2004, pp. 3207–3212.

[12] C. Zhou, S. Lin, and S. K. Nayar, "Coded aperture pairs for depth from defocus and defocus deblurring," *International journal of computer vision*, vol. 93, no. 1, pp. 53–72, 2011.

[13] K.-D. Kuhnert and M. Stommel, "Fusion of stereo-camera and pmd-camera data for real-time suited precise 3d environment reconstruction," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2006, pp. 4780–4785.

[14] G. Grunwald, G. Schreiber, A. Albu-Schäffer, and G. Hirzinger, "Touch: The intuitive type of human and robot interaction," in *Advances in Human-Robot Interaction*. Springer, 2005, pp. 9–21.

[15] M. Argyle, *Bodily communication*. Routledge, 2013.

[16] P. Ekman, "Facial expressions," *Handbook of cognition and emotion*, vol. 53, pp. 226–232, 1999.

[17] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[18] G. Yang and T. S. Huang, "Human face detection in a complex background," *Pattern recognition*, vol. 27, no. 1, pp. 53–63, 1994.

[19] I. Craw, D. Tock, and A. Bennett, "Finding face features," in *Computer VisionECCV'92*. Springer, 1992, pp. 92–96.

[20] C. Ding and D. Tao, "A comprehensive survey on pose-invariant face recognition," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, p. 37, 2016.

[21] M. P. Michalowski, S. Šabanović, C. DiSalvo, D. Busquets, L. M. Hiatt, N. A. Melchior, and R. Simmons, "Socially distributed perception: Grace plays social tag at aaai 2005," *Autonomous Robots*, vol. 22, no. 4, pp. 385–397, 2007.

[22] T. Germa, F. Lerasle, P. Danes, and L. Brethes, "Human/robot visual interaction for a tour-guide robot," in *Intel. Robots and Sys. IROS 2007. IEEE/RSJ International Conference on*, 2007, pp. 3448–3453.

[23] T. Spexard, A. Haasch, J. Fritsch, and G. Sagerer, "Human-like person tracking with an anthropomorphic robot," in *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*. IEEE, 2006, pp. 1286–1292.

[24] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer, "Multi-modal anchoring for human–robot interaction," *Robotics and Autonomous Systems*, vol. 43, no. 2, pp. 133–147, 2003.

[25] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[26] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human–robot interaction," *Image and Vision Computing*, vol. 25, no. 12, pp. 1875–1884, 2007.

[27] 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 *Second RSI/ISM Inte. Conf. on Robotics and Mechatronics (ICRoM), 2014*. IEEE, 2014, pp. 185–190.

[28] E. T. Hall, *The hidden dimension*. Anchor Books New York, 1969, vol. 1990.

[29] R. Mead and M. J. Matarić, "Perceptual models of human-robot proxemics," in *Experimental Robotics*. Springer, 2016, pp. 261–276.

[30] P. Holthaus, K. Pitsch, and S. Wachsmuth, "How can i help?" *International Journal of Social Robotics*, vol. 3, no. 4, pp. 383–393, 2011.

[31] D. Mazzei, L. Cominelli, N. Lazzeri, A. Zaraki, and D. De Rossi, "I-clips brain: A hybrid cognitive system for social robots," in *Biomimetic and Biohybrid Systems*. Springer, 2014, pp. 213–224.

[32] A. Betkowska, K. Shinoda, and S. Furui, "Robust speech recognition using factorial hmms for home environments," *Eurasip Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 10–10, 2007.

[33] L. Itti. (2015, Feb.) Visual salience. [Online]. Available: http://www.scholarpedia.org/article/Visual_salience

[34] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[35] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[36] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of vision*, vol. 8, no. 7, p. 32, 2008.

[37] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, "Visual saliency model for robot cameras," in *Inte. Conf. on Robotics and Automation, 2008*. IEEE, 2008, pp. 2398–2403.

[38] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: yet another robot platform," *International Journal on Advanced Robotics Systems*, vol. 3, no. 1, pp. 43–48, 2006.

[39] D. Mazzei. (2015, Feb.) Environment sensing. [Online]. Available: http://thingsoninternet.biz/products/toi-shield/

[40] Microsoft. (2016, July) Kinect features. [Online]. Available: http://www.microsoft.com/en-us/kinectforwindows/meetkinect/features.aspx

[41] F. IIS. (2016, July) Cognitive systems. [Online]. Available: http://www.iis.fraunhofer.de/en/ff/bsy/tech/bildanalyse.html

[42] C. Küblbeck and A. Ernst, "Face detection and tracking in video sequences using the modified census transformation," *Image and Vision Computing*, vol. 24, no. 6, pp. 564–572, 2006.

[43] M. Ogawa, T. Yonezawa, Y. Nishiyama, J. Nakazawa, and H. Tokuda, "A robot control system for video streaming services by using dynamic encoded qr-codes," in *Mobile Computing and Ubiquitous Networking (ICMU), 2015 Eighth Inte. Conf. on*. IEEE, 2015, pp. 86–87.

[44] J. Panero and M. Zelnik, *Human dimension and interior space: a source book of design reference standards*. Watson-Guptill, 2014.

[45] J. Fasola and M. Mataric, "A socially assistive robot exercise coach for the elderly," *Journal of HRI*, vol. 2, no. 2, pp. 3–32, 2013.

[46] K. Khoshelham, "Accuracy analysis of kinect depth data," in *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 38, no. 5, 2011, pp. 133–138.

[47] M. A. Livingston, J. Sebastian, Z. Ai, and J. W. Decker, "Performance measurements for the microsoft kinect skeleton," in *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*. IEEE, 2012, pp. 119–120.

[48] R. J. Wierzbicki, C. Tschoeppe, T. Ruf, and J.-U. Garbas, "Edis-emotion-driven interactive systems," *Other Publications of the AMEA Association*, no. 1, 2013.

[49] D. Hanson, A. Olney, S. Prilliman, E. Mathews, M. Zielke, D. Hammons, R. Fernandez, and H. Stephanou, "Upending the uncanny valley," in *Proceedings of the national conference on artificial intelligence*, vol. 20, no. 4. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005, p. 1728.

[50] D. Hanson, G. Pioggia, Y. Bar-Cohen, and D. De Rossi, "Androids: application of eap as artificial muscles to entertainment industry," in *SPIE's 8th Annual International Symposium on Smart Structures and Materials*. International Society for Optics and Photonics, 2001, pp. 375–379.

[51] D. Cameron, S. Fernando, E. Collins, A. Millings, R. Moore, A. Sharkey, V. Evers, and T. Prescott, "Presence of life-like robot expressions influences childrens enjoyment of human-robot interactions in the field," in *Proceedings of the AISB (Artificial Intelligence and Simulation of Behaviour) Symposium*, 2015.

[52] A. Zaraki, D. Mazzei, M. Giuliani, and D. De Rossi, "Designing and evaluating a social gaze-control system for a humanoid robot," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 157–168, February 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2016.2598423, IEEE Transactions on Cognitive and Developmental Systems



**Abolfazl Zaraki** graduated with a PhD degree in Automatic Robotic and Bioengineering from University of Pisa, in 2014. From Feb. 2014 to July 2016 was working as a Post-Doc Research Fellow at Research Center E. Piaggio of University of Pisa. Since July 2016, he is working as a senior research associate at the school of computer science, University of Hertfordshire, UK. He graduated with a B.S. Degree in Electronics-Electrical engineering from the University of Dezful, Iran, in 2006. He received his Master's Degree in Engineering in Mechatronics and Automatic Control from University Technology of Malaysia, in 2010. His current work is to define, implement and evaluate child-robot interaction application scenarios for developing specific socio-affective, communication and collaboration skills in autistic spectrum children.

**Michael Pieroni** PhD student in Automation, Robotics and Bioengineering at Research Center E. Piaggio, University of Pisa. At the same university, he received the M.S. Degree in Biomedical Engineering with honors in 2012. In 2014 he was Visiting Student Researcher at Stanford University (CA, USA). He published papers in field of bionics, biomimetics, artificial intelligence and Internet of Things. His research activity concerns the modelling and developing technology inspired by biological vision system.

**Danilo De Rossi** was awarded the title of Doctor in Chemical Engineering at the University of Genoa in 1976. From 1976 to 1981 he was researcher of the Institute of Clinical Physiology of C.N.R. He had appointments for teaching and research in France, USA, Brazil, Japan and Australia. Since 1982 he has been working in the School of Engineering of the University of Pisa, where he is presently Full Professor of Bioengineering. His scientific activities are related to the design of sensors and actuators for bioengineering and robotics and to the study and development of wearable systems for telemonitoring and telerehabilitation. He received the Young Investigator Forum Award from Biomedical Engineering Society (UK) in 1980 and from the American Society for Artificial Internal Organs in 1985. In 2012 he was awarded of the Order of the Cherubino from the University of Pisa for institutional and scientific values. He is author of over 300 peer reviewed papers on international science journals and peer reviewed proceedings, co-inventor of 14 patents and co-author of 10 books.

**Daniele Mazzei** PhD in Automatic Robotic and Bioengineering. Post-Doc researcher at Research Center E. Piaggio of University of Pisa. He is the scientific coordinator of the FACETeam, a group of the E.Piaggio focused on development of social robots and cognitive systems. He co-founded in 2014 a start-up called TOI ThingsOnInternet.biz. TOIs mission is to give everyone the tools to bring their Things Onto the Internet creating a new generation of social interactive objects and installations.

**Roberto Garofalo** Technical and developer with the FACETeam at Research Center E. Piaggio, University of Pisa. He received a Bachelors Degree in Electronic Engineering at the University of Pisa in June 2014 with a thesis aimed at developing of a modular platform for controlling the humanoid robot F.A.C.E. (Facial Automaton for Conveying Emotion) during human-robot social interaction. His main activities are focused on robotic developing for both hardware and software, in particular Robert designs and integrates robot body parts,and implement software architecture for its control.

**Lorenzo Cominelli** PhD Student, Biomedical Engineer and System Developer for Artificial Intelligence and Human Behavior Understanding. Currently working in the Face Team, he is developing the cognitive system of F.A.C.E. (Facial Automaton for Conveying Emotion) at Research Center E. Piaggio. The aim of his research is to create and exploit artificial emotions in social robots, grounded on human emotional and decision processes, psychology and philosophy.

**Maryam Banitalebi Dehkordi** has received her PhD in Perceptual Robotics and Innovative Technologies, in 2013, from Scuola Superiore Sant'Anna, Pisa, Italy, where at the same institute she has worked as a Post-Doc Research Fellow. She received her B.S. degree in Electronics-Electrical Engineering from Shahrekord University, Iran, in 2006. She received the Master of Engineering in Mechatronics and Automatic Control from University Technology of Malaysia, in 2009. She has been a visiting researcher at Technical University of Munich, Germany, in 2013. She has followed her research at Research Center E. Piaggio of University of Pisa, in 2015, on robot non-verbal communication, in which she focused on developing gestures for a humanlike robot in order to enable the robot to express emotions through body gesture in a natural and humanlike manner. Her research interests include assistive technologies using robots, human-robot interaction, and robot control.