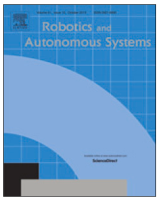
Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot



ChildBot: Multi-robot perception and interaction with children[☆]

Niki Efthymiou [a,b,*], Panagiotis P. Filntisis [a,b], Petros Koutras [a,b,1], Antigoni Tsiami [a,b], Jack Hadfield [a,b], Gerasimos Potamianos [a,c], Petros Maragos [a,b]

[a] Athena Research and Innovation Center, Maroussi 15125, Greece
[b] School of ECE, National Technical University of Athens, Athens 15780, Greece
[c] Department of ECE, University of Thessaly, Volos 38221, Greece

ARTICLE INFO

Article history:
Received 6 October 2020
Received in revised form 7 September 2021
Accepted 10 December 2021
Available online 24 December 2021

Keywords:
Child–Robot Interaction
Multi-robot perception
Visual activity recognition
Distant speech recognition
Audio-visual active speaker localization
6-DoF object tracking

ABSTRACT

In this paper, we present an integrated robotic system capable of participating in and performing a wide range of educational and entertainment tasks collaborating with one or more children. The system, called ChildBot, features multimodal perception modules and multiple robotic agents that monitor the interaction environment and can robustly coordinate complex Child–Robot Interaction use-cases. In order to validate the effectiveness of the system and its integrated modules, we have conducted multiple experiments with a total of 52 children. Our results show improved perception capabilities in comparison to our earlier works that ChildBot was based on. In addition, we have conducted a preliminary user experience study, employing some educational/entertainment tasks, that yields encouraging results regarding the technical validity of our system and initial insights on the user experience with it.

1. Introduction

Recently, robotic systems designed for Human–Robot Interaction (HRI) have been attracting significant interest, finding applications in various aspects of everyday life [1], such as entertainment [2], education [3,4], nursing and personal care [5], rehabilitation [6,7], and autism therapy [8,9], among others. Crucial to their usability and wider adoption is the achievement of resembling human-to-human interaction. For this purpose, robots need to have the ability to perceive and understand the different modalities that people use for communication, such as speech or body movements [10,11].

To date, most social robotics systems present two major deficiencies: First, they typically incorporate only specific modalities, forcing their users to adapt to the way the system perceives the environment instead of the opposite. Second, they are developed and designed for specific applications and tasks. With their expanding use in various application areas though, the need arises for integrated systems capable of dealing with multiple applications and real-world scenarios in challenging environments, offering natural interaction to their users. Such interaction involves creating smart adaptive integrated robotic systems capable of multitasking with a wide range of perceptual and actuation abilities, allowing HRI system users to design multiform interactive applications that can maintain their interest and engagement. This is especially important for children users in the context of education and entertainment (edutainment). In addition, systems leveraging multiple perceptual modalities allow their users to conduct HRI in their preferred communication modality.

One of the challenges in achieving the above is the fact that commercial social robots have different capabilities. For example, the NAO robot [12] is capable and adept in body movements, but incapable of facial expressions, Furhat [13] presents a large variety of facial expressions but cannot move, while the Zeno robot [14] is capable of movements and facial expressions, but not adept in both (Zeno's body movement lacks in comparison to NAO). Additionally, each social robot has different sensors, constraining the user to specific communication channels.

Motivated by the above, in this work, we present an integrated robotic system that can be used for multiple edutainment applications, which we will refer to as "ChildBot". To achieve this versatility, ChildBot incorporates: (i) multiple sensors and perception modules that allow the user to communicate with the robots via multiple channels, and (ii) multiple social robots, leveraging each other's strengths to circumvent their individual weaknesses.

An overview of the proposed system can be seen in Fig. 1. ChildBot is developed using a *Sense–Think–Act* paradigm [15] and

[☆] This research work was supported by the EU Horizon 2020 project BabyRobot, under grant 687831.

[*] Corresponding author at: School of ECE, National Technical University of Athens, Athens 15780, Greece.
E-mail address: nefthymiou@central.ntua.gr (N. Efthymiou).

[1] Work conducted while Petros Koutras was at NTUA and Athena RC, now he is with Snap Inc.

https://doi.org/10.1016/j.robot.2021.103975
0921-8890/© 2021 Elsevier B.V. All rights reserved.

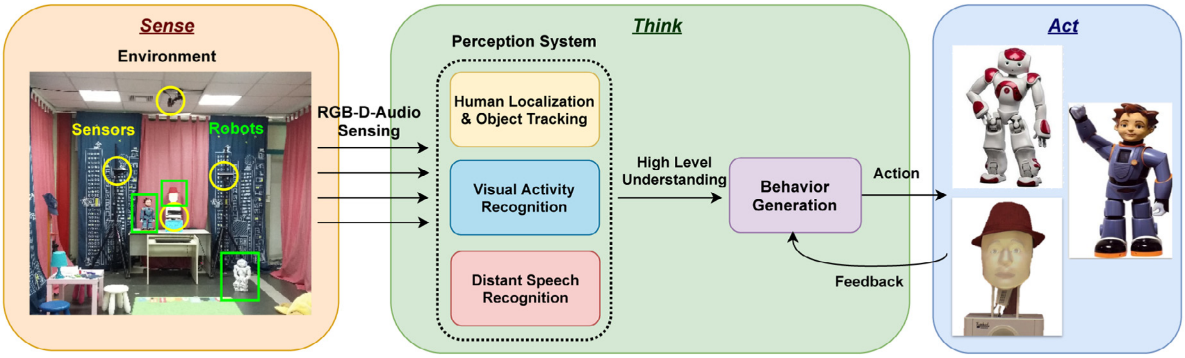


Fig. 1. Schematic overview of the ChildBot system during Child–Robot Interaction. The multimodal information of the child's action is received through a network of sensors placed around an interaction area. The perception system processes it and extracts high-level information about the action context. Based on this, the behavior generation module decides and controls the robotic agents.

is indoors based, allowing us to employ external sensors arranged inside a "smart space" where the interaction takes place. Robot-external sensing can overcome common HRI problems such as occlusions and allows the fusion of different data streams, improving the robustness and performance of the various perception modules. This way, we also achieve automatic perception of the interaction in a *robot-independent* fashion, bypassing the sensing limitations of individual robotic systems. In addition, this robot-agnostic architecture can easily accommodate new robots. The system coordinates a complex and continuous HRI procedure involving child actions and robot responses. Specifically, multimodal information flows from the sensors (Sense) to the perception modules. Then, high-level information about action context is extracted, and the appropriate response/action of the robot is decided (Think). Finally, the system transmits this decision to the robotic agent to act appropriately (Act).

The contribution of the proposed system lies both on the variety of its modules and the proposed integration of the system. Integration of perception modules is vital for accommodating more complex Child–Robot Interaction (CRI) scenarios that require multimodal HRI. This modularity allows the selection of different perception modules according to the desired application without affecting the functionality of the entire system. Furthermore, integration of multiple robots in a robot independent fashion allows switching between different robots according to the users' desires, as well as the addition of new social robots as they are released.

To showcase the versatility and capabilities of the integrated system and its individual perception modules, we have designed five different edutainment use-cases. These use-cases are *indicative* and have been designed to exploit different system components, showcasing the large variety of applications that can be accommodated with ChildBot. The data collected by a pool of 52 children while performing the aforementioned use-cases with the robots allow us to objectively evaluate the performance of each module of the ChildBot system regarding its perception capabilities. Furthermore, our crude evaluation of user experience yields encouraging results towards a complete well-designed subjective evaluation in the future.

ChildBot is an improved integrated extension of a set of preliminary works reported by the authors in conference publications on specific problems of multi-robot perception and interaction. It presents a wide-application CRI system able to manage multitasking interaction autonomously and accommodate a plethora of edutainment scenarios [16–19]. The work presented here has integrated our earlier research under a single and modular three-layer multi-robot architecture [15], includes improved perception modules, and is evaluated extensively on a larger corpus that contains spontaneous children data, more representative of CRI. To summarize, we highlight the most important contributions of the presented work:

- *An integrated system for HRI* has been designed and implemented by leveraging multiple robotic agents. The modular three-layer system architecture integrates multiple sensors, numerous perception modules, and different robotic agents, culminating in a multi-application autonomous HRI system.
- *Perception modules for multimodal scene understanding* have been developed and adjusted to specific CRI conditions by incorporating novel approaches and extensive studies. Audio-visual active speaker localization, six degrees-of-freedom (6-DoF) object tracking, visual activity recognition, and distant speech recognition are necessary for analyzing and tracking human behavior over time in the context of their surroundings. The perception modules of this system have been developed according to and sometimes exceed the state-of-the-art of the underlying technologies, as shown by our objective evaluations.
- *Spontaneous children data during CRI* have been collected and used for system evaluation. Indicative use-cases have been defined and implemented in order to showcase the large range of applications that ChildBot can be used for. The collected data have allowed for an extensive objective evaluation of the ChildBot capabilities during real use-case scenarios, as well as a preliminary user experience study.

2. Related work

Many research projects have aimed at developing robots both in ambient assisted living environments [20], as well as in well-defined and constrained environments, e.g., in bathrooms for assistive bathing [21]. Some robotic agents act as companions to improve quality of life, assist with mobility, or complete household tasks [22,23], while others are designed to help people live independently and serve themselves when they face difficulties due to disabilities or old age [24,25]. Nevertheless, the intervention of robots in human life remains a controversial issue [26–29].

Regarding educational CRI applications, many previous works focus on the theoretical exploration of different social robot behaviors in the learning experience, without delving deeply into the technical aspects, but mainly using off-the-shelf solutions for environment perception. An immediate result of this is the fact that the interaction space is constrained. In [30], a study involved children playing an educative mathematics scenario with a NAO robot. In [31], Saerbeck et al. studied the effect of social robot behavior on the subjects' learning performance in the context of a language learning task. Similar studies can be found in [32], while in [33] a humanoid robot was employed to interact with autistic children.

Notable works that have also focused on the robot perception aspect of CRI include the ALIZ-E project [34]. Belpaeme et al. studied a long-term, adaptive social CRI, emphasized the difficulties faced in real-world experiments at a school and a hospital, and developed a complete framework for multimodal CRI. Another interesting work evaluated in a hospital environment is the NAOTherapist platform [35], which focused on upper-limb rehabilitation sessions for children with physical impairments. The NAO robot performed physiotherapy sessions autonomously, observing the patient's pose through a Kinect sensor and giving appropriate corrective instructions if needed. A similar platform for creating autonomous interaction between a robot and Autism Spectrum Disorder (ASD) children was built by the INSIDE project [36]. A multimodal perception system was developed to identify the child's position and activity, as well as satisfaction when replying by the robot, and the state of the current activity.

Other similar projects include L2TOR [37], where a NAO robot capable of multimodal perception assumed the role of a second-language tutor, and the EASEL educational CRI project, where Vouloutsi et al. [38] presented a distributed adaptive control architecture of the robotic system developed and its four layers: somatic, reactive, adaptive, and contextual. Esteban et al. [39] built a multi-sensor system for autonomous interaction of a NAO robot with autistic children to perceive different features during an interaction, such as gaze estimation, action recognition, and object tracking. The system capabilities were sufficient for the presented tasks but limited for a more generic interaction, and the system lacked real use-case evaluation. In [40], Marinoiu et al. introduced an action and emotion recognition system by exploiting 2D and 3D pose estimation methods and evaluated it on a large-scale dataset of robot-assisted therapy sessions of children with autism. The ANIMATAS project, focusing on training researchers to advance human–machine interaction, is also worth mentioning. Specifically, in [41], Valipour et al. underlined the differences in social robot mind perception during virtual and real-world experiments and the importance of conducting real experiments to reveal all interaction parameters.

A general review of the perception methods used for HRI in social robots until 2014 is presented in [42]. Three important issues associated with perception systems are highlighted there: the need for developing perception systems in real environments with real data, the requirement of creating good representations following the context of the interaction, and the demand of combining an efficient perception system with reasonable robot responses in order to create pleasant HRI experiences. A recent review about how social robots perceive humans and their interactions is presented by Tapus et al. [43] and illustrates how social cues occur during HRI. Zaraki et al. [44] attempted to develop perception systems for HRI by combining low-level and high-level features to detect a range of human-relevant features that appear during a real use-case procedure. Valipour et al. [45] proposed a novel paradigm for incrementally improving the visual perception of a robot during an HRI experience.

Focusing on perception technologies for children, Kennedy et al. [46], after evaluating numerous automatic speech recognition systems, concluded that child speech recognition requires a multi-pronged approach to be efficient and achieve higher performance. An interesting result was noted by Yeung and Alwan [47], where it was found that even a single year difference in the kindergarten age impacts the performance of automatic speech recognition. Regarding action recognition, Chiang et al. [48] classified eight human actions performed by children and adults using Histograms of Oriented Gradients (HOG) features extracted from combined depth and motion color maps. Moreover, in [49], Zhang et al. developed a method that recognizes stereotyped actions of ASD children using Long Short-Term Memory networks on top of skeleton data. Finally, in [50], Wu et al. integrated object recognition in an educative robotic system that provided interesting and innovative second language learning services for preschool children in China.

Aiming to increase performance, flexibility, and robustness, ChildBot consists of multiple robots and multimodal perception modules designed for and adapted to children, allowing interaction inside a relatively large space for a variety of edutainment tasks. Parts of the ChildBot system are based on our previous preliminary works, where an early design was presented. More specifically, a preliminary setup and evaluation of a basic architecture in a few use-cases has been presented in [19], while [18] has focused mainly on multi-party interaction via speech. In [16], the techniques for a multi-view fusion of action recognition have been explored more in-depth. Finally, [17] has focused on developing tracking algorithms, essential in interactive tasks between a child and a robot. In all these previous works, limited and specific functionality of the system has been investigated, and evaluation has been carried out employing data acquired in a strictly controlled procedure and not spontaneous data from real interaction.

The current paper integrates all these different modules from previous works under the same unified system, using a three-layer architecture. The modules can now work both in isolation and in synergy, and due to the system modularity, adding or removing a component is an easy task. Moreover, much effort has been dedicated to improving the perception modules' performance apart from the integration. For this reason, we have included more sensors and carried out ablation studies in order to validate the plausibility of the employed modules. Most importantly, contrary to all previous works, the modules evaluation has been performed employing real-time spontaneous children data (see Section 4.2), which are more challenging.

3. System overview

This section presents the overall ChildBot perception system that provides global and effective CRI supervision. We first briefly provide an overview of the perception system and its modules. Then we delve into each perception module and its algorithms, describing them in depth. Finally, we detail the system architecture that constitutes the backbone of the perception system.

3.1. Perception system

The overview of the robot-agnostic perception system can be seen in Fig. 2, consisting of three main modules: *Audio-Visual Active Speaker Localization and 6-DoF Object Tracking*, *Visual Activity Recognition*, and *Distant Speech Recognition*. Four Kinect V2 sensors capture a detailed raw data representation of the environment and feed it into the perception system. The Kinect V2 sensors are placed at different positions and viewing angles to sufficiently cover the entire environment, tackle occlusion problems (self or from objects), and offer multiple viewpoints for visual perception. The sensors record RGB, depth, and 4-channel audio – the latter from the microphone array of each Kinect. The spatial arrangement of the sensors is presented in Fig. 3(a). Subsequently, we present an overview of each perception module:

Audio-Visual Active Speaker Localization and 6-DoF Object Tracking: To allow a natural interaction between robots and humans, robotic awareness of the active speaker location, as well as the detection and tracking of important objects, are essential.

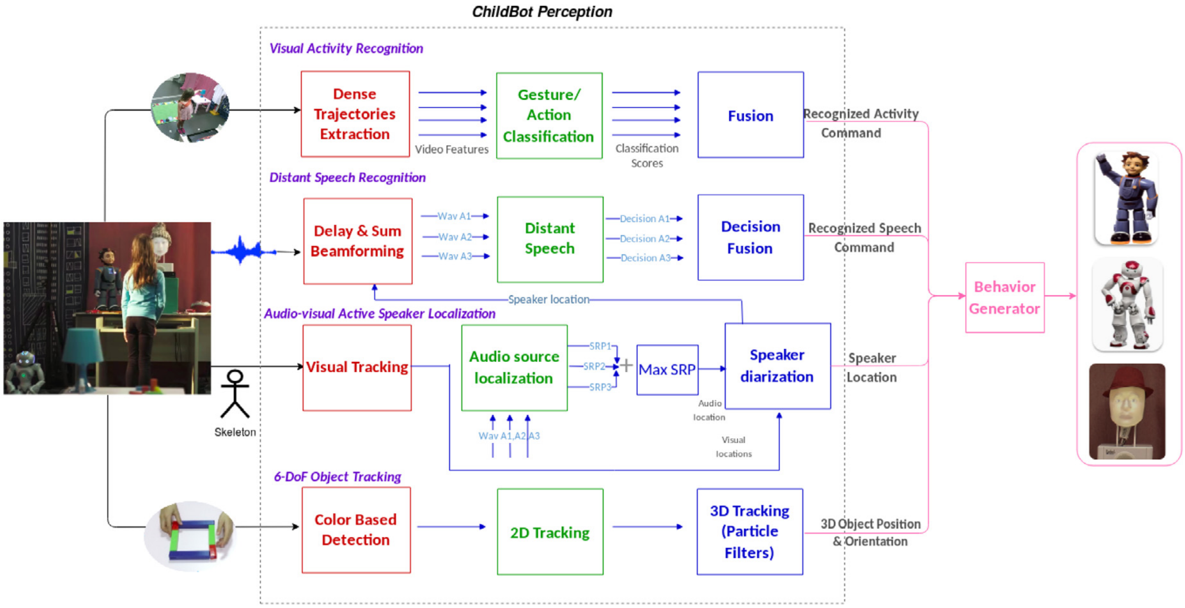


Fig. 2. Overview of the ChildBot perception modules including: a) *Audio-Visual Active Speaker Localization and 6-DoF Object Tracking*, b) *Visual Activity Recognition*, and c) *Distant Speech Recognition*. "A" refers to microphone array and "SRP" to Steered Response Power. The modules are employed during CRI to monitor the multiple aspects of human behavior, and then their outputs are fed to the robotic behavior generator module.

An effective audio-visual method for active speaker localization in HRI scenes has been developed by leveraging audio information in addition to visual information. Moreover, a module for object recognition that can detect multiple toys based on their colors and size has been incorporated into the system. Finally, a 3D tracking method has been designed for providing both the 3D location and the orientation of rigid objects.

Visual Activity Recognition: A visual frontend has been developed to recognize hand gestures accompanying everyday communication and more general body movements conveying specific meanings. The multiview visual activity recognition module can successfully recognize the child's activity while wandering around the room and interacting with the robots and objects. The gesture recognition version of the module aims at identifying hand gestures that deliver a conceptual message during the interaction, such as waving at the robot or asking the robot to come closer. On the other hand, the action recognition version targets child body movements that form complex meanings, such as pantomimic movements.

Distant Speech Recognition: A multisensory distant speech recognition (DSR) system in Greek has been developed to enable CRI via speech. As close-talking microphones are not convenient for children and restrict their movements, we take advantage of the multiple microphone arrays located around the room recording audio, while at the same time the children can move freely and communicate hands-free with the robots. In order to make the DSR more robust and exploit the distributed microphone arrays, we experiment with adaptation and fusion.

The high-level understanding obtained by the perception modules is fed to the Dialog Manager, along with extra input from a Touch Screen, which is used as an extra means of communication during the interaction. According to its input, the Behavioral Generator then decides on the multi-robot system action and forwards its decision to the actuators. The actuators, in turn, respond with information back to the system.

In order to create a detailed picture of the ChildBot system, we proceed by describing each perception module extensively. Following this, we present the system architecture, the intercommunications, and the dialog management module in order to describe our complete system for CRI.

3.2. Perception modules

Audio-visual active speaker localization

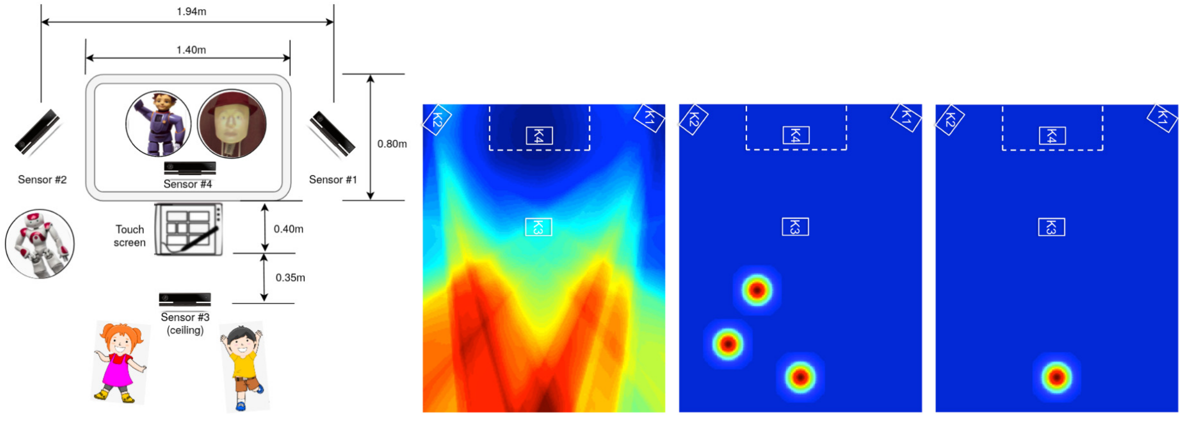
When analyzing and understanding an auditory or audio-visual scene that consists of multiple speakers, the sound and speaker localization are necessary for tracking. In addition, the speaker's location is required for beamforming and guiding the robot's attention/head in a multi-party scenario to achieve natural and intuitive interaction. Although visual tracking can be precise, it does not suffice when the active speaker/speakers have to be localized among other non-speaking persons in an audio-visual scene.

Various techniques for audio speaker localization [51] have been proposed in the literature. Some of them have been specifically adapted to HRI setups [52,53] for microphones mounted on robots. In our multi-robot case, microphones are external to the robots, and a fast algorithm is needed due to the real-time nature of our system. Thus, a real-time 3D audio localization SRP-PHAT (Steered Response Power-Phase Transform) system based on [54,55] has been developed, which is robust to noise and errors. Regarding audio-visual speaker localization [56–58], several methods have been developed for RGB cameras, most of them employing Bayesian filtering techniques or fusion between audio and video features. In our case, visual tracking is accomplished via skeleton tracking, developed for Kinect sensors.

Our audio-visual active speaker localization exploits the 3D skeleton (provided by Kinect V2) and the microphone arrays, and it is performed as follows: Person tracking is first achieved by retrieving the 3D skeletons of all persons present in the audio-visual scene. Auditory source localization via SRP-PHAT provides information concerning the speakers. The final active speaker localization is performed by choosing the visual locations closest to the auditory ones. The result is then used to guide the robot's attention by turning its head towards the active speaker. An example of audio-visual active speaker localization can be seen in Figs. 3(b)–3(d).

6-DoF Object Tracking

In certain CRI scenarios, children and robots may be expected to interact with various movable objects. Thus, aside from human localization described above, the robot must understand the



(a) Experimental setup

(b) Audio source localization

(c) Visual person localization

(d) Audio-visual active speaker localization

Fig. 3. (a) Spatial arrangement of the four Kinect sensors. (b–d) An example of audio-visual active speaker localization. The SRP output is shown with high values in red. Positions of the table and the four Kinects are also shown. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

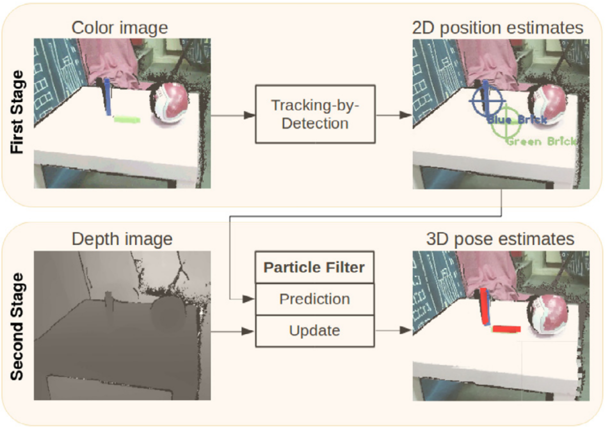


Fig. 4. Overview of the implemented 6-DoF object tracking module. In this example, the bricks are tracked. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

configuration of these objects. We have developed a method for robustly tracking the 6-DoF poses of multiple objects in real-time. The main idea is to crudely detect the objects computationally cheaply and then use the detected positions to infer each object's 3D pose. The used objects are known beforehand, meaning that their shape and appearance models are predefined. The developed tracker consists of two stages: the first involves a tracking-by-detection scheme upon the color stream to locate the objects on the image plane. The second operates on depth data to refine the first stage output and infer the remaining variables related to the object rotations. The basic architecture is presented in Fig. 4.

During the first stage, our approach uses a simple color histogram model to detect object regions, though depending on the object characteristics a variety of features could potentially be used. The histogram models are defined offline and remain unchanged during the entire tracking. The hue and saturation of the Hue Saturation Value (HSV) color space were used to introduce sufficient robustness to brightness changes. Assuming the histograms are normalized, they define a probability distribution over the color space. Therefore, a probability map can be generated over the latest color image. After thresholding and morphological filtering, a binary mask contains the most likely object regions in the image. We choose to retain the region with the largest area under the assumption that the remaining regions will correspond to noisy artifacts or irrelevant background objects. The center of the chosen region is taken as the object location, and a confidence score s_k is produced for object k.

Once the object locations have been detected on the image plane, the tracker's second stage consists of estimating the 6-DoF poses with the help of the newest depth image. The developed tracker employs particle filters and is closely based on the algorithm proposed in [59], where the hidden states are augmented with a set of binary variables that model the occlusions at each pixel. We transform the kth object's 2D position estimates $\mathbf{p}_k$ into 3D estimates $\mathbf{P}_k$, using the camera inverse perspective mapping and the depth image. The input vector for each object k is then $\mathbf{u}_k = (\mathbf{P}_k - \mathbf{r}_k) \cdot s_k$, where $\mathbf{r}_k$ is the particle's position estimate from the previous time step, and s_k is the confidence score produced by the tracking-by-detection module. Using a Rao–Blackwellisation technique [60], only the pose variables need to be sampled, while the occlusion variables can be marginalized out analytically. In order to prevent collisions in the object estimated configuration, the observation model is weighted by a factor that depends on the existence of mesh intersections in the particle estimates. If no intersections exist, this factor is set to 1 not to affect the initial model. Otherwise, it is set empirically to 0.01, which is small enough to penalize invalid configurations and other than zero to avoid eliminating the observation model, if all the samples seem to be intersections.

Visual activity recognition

For understanding nonverbal communication, an efficient multi-sensor visual activity recognition frontend has been developed by experimenting with Dense Trajectories features [61] along with different encoding methods and fusion schemes for visual information processing. Dense Trajectories have been chosen over convolutional neural network pretrained features, because the actions included in the state-of-the-art databases are not similar to those of children, and the fine-tuning of pretrained networks does not perform adequately since, in our case, we have limited data from real-world CRI [16].

Our main goal is to establish a robust framework for tackling different tasks, such as generic body movements performed by kids, with limited training data. We have implemented two different versions of the module in the ChildBot system that work independently, one for gesture recognition and one for action recognition. Although the pipeline for both versions is the same, they are trained, tested, and enabled separately for hand

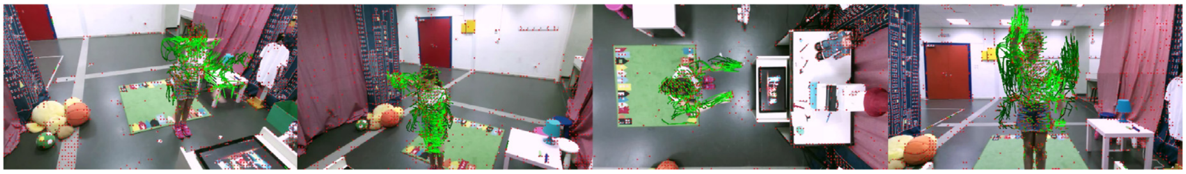


Fig. 5. Example of the extracted Dense Trajectories from different sensor perspectives while the child is performing the swimming pantomime (see Section 4.1).

gestures and more general body movements, respectively. An example of the extracted Dense Trajectories during a pantomime performance (see Section 4.1) is presented in Fig. 5.

In a more detailed view of the system, the recorded RGB video from each of the four RGB cameras is sampled frame by frame. Feature points are sampled for each frame on a grid and are tracked through time based on dense optical flow [62]. Multiple spatial scales are used for the sampling and the tracking independently, while the trajectories are pruned to a fixed length to avoid drifting. The computed features include the Histograms of Optical Flow (HOF) [63] and the Motion Boundary Histograms (MBH) [61] on both axes (MBHx, MBHy).

Afterwards, the features are encoded employing either the zero-order statistics Bag-of-Visual-Words (BoVW) [64] or the first-order statistics Vector of Locally Aggregated Descriptors (VLAD) [65]. Based on their BoVW representation, videos are classified using non-linear Support Vector Machines (SVMs) with the χ^2 kernel, following a similar approach as in [66]. In addition, the above different types of descriptors are combined with the Trajectory descriptor [61] and the Histograms of Oriented Gradients (HOG) [63] by computing distances between their corresponding BoVW histograms and adding the corresponding kernels. Alternatively, the encoded features that result from VLAD are classified employing linear SVMs.

After the feature extraction, we follow three different approaches – in multiple levels – to fuse the RGB information acquired by the multiple sensors: (i) feature fusion, (ii) encoding fusion, and (iii) score fusion. We modify the general frameworks of BoVW and VLAD to deal with our proposed multi-view approach for visual activity recognition.

Feature Fusion: In this method, the visual information is fused at an early stage where only low-level D-dimensional feature descriptors $\mathbf{x}_m^i \in \mathbb{R}^D$ are extracted, i.e., local descriptors alongside dense trajectory $m = 1, \ldots, M_i$, from each different sensor $i = 1, \ldots, S$. The codebook generation approach, which is based on the K-means algorithm, is modified in order to deal with the multi-view data. Given a set of feature descriptors $\mathbf{x}_m^i$, our goal is to partition the feature set into K clusters $[\mathbf{d}_1, \ldots, \mathbf{d}_K]$, where $\mathbf{d}_k \in \mathbb{R}^D$ is the centroid of the kth cluster. These $\mathbf{d}_k$ are shared between the features of all sensors. Using the notation of [64], if descriptor $\mathbf{x}_m^i$ is assigned to cluster k, then the indicator value $r_{m,i,k} = 1$ and $r_{m,i,\ell} = 0$ for $\ell \neq k$. The optimal $\mathbf{d}_k$ can be found by minimizing the objective function:

$$\min_{\mathbf{d}_k, r_{m,i,k}} \sum_{k=1}^{K} \sum_{i=1}^{S} \sum_{m=1}^{M_i} r_{m,i,k} \|\mathbf{x}_m^i - \mathbf{d}_k\|_2^2. \tag{1}$$

Then the encoding procedure is employed for both the BoVW and the VLAD method, resulting in an encoded feature representation $\mathbf{s}_{n_j}^i$ for each trajectory n_j of the jth video captured by sensor i, as explained in [16]. The global representation $\mathbf{h}$ of the multi-view video using a sum pooling scheme is given by:

$$\mathbf{h} = \sum_{i=1}^{S} \sum_{n_j=1}^{N_j} \mathbf{s}_{n_j}^i \tag{2}$$

Finally, for the BoVW approach, an $L2$ normalization scheme [67] is applied, while for the VLAD the intra-normalization strategy proposed in [68] is followed.

Encoding Fusion: In this approach, a different global vector $\mathbf{h}^i$ is created by encoding the dense trajectory features using a different codebook $\mathbf{D}^i$ for each sensor i. For the BoVW encoding, the multi-view fusion is applied by adding the χ^2 kernels:

$$K\left(\mathbf{h}_j, \mathbf{h}_q\right) = \sum_{i=1}^{S} \sum_{c=1}^{N_c} \exp\left(-\frac{1}{A_c} L\left(\mathbf{h}_j^{c,i}, \mathbf{h}_q^{c,i}\right)\right), \tag{3}$$

where $\mathbf{h}_j^{c,i}$, $\mathbf{h}_q^{c,i}$ denote the BoVW representations of the cth descriptor for the jth and qth video respectively captured by sensor i, and A_c is the mean value of χ^2 distances $L(\mathbf{h}_j^{c,i}, \mathbf{h}_q^{c,i})$ between all pairs of training samples from a specific sensor i. On the other hand, for the VLAD encoding, a simple concatenation of the vectors corresponding to the different sensors is applied as follows: $\mathbf{h} = [\mathbf{h}^1, \ldots, \mathbf{h}^S]$.

Score Fusion: For a given sensor i a different SVM is trained for all employed classes and obtains the probabilities $\mathbf{P}^i$ as described in [69]. Then a softmax normalization is applied to each sensor's SVM probabilities. For the fusion of the different sensor output probabilities an average fusion is employed: $\mathbf{P} = \frac{1}{S} \sum_{i=1}^{S} \mathbf{P}^i$. Finally, the class with the highest fused score is selected, following a one-against-all approach.

Distant speech recognition

In order to ensure natural communication between humans and robots in an HRI system, it is essential to incorporate a speech recognition module. A considerable amount of distance between the robot and the users imposes the need to employ a distant speech recognition system (DSR) [70,71] that will have to efficiently address challenging problems, such as noise and reverberation [72]. Especially when children are the end users that play and interact with the robots, speech recognition becomes much more challenging because of the special characteristics of children voices and the difficulty of acquiring quality data.

In our setup, the microphone arrays distributed in space are employed for the DSR task. Children can use a set of utterances adopted for the specific context of the employed use-cases to communicate with the robots, thus our speech recognition system is grammar-based. A continuous system would require a large amount of children data to be collected; that was unfeasible in our case. Also, a grammar-based speech recognition system is adequate to fulfill the requirements of the considered use-cases. The employed language is Greek, and the set of utterances contains possible children answers in some games and some general-purpose speech utterances.

The DSR system is able to detect and recognize the spoken utterances at any time, namely, it is always-listening. Since speech is usually corrupted by reverberation, noise, or other non-speech events, robustness is achieved via beamforming of the far-field signals and adaptation of the acoustic models.

More into the details, a sliding window of 2.5s duration with a 0.6s shift is used to process speech in time frames. A custom module has been developed and integrated with Robot Operating System (ROS), allowing raw audio processing from the Kinect microphones. Each speech frame is first denoised with simple delay-and-sum beamforming applied on each available 4-channel Kinect array: The insertion of delays to the different microphone signals $a_n(t)$, allows us to align them appropriately, in order to enhance speech coming from a specific direction. For uniform linear arrays with N microphones, which is also our case, if the desired direction is denoted by ϕ, the time-delay to be applied to each microphone is

$$\tau_n = \frac{(n-1)d\cos\phi}{c}, \tag{4}$$

where c is the speed of sound and d the space between microphones. The beamformed signal is denoted by:

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} a_n(t-\tau_n) \tag{5}$$

The denoised signal is then fed to the DSR module, where we enforce recognition of one of the pre-defined sentences. Regarding acoustic modeling, Gaussian Mixture Models (GMMs) and Hidden Markov Models built on cross-word tri-phone models have been trained using standard Mel-Frequency Cepstrum Coefficients-plus-derivatives features on the Logotypografia database [73] that contains clean, close-talk speech in Greek. Thus, we artificially distort the database by convolving the clean speech with room impulse responses and adding white Gaussian noise in order to match the far-field condition [71]. Maximum likelihood linear regression (MLLR) adaptation is employed to transform the GMM means, aiming to reduce the mismatch between the initial model and the adaptation data [74].

3.3. System architecture

In this subsection, we describe the backbone of the perception system: the hardware architecture, the interconnection and communication between the different modules, and how the interaction flow is managed.

The perception modules are integrated into the full perception system based on the following hardware architecture. The system runs on four distributed interconnected machines, three of which run the Linux operating system and the ROS, and one the Windows Operating System. Each of the three Linux machines is connected with a Kinect V2 sensor which provides raw data (i.e., color, depth, and audio). The Windows machine is also connected to a Kinect V2 sensor, and using the Microsoft SDK Kinect V2 API provides additional skeletal and tracking information. A touch screen is also connected to the Windows machine and sends feedback to the dialog module about the children's choices. The main data processing of the perception modules takes place on each of the three Linux machines, while the multi-view fusion is handled in one of the Linux machines. Streaming of data and communication between the system modules flows via events transmitted through the TCP/IP broker, which runs in the Windows machine and is provided by the IrisTK framework [75]. Under the IrisTK paradigm, we divide events into three classes:

- *Sense*: events that include information about what the sensors of the system perceive
- *Action*: events that order an actuator (i.e., a robot) to do something
- *Monitor*: background events that contain feedback information about the actions of the system (e.g., when a robot has ended speaking)

Similarly, the architecture of the system was designed based on the *Sense–Think–Act* principle [15], as shown in Fig. 1. The multi-sensor setup of the system represents the Sense part, while the perception modules are classified into the Think principle. Finally, the multiple robots belong to the Act part of the architecture. This three-layer architecture allows for high-level modularity because the different layers can be replaced/modified without affecting the others.

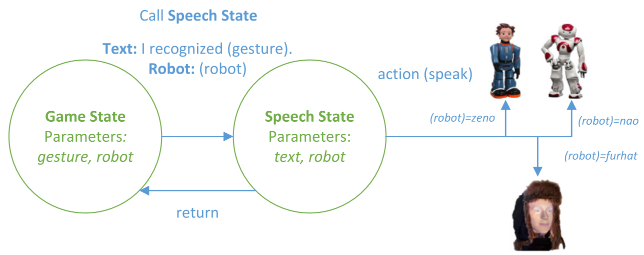


Fig. 6. The "Speak Action State" employed for announcing a gesture during the interaction.

The broker, along with the dialog management module which will be described next, acts as a central unit that receives events from all system modules and distributes them accordingly to the appropriate modules. This unit offers extra fine-grained modularity allowing modules to be easily removed or added to the architecture by simply defining the sets of events that the module should perceive or send back to the broker.

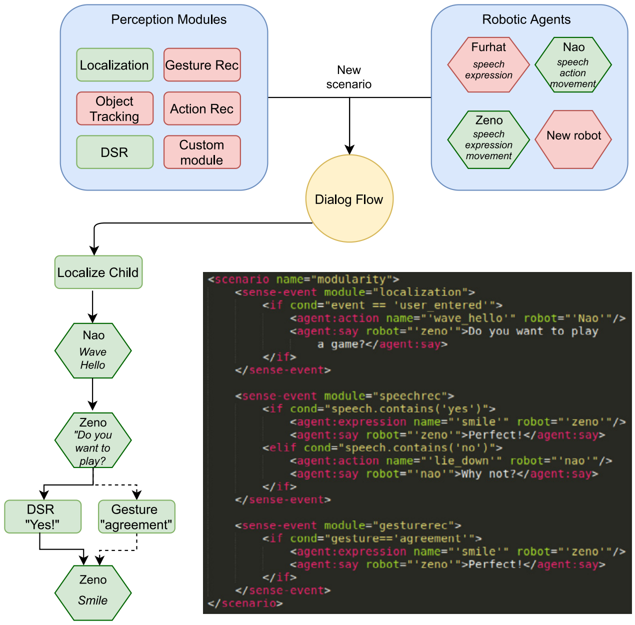
The dialog manager is the central module of the system and models the interaction flow between the user and the system. The interaction is modeled using Harel states [76]- states that can be hierarchically structured, executed conditionally, and contain parameters that alter the flow and transitions. In addition, states can be called as functions, which means that the flow of the execution will continue to the caller state after the callee has finished his execution.

We have included "action states", i.e., states that act as a mediator between the core dialog flow and the robots for the design and development of the statechart that models the dialog. These action states contain the information needed to instruct the system's robots to perform an action and include the robot as an additional state parameter. As a result, the core dialog flow is decoupled from robot-specific details, and we avoid defining multiple similar states for different robots. This extension also allows us to easily include new robots in the dialog flow by adding the robot-specific details in the action states and handling the event on the robot side. An example can be seen in Fig. 6 where a state in the core dialog flow "calls" the "speak action state", including the robot that is needed to speak as an additional parameter.

From the three robots that our multi-robot system uses, the Furhat robot head is already integrated into the IrisTK framework. For the NAO and Zeno robots, we developed intermediate APIs that we use for communication between the robots and the dialog.

Modularity and new scenarios:. We have described the modularity of the proposed system, which extends both in terms of the perception modules and switching robots and adding new robots to the system. As a result, the system can be seen both as an integrated framework, but also as an aggregation of different robots and modules, which can be switched on/off according to the desired application.

We show an example of this modularity in Fig. 7, where a simple scenario that employs only two robots (Nao and Zeno) and two perception modules (DSR and localization) is shown. It is evident that due to the system's architecture, apart from the Dialog Manager, all other perception modules and robots can be switched on/off without affecting the system functionality. For



```
<scenario name="modularity">
    <sense-event module="localization">
        <if cond="event == 'user_entered'">
            <agent:action name="'wave_hello'" robot="'Nao'"/>
            <agent:say robot="'zeno'">Do you want to play
                a game?</agent:say>
        </if>
    </sense-event>

    <sense-event module="speechrec">
        <if cond="speech.contains('yes')">
            <agent:expression name="'smile'" robot="'zeno'"/>
            <agent:say robot="'zeno'">Perfect!</agent:say>
        <elif cond="speech.contains('no')">
            <agent:action name="'lie_down'" robot="'nao'"/>
            <agent:say robot="'nao'">Why not?</agent:say>
        </if>
    </sense-event>

    <sense-event module="gesturerec">
        <if cond="gesture=='agreement'">
            <agent:expression name="'smile'" robot="'zeno'"/>
            <agent:say robot="'zeno'">Perfect!</agent:say>
        </if>
    </sense-event>
</scenario>
```

Fig. 7. Showcase of the system modularity system through a simple custom scenario using two robots and two perception modules.

example, in the shown scenario, after Zeno asks the child, "Do you want to play?", the scenario can include either the DSR module or the Gesture recognition module (or it could be both) to recognize the child's answer.

Note that while we have described the entire architecture of our multi-sensor and robot integrated framework in this section, the only hard requirements for running the system include a machine running ROS and the IrisTK-based Dialog Manager and Broker, both of which are open source. Furthermore, while we have used Kinect V2 sensors (which also carry the inherent limitation of one sensor per machine), the system can use any RGB sensor and microphone supported by ROS.

4. Use-cases for CRI

This section presents the indicative use–cases that we have designed and implemented in order to showcase and evaluate the ChildBot capabilities. Moreover, we describe the collected data of the database and the protocols of the conducted experiments.

4.1. Tasks description

A set of scenarios has been designed to highlight the system's capabilities during an amusing and educative multimodal interaction between children and robots. As explained extensively, our integrated system perceives various events that occur during the interaction, such as children speech and activities, children locations in the room, and tracking of objects. Each task employs different technologies and combines them appropriately to create a smooth interaction. Children are asked to complete the following tasks-games: (i) "Show me the Gesture", (ii) "Express the Feeling", (iii) "Pantomime", (iv) "Assembly Game", (v) "Form a Farm".

In the first task, "Show me the Gesture", a child interacts with the robot via gestures and speech. The robot requests the child to perform a gesture that usually denotes a meaning and tries to recognize it. It then asks the child for confirmation of its guess. The different gestures of this game are: (i) stating an agreement, (ii) calling the robot to come closer, (iii) asking the robot to sit down, (iv) pointing an object in the room, (v) asking the robot to stop, and (vi) drawing a circle in the air. Except for the first gesture that is usually performed by nodding, the rest are hand gestures. The children are allowed to gesture spontaneously, as they would do when interacting with another human.

The "Express the Feeling" game motivates children to reveal their feelings using both their face and body during an entertaining interaction with the robot. In this game, the child selects one of the cards depicted on the touch screen and expresses the chosen feeling. The emotions included in this game are happiness, sadness, fear, anger, surprise, and disgust. After the child's reaction, the robot also expresses the same feeling using its body and face.

"Pantomime" is a popular game, during which, one person mimes handwork and the other figures out the depicting handwork. The child can use the whole body to mimic an activity and interact extensively with the robot. The robot and the child repeatedly swap the roles of the mime and the guesser. The twelve activities used in this game are the following: (i) cleaning a window, (ii) driving a bus, (iii) hammering a nail, (iv) swimming, (v) working out, (vi) dancing, (vii) reading a book, (viii) digging a hole, (ix) playing the guitar, (x) wiping the floor, (xi) dancing, and (xii) ironing a shirt.

For the "Assembly Game", one or more children are asked to complete an assembly under robot supervision. Six interconnectable 3D printed bricks of different lengths are used to create rectangles and squares. The bricks are placed on a table in front of the child, with the robot standing close by. The child is responsible for manipulating the assembly subcomponents while the robot provides instructions and feedback. If the child correctly completes a connection, the robot congratulates the child and gives the next instruction. However, if the child makes a mistake, the robot will attempt to recognize this mistake and react accordingly. Aside from verbal instructions, the robot also looks and points at the bricks that it refers to, for clarity.

The "Form a Farm" game is a multi-party game scenario involving two roles that can be interchanged and equally played by both the robot and the children, aiming to entertain, educate, and establish a natural interaction between all parties. The game involves two different roles: the picker and the guesser. The picker chooses an animal and utters its characteristics. The guesser has to guess the picked animal. The interaction proceeds as follows: At first, the robot chooses a random animal, and the human players take turns guessing the chosen animal. In case of a wrong guess, the robot reveals more animal characteristics (animal color, number of legs, animal class, e.g., mammals, reptiles). In case of correctly identifying the animals, the robot asks the children to properly place the animal in a farm with some distinct segmented areas, which appears on a touch screen in front of them. In the second round, the roles are reversed: children discuss and pick an animal and reveal one characteristic. The robot then tries to guess the picked animal. If the robot guesses correctly, the children are again asked to place the animal inside the farm, else they reveal more animal characteristics, one at a time. The game continues by interchanging the role of the guesser between children and robots in each round. The game features a total of 19 animals, and their characteristics belong to five different classes: color, size, species, number of legs, and a distinctive feature, i.e. for the dog: "it's the human's best friend".

The tasks mentioned above aim to create a proper framework for multimodal communication between children and robots, as it happens between humans. This way, the tasks demonstrate the system's capabilities and give some examples of how ChildBot can be used and can be employed for system evaluation. Even though each task focuses on one of the system perception technologies, more than one modules are used in parallel. In Table 1, the used

Table 1
Used ChildBot technologies in each use-case scenario and the eligibility of each robotic agent for participating.

	Distant Speech	Detect & Track	Speaker Local.	Visual Activity Rec.		Touch Screen	Behavioral	Robots		
	Recognition			Gesture	Action		Generation	NAO	Furhat	Zeno
Show me the Gesture	✓		✓	✓			✓	✓	✓	✓
Pantomime	✓		✓		✓	✓	✓	✓		
Assembly Game		✓	✓				✓	✓		✓
Form a Farm	✓		✓	✓		✓	✓	✓	✓	✓
Express the Feeling						✓	✓		✓	✓

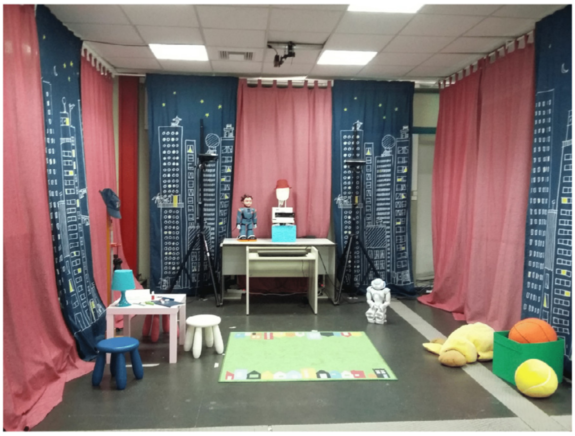


Fig. 8. Data collection room and experimental setup.

modules are summarized along with the eligibility of the robots to participate in each task.

Considering that the tasks are intended for children, the use-cases were designed under the supervision of psychologists and in collaboration with the consortium of the BabyRobot project. Specifically for the single games, we conducted pilot studies with eight children in our specially designed room, which led to the presented scenarios. An example of how the use-cases are improved and adapted to the children's needs is the game "Form a Farm". It was initially proposed as a game called "move together" where the participants collaborate to decorate their house. After pilot studies in two countries, Greece and Sweden, it was concluded that while the game format (including interaction with the touch screen) was suitable, the subject was complex for children less than 12 years old. As a result, the subject of the game was changed to a more familiar theme for children, farm animals. We encourage readers who want to delve deep into the experimental conditions and protocols to read the relevant deliverables of the BabyRobot project (D 4.1–4.4, 5.1–5.4) [77].

4.2. Database

Real data obtained through HRI prove to be especially important while developing a system, from the training to the evaluation stage. Such data contribute to an adaptation of the system to actual circumstances and spontaneous human behavior. Thus, extensive data collection has taken place with the participation of a pool of 52 children, aged from six to eleven years old, in a specially designed room and a school classroom.

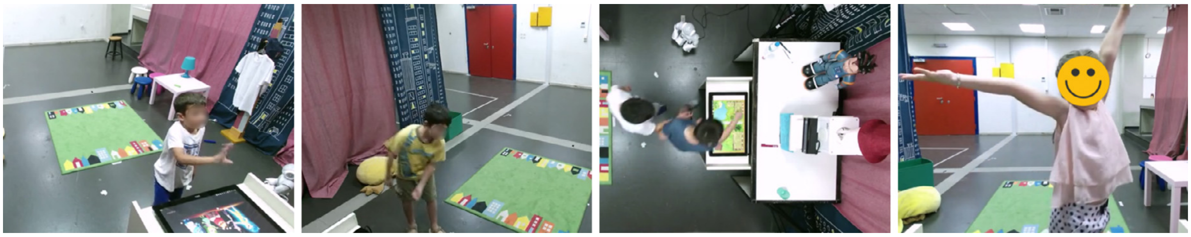
Most of the data have been collected in a room that resembles a child's room and where the robotic agents and the sensors have been located, as is presented in Fig. 8. There, the data collection has been carried out in two phases. In the first one, children data have been recorded while performing certain actions and uttering certain phrases that are expected to arise throughout the interplay between them and the robots, in a strictly controlled way, when asked to do so. These data will be referred to below as *development data* since they have been used to develop the system. In the second phase, the data were collected during the experimental procedure where children interact with robots without interruption or other people's intervention. The latter data will be referred to as *use-case related data*. Both types of data are equally important for CRI, as the first one is indispensable for training the perception modules on data that are relevant to use-cases, while the second one is essential for the testing of the behavioral monitoring software during CRI. Table 2 presents the most important recorded events during the two phases and the total number of their occurrences.

The information we collected during the data collection includes Full HD (1920 × 1080) RGB and depth (512 × 424) video streams from all four Kinect cameras, running at 30fps, as well as raw audio from the microphone array embedded in each Kinect sensor. By exploiting the Kinect v2 API we have also captured the following streams from the frontal Kinect sensor: (a) Skeletal information both in 2D (image) and 3D (world) coordinates; (b) Bounding boxes from face detection, facial landmarks, and a facial 3D mesh.

For the development data, 28 children have participated by performing seven gestures and twelve pantomimes, and uttering 40 phrases from a vocabulary of 120 phrases. This phase is crucial for developing the perception models and adapting them to children since they focus on speech, gestures, and actions relevant to the use-cases. Specifically, children are more spontaneous and expressive than adults and their speech is usually brief and low-voiced. Thus, in order to test the performance of ChildBot modules, it is necessary to have a plethora of children activities and utterances. Moreover, adults' data have been collected to augment the use-cases' data, validate, and highlight the need for children data for enhancing performance in the perception models.

As far as the use-case related data are concerned, 31 children with an average of 8.6 years old, 10 girls and 21 boys, had been chosen randomly from a set of volunteers that met our team in a dissemination event. From six to eleven years old, all children spoke Greek and were able to read and write. Each child accompanied by his/her parents entered the specifically designed room and was introduced to the robots by a researcher. The child got familiarized with the room and the robots while the researcher explained the procedure's structure and the tasks presented above. Afterwards, the parents and the researcher exited the interaction space, and the child played individual games with the robots. After completing the individual interaction, a second child (who had completed the same interaction previously) entered the space and collaborated with the other child while playing the "Form a Farm" task (see Fig. 9). In cases where there was no second child available, an adult took its place. However, these data were removed from the subsequent evaluation. Finally, after completing the procedure, the children were asked to complete a questionnaire that included subjective statements regarding their experience. The questionnaire will be described and discussed in Section 5.2.

The Ethics Committee of the Athena Research Center has approved the above procedure, and a consent form sent by email



(a) "Show me the Gesture", Kinect#1 (b) "Pantomime", Kinect#2 (c) "Form a Farm", Kinect#3 (d) "Express the Feeling", Kinect#4

Fig. 9. The four different use-cases that took place in our laboratory, each one presented from one the four different camera viewpoints of ChildBot.

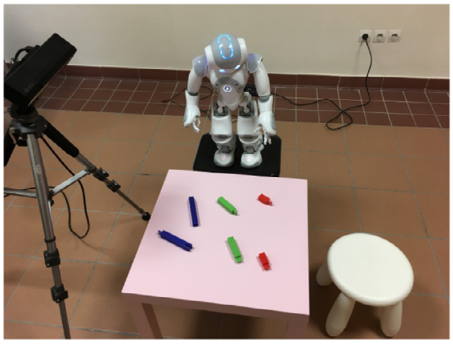


Fig. 10. Setup of the "Assembly Game" at a Greek primary school.

Table 2
Statistics of the most important child activities during the data collection.

Collected data	Event type	Number of events
Development data	Utterances Gestures Pantomimes	977 196 336
Use-case related data	Utterances Gestures Pantomimes	641 143 109

to the parents before conducting the experiments. In addition, all experiments have been supervised by an experienced child psychologist.

The use-case related data regarding the "Assembly Game" were collected in a Greek primary school from 21 students with an age span between nine and ten years old. Six students participated individually and the remaining were organized in groups of five, according to the teachers' advice, in order to reinforce their collaborative skills. For this task, a single Kinect camera and one robotic agent (NAO robot) have been chosen as a lightweight version of the system to accommodate the educational process. Such a version can be easily installed in a typical classroom and help the teacher give a vivid lesson through a CRI experience (Fig. 10).

5. CRI evaluation

Each perception module of the ChildBot system has been evaluated by measuring its performance in efficient multimodal scene understanding using the collected data. In addition, we have performed a preliminary user experience study to assess how the children interact and perceive the system and collect insights towards carrying out a complete subjective evaluation in the future.

Table 3
Evaluation of the audio-visual active speaker module.

Audio source localization			Audio-visual active speaker localization
Pcor	RMSE	RMSEf	Pcor
45.51%	0.60 m	0.35 m	85.58%

5.1. Perception module evaluation

Audio-visual active speaker localization

The evaluation results of audio-visual speaker localization are presented in Table 3. For audio-only speaker localization, the employed metrics are Pcor (Percentage correct) which is the percentage of correct estimations (deviation from ground truth less than 0.5m) over all estimations, RMSE (Root Mean Square Error) between the estimation and the ground truth, and RMSEf (RMSE for estimations with error less than 0.5m - i.e., 'fine errors'). For audio-visual speaker localization, since person locations are estimated by the Kinect skeleton, the problem is essentially transformed into an active speaker localization problem. Thus, evaluation is performed in terms of correct speaker estimation, where Pcor is used, denoting the correct speaker estimations over all estimations. Audio-only localization does not perform sufficiently well yielding a Pcor of 45%, but the average RMSE is 60cm, meaning that the average source localization error is 60cm which is not very large. If both audio and visual information are used, then the active speaker localization performance is boosted to a Pcor of 86%.

6-DoF Object Tracking

For object tracking, we have performed both an objective evaluation and a subjective evaluation to assess the performance of the 3D visual tracking module.

During the objective evaluation, because it is difficult to annotate and obtain ground truth poses for 3D tracking, we have placed two static objects on a table, along with obstacles, in order to add occlusions. We have also moved a camera around the objects with sudden movement bursts to establish the tracker's robustness. We have compared our method with an SDF (Signed Distance Function) tracker [78]. Our results have showed that, although the SDF tracker has produced a steadier output than our tracker in cases of partial occlusions and slow camera movements, when we have introduced sudden jolts and full occlusion, the SDF tracker has been unable to continue tracking and has failed, without recovering, even when the normal conditions (no occlusion-jolts) were restored. On the other hand, our tracker has been able to successfully track the object with low error, even under full occlusion and fast movements, and recover in rare cases where the tracking has been lost, without the need for reinitialization proving its robustness. More information on these objective evaluations can be found in [17].

We used the NAO robot as a supervisor for the Assembly Game during the subjective evaluation, which is described in Section 4.

Table 4
Statistics about the performance of the 6-DoF object tracking employed in "the Assembly Game".

	Total connections (Recall %)		Required connections (Recall %)	
Identification Time	5s	20s	5s	20s
Rectangle	70.00	80.00	50.0	56.25
Square	39.39	57.58	43.24	59.46

Table 5
Examination of activity recognition modules performance (accuracy (%)) when using different age groups for model training (adults, children, or mixed data). Bold values denote the best obtained accuracy for each testing age group and the activity recognition module (gesture, action). Scores shown were obtained after the fusion of individual camera scores using MBH features and BoVW encoding.

	Testing group	Training group		
		Adults	Children	Mixed
Gesture recognition	Adults	92.19	62.08	**95.10**
	Children	56.25	**83.80**	80.09
Action recognition	Adults	**87.36**	72.53	86.26
	Children	56.51	**74.46**	74.26

Of the 21 participants, six played the game on their own, while the remaining children played in groups of five. The children were required to complete two different rectangles and one square by choosing and connecting items from six different brick objects.

In Table 4 we can see the results from the experiments in the form of statistics for the different assemblies. We present the percentage of the total and required connections that the system recognized within a time interval of 5s and 20s, referred to as identification time. The term "total connections" includes both correct connections that the child completed and mistaken connections, In contrast, "required connections" refers only to the correct connections needed to complete the assembly.

Visual activity recognition

First, we have examined if the age group of the participants (adults or children) impacts the accuracy of visual activity recognition. For both visual activity recognition tasks, we trained separate models using as the training set: (a) children, (b) adults, (c) mixed (both adults and children) and as the testing set: (a) children, (b) adults. In Table 5, it can be noticed that the use of children training data is imperative for achieving high accuracy in children activity recognition, irrespectively of the task. This result justifies our choice for collecting development data from children movements. On the other hand, recognition models trained on mixed age groups perform better for adult gesture recognition since the diversity with which children perform the gestures accommodates the model's generalization. For the action recognition task, children employ a wider range of different movements that adults do not use as they act stereotypically, and the mixed age training models perform worse than the adults' models.

Furthermore, Tables 6 and 7 summarize the evaluation of gesture and action recognition respectively, for several combinations of different features, encodings in both the single-view case and multi-view case along with the multi-level fusion. Also, the recognition models have been trained on children development data and tested on both development and use-case related data separately, using the leave-one-out cross-validation approach.

Specifically, Table 6 presents average accuracy results (%) for the seven gestures and a background model. Results indicate that the best multi-view model outperforms the best single-view model by about 7%, underlining the need for a multi-view system for unrestrained CRI. The development data shows that the combination of different types of features performs better than HOF and MBH features individually. Among the single-sensor cases, Kinect#1 (right side view) performs best as most of the kids are right-handed and stood at approximately the same location while performing the gestures. The best recognition accuracy of 85.19% is noticed for the fusion in the final step of the procedure with the VLAD encodings and the feature combination. As far as the use-case related data are concerned, the accuracy for the single streams is moderately lower than the previous ones, which reveals the difficulty that children faced while trying to perform the gesture spontaneously. Also, as the children stand at a completely different location, usually closer to the cameras, the best single stream result appears for Kinect#3 (floor plan view). Regarding the fusion of the different streams, recognition performance is slightly better for the encodings fusion than the scores fusion, and it approaches 74%. More generally, VLAD encodes more effectively the visual information than the BoVW, since it contains rich information about the distribution of the visual words. Finally, we have to note that, as nodding requires a gentle movement, it is usually confused with the background movement.

In order to verify the appropriateness of the proposed visual activity recognition system in more challenging tasks, we evaluate the visual activity recognition system for the pantomimes. Table 7 presents the average accuracy results (%) for the 12 pantomimes and the background model. The fusion of the single-view information remarkably enhances the recognition's performance, as was observed in the gesture case. The highest accuracy for testing on development data appears with VLAD encodings in scores fusion since the visual information in these data is more consistent than use-case related data, e.g. similar time duration of a pantomime or similar children locations in the room. Regarding the single-view case, in both types of data, the right-side view Kinect#1 appears to be the best perspective for the trained models. It can be noticed that in use-case related data, MBH yields slightly better results than feature combination. Moreover, feature fusion, i.e. the fusion of the information at an early step of the entire procedure results in the best performance regardless of the type of the encodings.

In conclusion, the accuracy of the visual activity recognition is lower in use-case related data since children act more spontaneously while they move around the room and interact freely with the robots. Furthermore, since the variation of the visual information in the pantomime task is larger than in the gesture task, the early fusion of the features performs better for the pantomime while the scores fusion is satisfactory for the "Show me the Gesture" since only one camera is adequate to recognize the gesture.

Distant speech recognition

Two sets of data have been employed for the offline evaluation of the DSR task: the development data and the use-case related data, both consisting of children data. As stated before, the DSR system is grammar-based, namely depending on the context of the application, and there is a specific set of commands that the users adopt to communicate with the robot. Thus, we have designed one set for the "Show me the Gesture", the "Express the Feeling", and the "Pantomime" games and another one for the cooperative game, i.e. the "Form a Farm" game. The grammar size and other statistics concerning the two datasets can be found in Table 8.

The development data have been used for adapting speech models and testing them. Results are presented in Table 9 in terms of word and sentence accuracies, denoted by WCOR and SCOR, respectively. Four different adaptation schemes have been tested for comparison: In the "No-adapt" case, the employed models have been trained on the Logotypografia database, which contains adult data. The available children data included in the

Table 6
Average classification accuracy (%) for the employed 8 gestures. Results on both development and use-case related data are shown for the different features, encoding, and fusion methods of the activity recognition module. Bold values denote the best single camera score and the best fusion scheme scores.

Features	Development data									
	Single camera				Fusion					
	Kinect #1	Kinect #2	Kinect #3	Kinect #4	Features		Encodings		Scores	
	BoVW				BoVW	VLAD	BoVW	VLAD	BoVW	VLAD
HOF	70.83	70.37	69.21	63.43	71.76	74.07	77.78	81.48	75.93	81.94
MBH	76.85	67.82	68.29	65.28	76.39	76.85	81.02	81.48	82.87	83.80
Traj.+HOG+HOF+MBH	**77.78**	73.84	73.61	75.00	81.48	82.87	82.87	83.80	82.87	**85.19**
Features	Use-case related data									
	Single camera				Fusion					
	Kinect #1	Kinect #2	Kinect #3	Kinect #4	Features		Encodings		Scores	
	BoVW				BoVW	VLAD	BoVW	VLAD	BoVW	VLAD
HOF	56.92	54.49	57.10	51.97	54.56	71.61	58.01	74.73	63.26	74.83
MBH	62.70	56.47	60.15	54.25	65.32	72.70	67.72	72.52	66.73	72.72
Traj.+HOG+HOF+MBH	57.96	54.08	**67.03**	59.16	61.51	69.85	63.38	**73.95**	64.82	**73.35**

Table 7
Average classification accuracy (%) for the employed 13 pantomimes. Results on both development and use-case related data for the different features, encoding and fusion methods of the activity recognition module are depicted. Bold values denote the best single camera score and the best fusion scheme scores.

Features	Development Data									
	Single camera				Fusion					
	Kinect #1	Kinect #2	Kinect #3	Kinect #4	Features		Encodings		Scores	
	BoVW				BoVW	VLAD	BoVW	VLAD	BoVW	VLAD
HOF	68.31	56.31	48.62	53.85	66.77	67.08	68.00	69.23	68.62	75.50
MBH	70.77	60.92	61.85	55.22	76.00	76.69	76.92	76.92	74.46	76.50
Traj.+HOG+HOF+MBH	**73.85**	63.38	60.00	61.45	75.08	76.92	77.23	77.85	75.08	**79.00**
Features	Use-case related data									
	Single camera				Fusion					
	Kinect #1	Kinect #2	Kinect #3	Kinect #4	Features		Encodings		Scores	
	BoVW				BoVW	VLAD	BoVW	VLAD	BoVW	VLAD
HOF	46.34	46.19	25.50	47.70	63.08	61.02	49.87	56.17	52.59	57.99
MBH	**61.42**	46.28	31.59	45.57	**70.25**	67.97	57.70	59.04	62.18	62.49
Traj.+HOG+HOF+MBH	52.59	46.74	36.62	48.16	63.52	**69.37**	60.75	61.55	55.00	64.90

development data have been used for testing. In the “Adults" case, speech models have been adapted to a small amount of adult data and tested both with adult and children data. In the “Children" case, data from 20 out of 28 participants of the development data have been used to adapt speech models globally, i.e. data from the Kinect arrays have been used to adapt a single model. The remaining eight participants form the testing set. The adaptation and testing have been four-fold cross-validated. Finally, in the “Mixed" case, we have used both adult and children data to adapt the models and then test them separately on adult and children data.

Speech recognition achieves satisfactory performance for adults even without adaptation. However, adaptation indeed improves performance for all cases, even when performed in a different age group than testing. Results indicate that the best performance is obtained for adapting and testing on the same group, which was expected. The best achieved results are 98.87% for adults and 95.50% for children in terms of SCOR. The results concerning children underline the need and importance of collecting children data. Performance is boosted, from 75.3% to 97.8% for WCOR and from 71.2% to 95.5% for SCOR, when children data have been used for adaptation and testing, who are indeed the target group of the system. All the above results refer to the development data where data collection was controlled and guided. On the other hand, use-case related data contain data collected in the wild, i.e., while children were playing with the robots. Although the children had received some instructions concerning the utterances they could use, it is obvious that they were not followed in most cases. Thus, after the data annotation, new grammar sets have been formed in order to incorporate new phrases.

The use-case related data results are depicted in Table 10, in terms of WCOR, SCOR, and LabelCOR. LabelCOR refers to the percentage of correct recognition of the semantic content. For example, there can be various ways to express a negation: “no", “no, I don’t know", “no, I did not find it", etc. All similar phrases in terms of content are given a specific label, and after speech recognition data post-processing calculates the score of the correct recognition of labels. Adaptation has been performed using the development data. We notice that adaptation boosts the performance, achieving a percentage of 59.64% for the gesture and pantomime games and 78% for the farm game in terms of WCOR. Both results indicate that children speech recognition is very challenging in real conditions, because children voice and articulation can be unintelligible and unclear when they are speaking in a spontaneous, continuous way. The lower percentage of single games can be attributed to the larger grammar size, the distance between the speakers and the microphones, and the relatively large variations of speaker orientation.

5.2. User experience study

Objective Statistics: Regarding the individual tasks (“Show me the Gesture”, “Express the feeling”, “Pantomime”), where each of the 31 kids participated alone, all of them were able to complete the games successfully. The average duration of completion, including the introduction by the robots, is nine minutes, with a variance of two minutes.

Table 8
Data statistics for the children DSR task. Number of children speakers, utterances recorded, and grammar size in the development and use-case related data for each group of games.

	Development data			Use-case related data		
	# speakers	# utterances	grammar size	# speakers	# utterances	grammar size
Single games	28	642	75	31	426	157
Cooperative game	28	335	58	9 pairs	215	113

Table 9
Evaluation of the DSR recognition on the development data. Accuracy (%) of the average Word recognition (WCOR) and Sentence recognition (SCOR) for the different adaptation schemes and each testing age group. Bold values denote the best recognition results for each testing age group.

Test	DSR-Adaptation scheme							
	No-adapt		Adults		Children		Mixed	
	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Adults	97.54	91.25	99.58	**98.87**	96.73	93.20	99.50	98.43
Children	79.06	69.95	75.31	71.20	97.81	**95.50**	90.71	82.06

Table 10
Accuracy (%) of the average word (WCOR), sentence accuracy (SCOR), and Label accuracy (LabelCOR) for the children DSR task on the use-case related data.

	No-adapt			Adapt-all		
	WCOR	SCOR	LabelCOR	WCOR	SCOR	LabelCOR
Single games	56.68	29.52	55.12	59.64	43.77	55.12
Cooperative game	72.95	61.02	63.16	78.00	67.69	70.51

In 32% of the cases, a human (verbal) intervention was required up to two times during the experimental flow when the children became confused or had questions about the procedure. For example, some children asked for a confirmation about what to do or needed a prompt in order to act. Such possible deviations from the designed scenario have been overcome by enabling the dialog manager to recognize these cases (e.g., if the child is silent for a reasonable time) and getting the robots to prompt the children or ask them to repeat their utterance/activity. In cases where children were expected to say something or their speech was not recognized, robots requested for repetitions up to two consecutive times, while in the case of a child's action, the robots asked for repetition only once.

For the collaborative "Form a Farm" game, played by two children, it was observed that younger children faced difficulties with the game's rules, even though primary school children are familiar with the animals of a farm. As a result, kids aged six and seven played the game following the guidelines offered by one adult. The rest of the children played the game without any guidance. The average duration of the game was eight minutes. In total, children assumed the role of the guesser for 24 rounds and found the correct answer using 2.4 guesses on average and four guesses maximum. On the other hand, the robot assumed the guesser role for 22 rounds and found the correct answer in 2.2 guesses on average, with a maximum of six. Children did not identify the picked animal in 4% of the guesses, while the robot in 32%. Generally, the children managed to guess the picked animal more easily since the robot was programmed to always reveal more general animal characteristics in the beginning and proceeding with more specific details.

We have also performed a user experience study which included a pool of 52 children, from six to eleven years old, participating in the designed interaction described in Section 4.2. The purpose of this study is to collect objective statistics and insights and get a measure of the system's ability to accommodate a complete CRI.

User experience assessment: Regarding subjective evaluation of the experience, children were asked to complete a questionnaire containing the subjective statements that can be seen in Fig. 11. Each statement was accompanied by a 5-point Likert type ordinal scale labeled from "disagree" to "agree", using smiley faces [79].

We also included two multiple-choice questions asking children to justify which use-case was the most preferable and why, and which perception ability of the robots makes them popular to the kids to get more insight into their preferences. In general, the favorite use-case of the children, with 12 out of 31 preferring it, was "Pantomime" because they liked the robot's movements. As we can see in Fig. 11, most of the children (27/31) stated that they like playing with the robots, while 22 enjoyed playing because robots understood both their movements and speech. Many of them (21/31) also found the interaction and use-cases easy to follow, without external help (19/31). Furthermore, children tended to agree (20 positive answers out of 31) that robots behave like humans. By analyzing the questionnaire responses, we noticed that older children stated that they did not need prior knowledge to play with the robots, compared to younger children who stated that they did.

Similarly, in the assembly task that was evaluated in the primary school, 21 children were asked to express their opinion for the interaction. The questions are presented in Table 11, and the available responses were a 3-point Likert scale (Disagree–Neutral–Agree). The Table also presents the questionnaire results after being mapped to a scale of 0–2, with 0 being the most negative. Their answers indicate that children were pleased with the interaction (1.81 MOS on whether they would like to play again with the robot and the comfortableness of the interaction). However, clearly the robot supervision for the assembly task has room for improvements, since although the instructions of the robot were very clear (1.95 MOS), children were neutral on whether they were helpful (1.10 MOS), or wrong (0.95 MOS).

It is important to note that the questions presented in Fig. 11 and Table 11 are not translated precisely from the Greek language in English (we use more formal language in this text), and the original ones were adapted to the children's knowledge level. Rarely, younger children (mainly those of six to seven years old) were helped by their parents in explaining to them if some question seemed to be fuzzy.

Discussion: In general, the evaluation of user experience during interaction with our multi-robot, multi-tasking, and multi-sensor robotic system provided encouraging results. It is found that the system is technically capable of accommodating a complete CRI experience, with some adult intervention needed in

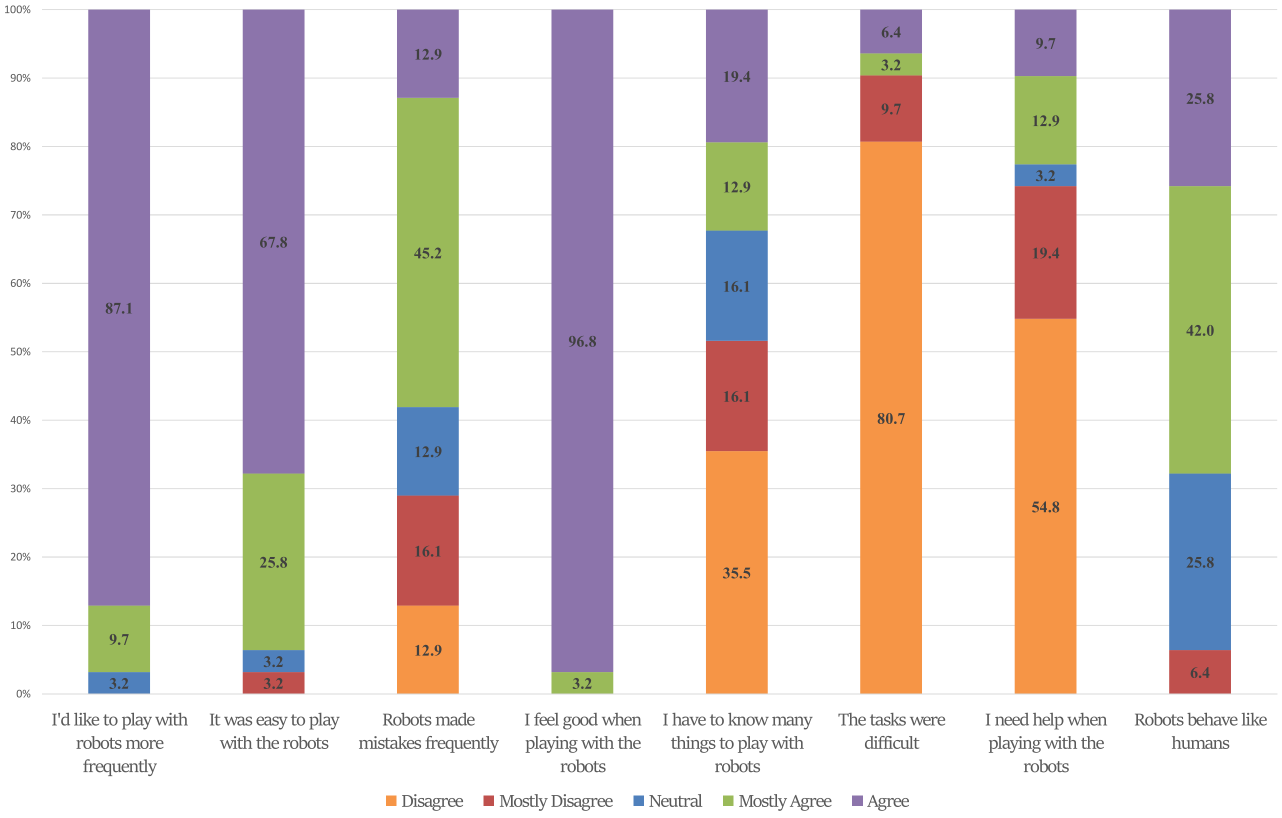


Fig. 11. User experience of the entire ChildBot system. After each completed interaction, children were asked to fill a questionnaire with the shown questions on a 5-point Likert-type scale labeled from "Disagree" to "Agree" as shown with color codes.

Table 11
Questions and results of the questionnaire presented to the children, following their "Assembly Game" with the robot. (Opinion scale 0–2)

Question	Mean opinion score
Were you comfortable working with the robot?	1.81
Would you play with a robot again, sometime?	1.81
Was the robot helpful?	1.10
Did the robot make a lot of mistakes?	0.95
Were the robot's instructions clear?	1.95

certain instances and mainly for the collaborative task. Of course, there is room for improvement since many children stated that robots frequently made mistakes (18/31). In the future, we also aim to conduct a subjective evaluation focused on the pedagogical aspect of the system, based on the insights collected during this initial study.

6. Conclusion

In this paper, we have presented ChildBot, a multimodal perception framework that is the culmination and the extension of several earlier works by the authors in multimodal perception and CRI. ChildBot constitutes a CRI framework with multiple robotic agents that can be successfully used for edutainment purposes, and its perception system includes several different modules: audio-visual active speaker localization, and 6-DoF object tracking, visual activity recognition, and distant speech recognition. The architecture of ChildBot follows a modular approach, allowing the user to easily switch on and off its modules, according to the target application, without critically affecting the functionality of the system. The system dialog is also decoupled from specific details and follows a robot-independent scheme, thus allowing new social robots to be easily added to the system. The effectiveness and successful interconnection of the modules has been demonstrated via five indicative edutainment CRI use-cases, each using a different subset of the various perception modules.

In order to validate the performance and the capabilities of our system for CRI, we have carried out an extensive, objective evaluation of the developed perception modules, as well as a user experience study that provides valuable initial insights for the interaction with ChildBot. The experiments took place in a specially designed area that was decorated to resemble a child's room. We collected both development data necessary for training the individual system modules, and use-case related data essential for testing the system performance during actual CRI.

Our results have shown that the individual perception technologies successfully capture the environment surrounding the interaction with high accuracy, while the user experience study showed that children enjoyed playing with different robots.

For future work, we would like to extend ChildBot for other applications. It would be interesting to see how some of the novel perception methods we presented can generalize to other fields, such as rehabilitation or assistive applications for ASD children. Further, we aim to conduct a more thorough subjective evaluation of the pedagogical aspect of the system.

In conclusion, our work shows that through the integration of multiple robots, sensors, and modalities, we can achieve a high level of unconstrained and autonomous CRI, opening up new prospects for future educational and entertainment social robotics.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been funded by the BabyRobot project, supported by the EU Horizon 2020 under grant 687831. The authors wish to thank Professor Christina Papaeliou and Dr. Asimenia Papoulidi for their help in designing the use-cases, supervising and evaluating the experiments with the children, and their useful remarks. We would like also to thank the "Nea Genia Ziridi" school for their collaboration in the experimental procedure. Further, we wish to thank the BabyRobot project coordinators Professor Costas Tzafestas and Professor Alexandros Potamianos, Niko Kardari for providing his original code functions, Dr. Georgia Chalvatzaki for her help during the experiments, and the members of the NTUA IRAL.

References

[1] M. Goodrich, A. Schultz, Human-robot interaction: a survey, Found. Trends Human-Comput. Interact. 1 (3) (2007) 203–275.

[2] A. Sullivan, M.U. Bers, Dancing robots: integrating art, music, and robotics in Singapore's early childhood centers, Int. J. Technol. Des. Educ. 28 (2) (2018) 325–346.

[3] T. Kanda, T. Hirano, D. Eaton, H. Ishiguro, Interactive robots as social partners and peer tutors for children: A field trial, Human–Comput. Interact. 19 (1–2) (2004) 61–84.

[4] T. Pachidis, E. Vrochidou, V. Kaburlasos, S. Kostova, M. Bonković, V. Papić, Social robotics in education: state-of-the-art and directions, in: Proc. International Conference on Robotics in Alpe-Adria Danube Region, 2018.

[5] M. Gombolay, X.J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen, J. Shah, Robotic assistance in the coordination of patient care, Int. J. Robot. Res. 37 (10) (2018) 1300–1316.

[6] W. Huo, S. Mohammed, J.C. Moreno, Y. Amirat, Lower limb wearable robots for assistance and rehabilitation: A state of the art, IEEE Syst. J. 10 (3) (2016) 1068–1081.

[7] Z. Qian, Z. Bi, Recent development of rehabilitation robots, Adv. Mech. Eng. 7 (2) (2015).

[8] S.M. Anzalone, E. Tilmont, S. Boucenna, J. Xavier, A.-L. Jouen, N. Bodeau, K. Maharatna, M. Chetouani, D. Cohen, M.S. Group, et al., How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot, Res. Autism Spectr. Disord. 8 (7) (2014) 814–826.

[9] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A.S. Rusu, D.O. David, Children with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments, Interact. Stud. 13 (3) (2012) 315–347.

[10] L. Lucignano, F. Cutugno, S. Rossi, A. Finzi, A dialogue system for multimodal human-robot interaction, in: Proc. ICMI, 2013.

[11] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, A. Waibel, Natural human-robot interaction using speech, head pose and gestures, in: Proc. IROS, 2004.

[12] NAO, Softbank Robotics, https://www.softbankrobotics.com/.

[13] Furhat Robotics, http://furhatrobotics.com.

[14] Robokind. Advanced Social Robots, http://robokind.com/.

[15] E. Gat, R.P. Bonnasso, R. Murphy, et al., On three-layer architectures, Artif. Intell. Mob. Robots 195 (1998) 210.

[16] N. Efthymiou, P. Koutras, P.P. Filntisis, G. Potamianos, P. Maragos, Multi-view fusion for action recognition in child-robot interaction, in: Proc. ICIP, 2018.

[17] J. Hadfield, P. Koutras, N. Efthymiou, G. Potamianos, C.S. Tzafestas, P. Maragos, Object assembly guidance in child-robot interaction using RGB-D based 3D tracking, in: Proc. IROS, 2018.

[18] A. Tsiami, P.P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, P. Maragos, Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults, in: Proc. ICASSP, 2018.

[19] A. Tsiami, P. Koutras, N. Efthymiou, P.P. Filntisis, G. Potamianos, P. Maragos, Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots, in: Proc. ICRA, 2018.

[20] P. Mayer, C. Beck, P. Panek, Examples of multimodal user interfaces for socially assistive robots in Ambient Assisted Living environments, in: Proc. CogInfoCom, 2012.

[21] A. Zlatintsi, I. Rodomagoulakis, V. Pitsikalis, P. Koutras, N. Kardaris, X. Papageorgiou, C. Tzafestas, P. Maragos, Social human-robot interaction for the elderly: two real-life use cases, in: Proc. HRI, 2017.

[22] D. Fischinger, P. Einramhof, K. Papoutsakis, W. Wohlkinger, P. Mayer, P. Panek, S. Hofmann, T. Koertner, A. Weiss, A. Argyros, M. Vincze, Hobbit, a care robot supporting independent living at home: First prototype and lessons learned, Robot. Auton. Syst. 75 (2016) 60–78.

[23] M. Nani, P. Caleb-Solly, S. Dogramadzi, T. Fear, H. van den Heuvel, MOBISERV: an integrated intelligent home environment for the provision of health, nutrition and mobility services to the elderly, in: Proc. 4th Companion Robotics Workshop, 2010.

[24] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, P. Maragos, A Platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands, in: Proc. ACMMM, 2016.

[25] M. V. Soler, L. Agüera-Ortiz, J. O. Rodríguez, C. M. Rebolledo, A. P. Muñoz et al., Social robots in advanced dementia, Front. Aging Neurosci. 7 (2015) 133.

[26] S. Frennert, B. Östlund, Review: Seven matters of concern of social robots and older people, Int. J. Soc. Robot. 6 (2014) 299–310.

[27] H. Robinson, B. MacDonald, N. Kerse, E. Broadbent, The psychosocial effects of a companion robot: A randomized controlled trial, J. Am. Med. Directors Assoc. 14 (2013) 661–667.

[28] M. Shishehgar, D. Kerr, J. Blake, A systematic review of research into how robotic technology can help older people, Smart Health 7 (2018) 1–18.

[29] Y.H. Wu, C. Fassert, A.S. Rigaud, Designing robots for the elderly: appearance issue and beyond, Arch. Gerontol. Geriat. 54 (2012) 121–126.

[30] J. Kennedy, P. Baxter, E. Senft, T. Belpaeme, Higher nonverbal immediacy leads to greater learning gains in child-robot tutoring interactions, in: Proc. ICSR, 2015.

[31] M. Saerbeck, T. Schut, C. Bartneck, M. Janse, Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor, in: Proc. CHI, 2010.

[32] G. Gordon, C. Breazeal, S. Engel, Can children catch curiosity from a social robot? in: Proc. HRI, 2015.

[33] B. Robins, K. Dautenhahn, R. Te Boekhorst, A. Billard, Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? Univers. Access Inf. Soc. 4 (2) (2005) 105–120.

[34] T. Belpaeme, P. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, et al., Multimodal child-robot interaction: Building social bonds, J. Human-Robot Interact. 1 (2) (2012) 33–53.

[35] J.C. Pulido, J.C. González, C. Suárez-Mejías, A. Bandera, P. Bustos, F. Fernández, Evaluating the child–robot interaction of the NAOTherapist platform in pediatric rehabilitation, Int. J. Soc. Robot. 9 (3) (2017) 343–358.

[36] F.S. Melo, A. Sardinha, D. Belo, M. Couto, M. Faria, A. Farias, H. Gambôa, C. Jesus, M. Kinarullathil, P. Lima, L. Luz, A. Mateus, I. Melo, P. Moreno, D. Osório, A. Paiva, J. Pimentel, J. Rodrigues, P. Sequeira, R. Solera-Ureña, M. Vasco, M. Veloso, R. Ventura, Project INSIDE: towards autonomous semi-unstructured human–robot social interaction in autism therapy, Artif. Intell. Med. 96 (2019) 198–216.

[37] T. Belpaeme, J. Kennedy, P. Baxter, P. Vogt, E.E. Krahmer, S. Kopp, K. Bergmann, P. Leseman, A.C. Küntay, T. Göksun, et al., L2TOR-second language tutoring using social robots, in: Proc. of the ICSR 2015 WONDER Workshop, 2015.

[38] V. Vouloutsi, M. Blancas, R. Zucca, P. Omedas, D. Reidsma, D. Davison, V. Charisi, F. Wijnen, J. van der Meij, V. Evers, et al., Towards a synthetic tutor assistant: the EASEL project and its architecture, in: Conference on Biomimetic and Biohybrid Systems, 2016.

[39] P.G. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H.L. Cao, M. Coeckelbergh, C. Costescu, D. David, A. De Beir, et al., How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder, Paladyn J. Behav. Robot. 8 (1) (2017) 18–38.

[40] E. Marinoiu, M. Zanfir, V. Olaru, C. Sminchisescu, 3d human sensing, action and emotion recognition in robot assisted therapy of children with autism, in: Proc. CVPR, 2018.

[41] S. Wallkötter, R. Stower, A. Kappas, G. Castellano, A robot by any other frame: framing and behaviour influence mind perception in virtual but not real-world environments, in: Proc. HRI, 2020.

[42] H. Yan, M.H. Ang, A.N. Poo, A survey on perception methods for human–robot interaction in social robots, Int. J. Soc. Robot. 6 (1) (2014) 85–119.

[43] A. Tapus, A. Bandera, R. Vazquez-Martin, L.V. Calderita, Perceiving the person and their interactions with the others for social robotics–a review, Pattern Recognit. Lett. 118 (2019) 3–13.

[44] A. Zaraki, M. Pieroni, D. De Rossi, D. Mazzei, R. Garofalo, L. Cominelli, M.B. Dehkordi, Design and evaluation of a unique social perception system for human–robot interaction, IEEE Trans. Cogn. Dev. Syst. 9 (4) (2017) 341–355.

[45] S. Valipour, C. Perez, M. Jagersand, Incremental learning for robot perception through HRI, in: Proc. IROS, 2017.

[46] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, T. Belpaeme, Child speech recognition in human-robot interaction: evaluations and recommendations, in: Proc. HRI, 2017.

[47] G. Yeung, A. Alwan, On the difficulties of automatic speech recognition for kindergarten-aged children, in: Proc. Interspeech, 2018.

[48] M.L. Chiang, J. Feng, W.L. Zeng, C.Y. Fang, S.W. Chen, A vision-based human action recognition system for companion robots and human interaction, in: Proc. ICCC, 2018.

[49] Y. Zhang, Y. Tian, P. Wu, D. Chen, Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder, Sensors 21 (2021) 411.

[50] Q. Wu, S. Wang, J. Cao, B. He, C. Yu, J. Zheng, Object recognition-based second language learning educational robot system for chinese preschool children, IEEE Access 7 (2019) 7301–7312.

[51] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: A review of recent research, IEEE Trans. Audio Speech Language Process. 20 (2) (2012) 356–370.

[52] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, R. Horaud, Active-speaker detection and localization with microphones and cameras embedded into a robotic head, in: Proc. Humanoid Robots, 2013.

[53] C. Evers, Y. Dorfan, S. Gannot, P. Naylor, Source tracking using moving microphone arrays for robot audition, in: Proc. ICASSP, 2017.

[54] A. Brutti, M. Omologo, P. Svaizer, C. Zieger, Classification of Acoustic Maps to determine speaker position and orientation from a distributed microphone network, in: Proc. ICASSP, 2007.

[55] H. Do, H. Silverman, Y. Yu, A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array, in: Proc. ICASSP, 2007.

[56] G. Garau, A. Dielmann, H. Bourlard, Audio-visual synchronisation for speaker diarisation, in: Proc. Interspeech, 2010.

[57] I. Gebru, C. Evers, P. Naylor, R. Horaud, Audio-visual tracking by density approximation in a sequential Bayesian filtering framework, in: Proc. HSCMA, 2017.

[58] V. Minotto, C. Jung, B. Lee, Multimodal multi-channel on-line speaker diarization using sensor fusion through SVM, IEEE Trans. Multimedia 17 (10) (2015) 1694–1705.

[59] M. Wüthrich, P. Pastor, M. Kalakrishnan, J. Bohg, S. Schaal, Probabilistic object tracking using a range camera, in: Proc. IROS, 2013.

[60] K. Murphy, S. Russell, Rao-Blackwellised particle filtering for dynamic Bayesian networks, in: Sequential Monte Carlo Methods in Practice, Springer, 2001, pp. 499–515.

[61] H. Wang, A. Klaser, C. Schmid, C. Liu, Action recognition by dense trajectories, in: Proc. CVPR, 2011.

[62] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian Conference on Image Analysis, 2003.

[63] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proc. CVPR, 2008.

[64] X. Peng, L. Wang, X. Wang, Y. Qiao, Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Comput. Vis. Image Underst. 150 (2016) 109–125.

[65] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proc. CVPR, 2010.

[66] H. Wang, M.M. Ullah, A. Kläser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proc. BMVC, 2009.

[67] F. Perronnin, J. Sánchez, T. Mensink, 2010. Improving the Fisher kernel for large-scale image classification, in: Proc. ECCV.

[68] R. Arandjelovic, A. Zisserman, All about VLAD, in: Proc. CVPR, 2013.

[69] C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27.

[70] M. Wölfel, J. McDonough, Distant Speech Recognition, John Wiley & Sons, 2009.

[71] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, P. Maragos, Room-localized spoken command recognition in multi-room, multi-microphone environments, Comput. Speech Lang. 46 (2017) 419–443.

[72] C. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, N. Hagita, A robust speech recognition system for communication robots in noisy environments, IEEE Trans. Robot. 24 (3) (2008) 759–763.

[73] V. Digalakis, D. Oikonomidis, D. Pratsolis, N. Tsourakis, C. Vosnidis, N. Chatzichrisafis, V. Diakoloukas, Large vocabulary continuous speech recognition in Greek: Corpus and an automatic dictation system, in: Proc. Interspeech, 2003.

[74] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, P. Woodland, The HTK Book Version 3.4, Cambridge University Press, 2006.

[75] G. Skantze, S. Al Moubayed, IrisTK: a statechart-based toolkit for multi-party face-to-face interaction, in: Proc. ICMI, 2012.

[76] D. Harel, Statecharts: A visual formalism for complex systems, Sci. Comput. Progr. 8 (3) (1987) 231–274.

[77] BabyRobot project http://babyrobot.eu.

[78] C.Y. Ren, V. Prisacariu, O. Kaehler, I. Reid, D. Murray, 3D tracking of multiple objects with identical appearance using RGB-D input, in: Proc. International Conference on 3D Vision, 2014.

[79] L. Hall, C. Hume, S. Tazzyman, Five degrees of happiness: Effective smiley face Likert scales for evaluating with children, in: Proc. 15th International Conference on Interaction Design and Children, 2016.

Niki Efthymiou is a Ph.D. student at the school of Electrical and Computer Engineering, National Technical University of Athens (NTUA), under the supervision of Prof. Petros Maragos. She is working primarily in computer vision problems associated during Human–Robot Interaction. She is a researcher at the Computer Vision, Speech Communication and Signal Processing Group at NTUA and her research interests lie in the fields of gesture, action and emotion recognition, with a focus on Child–Robot Interaction. She received her diploma degree in Applied Mathematics and Masters degree in Computational Mechanics from NTUA.

Panagiotis P. Filntisis was born in Athens in 1989. He received his Diploma Degree in Electrical Engineering and Computer Science from the National Technical University of Athens in October 2015. Since November 2015 he pursues a Ph.D. in the CVSP Group, school of ECE, NTUA, under the supervision of Prof. Petros Maragos. His research interests lie in the fields of affective computing and multimodal signal processing and their application in the domain of human and child robot interaction.

Petros Koutras received his Ph.D. from the National Technical University of Athens in 2019 and is now a computer vision engineer in Snap Inc. He also received the Diploma degree in E.C.E from NTUA in 2012. Since 2013 he had been a researcher at the CVSP Group at NTUA working in several EU and Greek projects. His research interests lie primary in the fields of machine learning for video understanding and computer vision for robotics. His research contributions include the development of multitask networks for automatic video understanding and the integration of multisensory and multimodal robotic perception systems in robotics applications.

Dr. Antigoni Tsiami received her M.Eng. Diploma degree in ECE in 2012 and her Ph.D. degree in 2019, both from NTUA. Her research interests and expertise include audiovisual saliency estimation and attention modeling, machine learning, speech recognition, multichannel speech processing and dialog systems. Since 2013, she is a research assistant at the Robotic Perception and Interaction Unit of the Athena Research Center and at the NTUA Computer Vision, Speech Communication & Signal Processing Group. Since 2020 she is a PostDoc Researcher at the above groups.

Jack Hadfield received his Diploma Degree in Electrical Engineering and Computer Science from the National Technical University of Athens. He was a researcher at the Computer Vision, Speech Communication and Signal Processing Group at NTUA.

Gerasimos Potamianos received his Ph.D. in ECE from the Johns Hopkins University, in 1994. He is currently an Associate Professor at the ECE Dept. at the University of Thessaly, in Volos, Greece and collaborating member of Athena Research and Innovation Center in Athens, Greece. His research interests span the fields of audio-visual speech processing, automatic speech recognition, sign language recognition, multimedia signal processing and fusion, as well as multimodal scene analysis. He has published over 140 articles in these areas that have received over 5.7k citations and holds 7 U.S. Patents. He is a member of IEEE, EURASIP, and ISCA.

Petros Maragos is a full Professor of ECE at NTUA, Director of the Intelligent Robotics and Automation Lab and the CVSP Group. He has also worked as professor at USA universities, including Harvard University (1985–93) and Georgia Tech (1993–98). His teaching and research interests include signal processing and machine learning, computer vision, speech/language, and robotics. He has been the PI of several USA, European and Greek research projects. He is a Fellow of IEEE and EURASIP.