Multi3: Multi-sensory Perception System for Multi-modal Child Interaction with Multiple Robots

Antigoni Tsiami[1,2], Petros Koutras[1,2], Niki Efthymiou[1,2], Panagiotis Paraskevas Filntisis[1,2], Gerasimos Potamianos[1,3], Petros Maragos[1,2]

***Abstract*— Child-robot interaction is an interdisciplinary research area that has been attracting growing interest, primarily focusing on edutainment applications. A crucial factor to the successful deployment and wide adoption of such applications remains the robust perception of the child's multi-modal actions, when interacting with the robot in a natural and untethered fashion. Since robotic sensory and perception capabilities are platform-dependent and most often rather limited, we propose a multiple Kinect-based system to perceive the child-robot interaction scene that is robot-independent and suitable for indoors interaction scenarios. The audio-visual input from the Kinect sensors is fed into speech, gesture, and action recognition modules, appropriately developed in this paper to address the challenging nature of child-robot interaction. For this purpose, data from multiple children are collected and used for module training or adaptation. Further, information from the multiple sensors is fused to enhance module performance. The perception system is integrated in a modular multi-robot architecture demonstrating its flexibility and scalability with different robotic platforms. The whole system, called Multi3, is evaluated, both objectively at the module level and subjectively in its entirety, under appropriate child-robot interaction scenarios containing several carefully designed games between children and robots.**

I. INTRODUCTION

Human-robot interaction (HRI) has drawn research interest during the last few years, mostly due to the growing intrusion of social robots in everyday life [11]. Although significant progress has been made that further developed use cases like assisted living [4], [33], the need for more intuitive, natural communication and better human behavior tracking has raised new challenges. Humans communicate mostly with the exchange of audio-visual information, thus gesture and speech recognition are in the spotlight of HRI research.

At the same time, social robots are increasingly entering our lives for entertainment and educational purposes [16], [27], especially childrens' [3], [24], thus creating new challenges. Children are very adaptive, quick learners and familiarized with new technologies. They have unique communication skills, as they can easily convey or share complex information with little spoken language. They constitute a perfect target group for studying the effect of HRI on the development and enhancement of communicative abilities [12]. Thus, child-robot interaction (CRI) is an emerging research field that at the same time presents new challenges, e.g., action and speech recognition for children [10], [17], because the majority of perception systems are designed and developed for adults. However, children behave differently than adults, and action and speech recognition models trained on adults usually do not perform sufficiently well in the case of children. Thus children data, though difficult and time-consuming to collect, are essential for building children-specific systems.

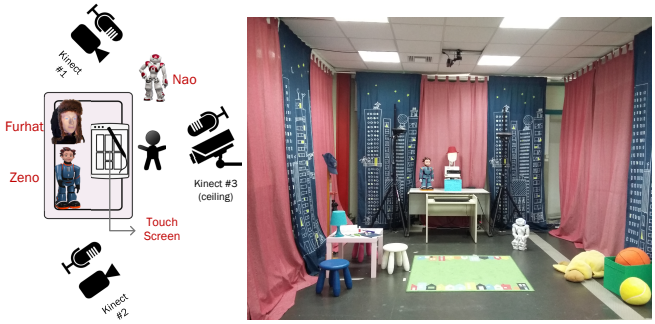

Fig. 1: Left: Diagram of the spatial arrangement of sensors and robots in the Multi3 system. Right: Picture of the setup.

The development of technologies like action, gesture, and speech recognition, in parallel with the development of new sensors, like Kinect cameras, microphone arrays, etc., has launched a new era of multi-modal and multi-sensory systems that can increase robustness and overall performance [15]. More and more often, traditionally uni-modal systems incorporate additional modalities in order to improve performance via fusion [2], [21]. However, the perception capabilities are platform-dependent and rather limited, since audio-visual sensors are often embedded on robotic platforms. In order to tackle this problem, recent HRI works have employed external sensors mounted in a space where the interaction takes place [22], [9], without however evaluating the perception results from combining and fusing the different sensors.

In this paper we propose a multi-robot, multi-modal, and multi-sensory system for robot perception, specifically developed for CRI, which we call the "Multi3" system. The child can interact with the social robots naturally, using body actions, gestures, and speech. The system employs multiple Kinects, using both their visual sensors and microphone arrays to unobtrusively capture from the far-field the child's activity, allowing natural, untethered CRI. The contributions

*This work was supported by the EU Horizon 2020 project Babyrobot, under grant 687831.

[1]Athena Research and Innovation Center, Maroussi 15125, Greece
[2]School of ECE, National Technical Univ. of Athens, 15773 Athens, Greece {antsiami,pkoutras,maragos}@cs.ntua.gr, {nefthymiou,filby}@central.ntua.gr
[3]Dept. of ECE, University of Thessaly, 38221 Volos, Greece gpotam@ieee.org

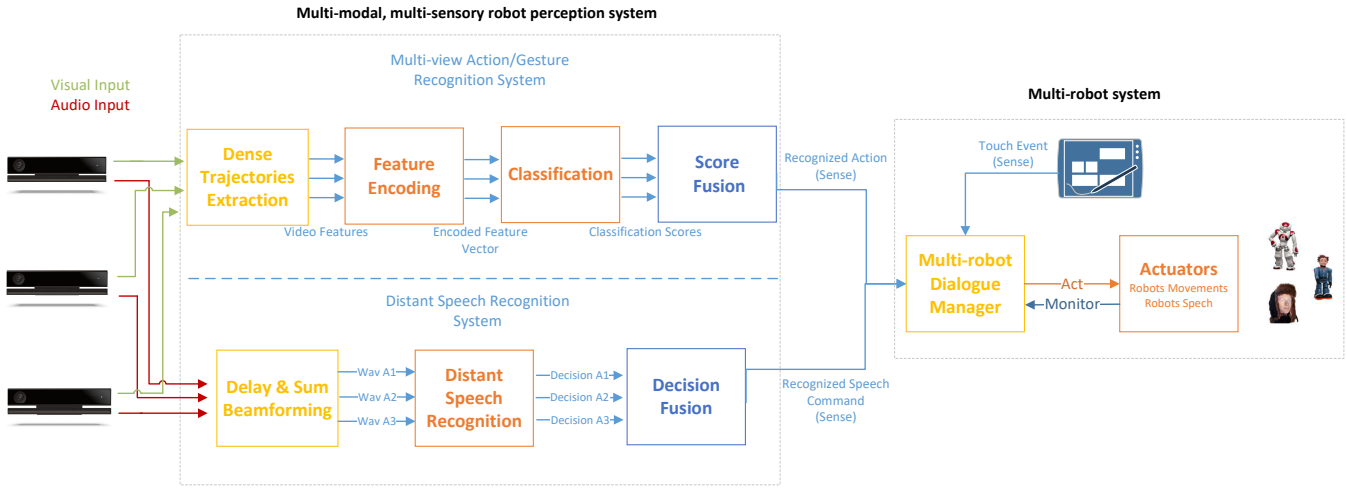

Fig. 2: Multi-modal, multi-sensory robot perception and multi-robot system architecture. "A" refers to the Kinect microphone array.

of the paper can be summarized as follows:

- The development of a multi-modal (action, gesture, speech), multi-sensory (multiple Kinects) robot perception system for child-robot interaction, the Multi3 system, presented in detail in Section II.
- The development of a multi-robot modular cross-platform architecture in ROS (Robot Operating System [1]) that handles the dialog and component communication, described in Section III.
- The design of a use case scenario for CRI, described in Section IV, that involves three games between the child and robots and aims both at engaging the child but also testing and showcasing the Multi3 system.
- Children data collection for model training/adaptation and evaluation, both of the Multi3 system in an objective way, but also of the CRI use case in a subjective way. Results can be found in Section V.

II. Multi-modal, Multi-sensory Robot Perception

As described above, our setup aims at allowing multiple CRI modalities, namely visual actions and speech, necessitating the use of multiple sensors for video and audio capture. During CRI, occlusions and pose variations are expected, thus the use of a single camera is not adequate. The same applies also for audio: Children may be far from the microphones as they play or move. For this reason, we employ distributed cameras and microphone arrays aspiring to monitor as much of the playing area as possible. The spatial arrangement of the sensors and the system in place along with the three robots currently integrated in the system (which are described in Section IV-A) can be seen in Fig. 1. The architecture of our multi-sensor system uses three Kinect V2 sensors both for video and audio capture. Each one includes: A full high definition color camera (with a 30FPS framerate), a depth sensor using a Time-of-Flight method, and four microphones. Two of the Kinect sensors are positioned sideways and symmetrically with respect to the central interaction area, and the third is mounted on the ceiling facing the floor, in order to get the floor plan view.

The multi-modal, multi-sensory robot perception system, which is depicted in Fig. 2, has been designed and trained/adapted specifically for children. The upper part of the perception system refers to action and gesture recognition, while the lower part to distant speech recognition. Results from the perception modules are forwarded to the multi-robot architecture described in Section III.

A. Multi-view Action and Gesture Recognition

Our multi-sensory action and gesture frontend employs Dense Trajectories features along with the popular Bag-of-Visual-Words (BoVW) framework. The Dense Trajectories method [29] has received attention due to its superior performance on challenging datasets. Its main concept consists of sampling feature points from each RGB frame on a regular grid and tracking them through time based on optical flow. Tracking is performed in multiple spatial scales, and trajectories are pruned to a fixed length L to avoid drifting.

Hence, the algorithm for each sampled feature point with index n results in a trajectory $\mathbf{x}_n = (P_1, P_2, \ldots, P_L)$, which is a sequence of points P_1,..,P_L in consecutive frames, beginning at frame $t_b(\mathbf{x}_n)$ and ending at $t_e(\mathbf{x}_n)$ (see also Fig. 3). Following the trajectory extraction, different visual features are computed within space-time volumes along each trajectory. More specifically, the following are used: the Trajectory descriptor [29], Histograms of Oriented Gradients (HOG) [19], Histograms of Optical Flow (HOF) [19], and Motion Boundary Histograms (MBH) [29] computed on both axes (MBHx, MBHy). The trajectory descriptor encodes the shape of the trajectories, while histographic features describe the local shape, appearance, and motion along each trajectory.

In more detail, the Trajectory descriptor consists of the sequence of displacement vectors between consecutive trajectory points, $\Delta P_l = P_{l+1} - P_l$, normalized by the sum of the displacement vector magnitudes over the trajectory. The HOG descriptor can model the local static appearance based on the orientation and magnitude of the image intensity gradient. HOF captures motion information using the orientation and magnitude of the optical flow. Finally, MBHx/MBHy are

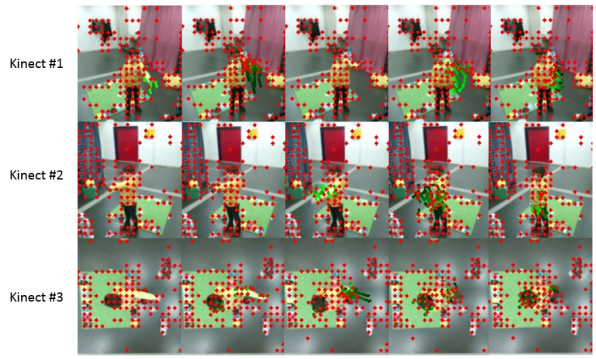

Fig. 3: An example of Dense Trajectories for a recording of the "Show me the Gesture" game (discussed in Section IV-B). Each row corresponds to a different Kinect camera view.

computed on the gradient of the horizontal/vertical optical flow components, and MBH is their concatenation, being more robust to camera motion.

A human detector may optionally be used to improve the representation of foreground action. For this purpose we have employed a simple person detector based on HOG descriptors [7]. This way, feature extraction time can also be reduced, since the Dense Trajectories are computed in a small region of the full HD Kinect frame.

For encoding purposes, codebooks for each descriptor (Trajectory, HOG, HOF, MBHx, MBHy) are constructed during the training phase from a subset of randomly selected training features using K-means. The centroid of each cluster is defined as a visual word, and each trajectory is assigned to its closest visual word using the Euclidean distance. We use BoVW encoding, i.e., a histogram of visual word occurrences, yielding a sparse K-dimensional video representation, which is essentially the histogram of visual word occurrence frequencies over the space-time volume. Videos are classified based on their BoVW representation, using non-linear support vector machines (SVMs) with the χ^2 kernel [30]. In addition, different descriptors are combined, by computing distances between their corresponding BoVW histograms as:

$$Q(\mathbf{h}_i, \mathbf{h}_j) = \sum_m \exp\left(-\frac{1}{A_m} D\left(\mathbf{h}_i^m, \mathbf{h}_j^m\right)\right), \qquad (1)$$

where $\mathbf{h}_i^m$ denotes the BoVW representation of the m-th descriptor of the i-th video, and A_m is the mean value of χ^2 distances $D(\mathbf{h}_i^m, \mathbf{h}_j^m)$ between all pairs of training samples. Since we face multiclass classification problems, we follow the one-against-all approach and select the class with the highest score.

For a given Kinect sensor we have trained a different SVM for all employed classes and obtain the probabilities as described in [6]. Then we apply a soft-max normalization to each sensor's probabilities. For the fusion of the three sensor output probabilities we have experimented with three widely used functions: min, max, mean, which correspond to different approaches in integrating decision scores for each class across sensors. With max we decide based on the most confident sensor, min selects the class with high scores for all sensors, whereas mean lies somewhere in the middle. Finally, we select the class with the highest fused score, following the same approach as in the single-sensor case.

B. Multi-microphone Distant Speech Recognition

Incorporating a speech module in a HRI system is essential for a natural communication between humans and the robot, let alone children. At the same time it presents challenges, such as noise, reverberation, and distance between the robot and the speaker [14]. Our CRI use case necessitates child's movement in the space in order to interact and play with the robots. For this reason we employ distant speech recognition (DSR) [31], [25] with three microphone arrays (Kinect) distributed in space. Children communicate with the robot via a set of utterances adopted for the context of the specific use case (see Section IV). A continuous speech recognition system would require a large amount of data to train/adapt for children in order to perform well, and at the same time children do not speak continuously. Thus our speech recognition module is grammar-based.

The DSR module is always-listening, namely it is able to detect and recognize the utterances spoken by the user at any time, among other speech and non-speech events, possibly degraded by environmental noise and reverberation. The specific set of utterances contains the child's possible answers in some games of the use case scenarios, as well as a few general utterances. More details can be found in Section V. The employed language is Greek. We target robustness via denoising of the far-field signals (beamforming) and adaptation of the acoustic models.

To detect one of the target utterances, we use a sliding window of 2.5 *sec* duration with a 0.6 *sec* shift, in which we enforce recognition of one of the pre-defined sentences against other garbage phrases.To improve the quality of the noisy and reverberant far-field children speech, we employ a simple delay-and-sum beamforming on each available 4-microphone Kinect array. The main idea is the insertion of delays to the different microphone signals a_n to align them in order to enhance speech coming from a specific direction. Thus, for uniform linear arrays with N microphones, if the desired direction is denoted by ϕ, the time-delay to be applied to each microphone is

$$\tau_n = \frac{(n-1)d\cos\phi}{c}, \qquad (2)$$

where c is the speed of sound and d the space between microphones. The beamformed signal then results as:

$$y(t) = \frac{1}{N}\sum_{n=1}^{N} a_n(t-\tau_n) \qquad (3)$$

Regarding acoustic modeling, GMM-HMM cross-word tri-phone models based on a standard MFCC-plus-derivatives features have been trained on the Logotypografia database [8]. The corpus was artificially distorted by convolving its clean speech with room impulse responses and adding white Gaussian noise in order to match the far-field condition [25]. Maximum likelihood linear regression

(MLLR) adaptation was employed to transform the means of the HMM state observation Gaussians, aiming to reduce the mismatch between the initial model and the adaptation data [32].

In order to make the DSR module more robust and exploit the distributed microphone arrays, we fuse the decisions of the three different microphone arrays, employing a simple majority voting for the best hypothesis results. In case all three microphone arrays disagree, the second best hypothesis results are taken into consideration.

III. Multi-robot System Architecture

In this section the dialog management and the intra-communication between the system modules are presented, as well as the integration of all components under a unified system that performs real-time.

A. Event-driven Communication and Dialog Management

Dialog management and communication between modules constitutes the backbone of our CRI system and raises several challenges when multiple robotic agents are considered. Here, we use the IrisTK [28] framework and extend it for a multi-robot system. The system follows a completely modular architecture, and communication between its different modules takes place through events. Events are divided into three categories (following the IrisTK framework):

- *Action*: i.e., things the multi-robot system should do
- *Sense*: i.e., what the system perceives from its surroundings
- *Monitor*: i.e., the feedback to the actions executed by the system

Similarly, system modules can be divided into sensors (gesture recognition, speech recognition, etc.) and actuators (robot speech, robot movement, etc.). The system core module is dialog management that handles events. Sensors analyze the environment continuously and send sense events to the dialog module, while actuators receive action events from the dialog module and send back feedback (e.g., when they finish an action). Fig. 2 presents this flow of events.

Dialog follows a variation of the Harel statechart [13] for event-driven dialog systems. Dialog states can be hierarchically structured, and each state can contain a number of parameters that modify its execution and greatly reduce the required number of states.

A naive approach towards a multi-robot system with three robotic platforms would be to create three different states, one for each robot. Using parameters in states leads us towards a robot agnostic approach, where the acting robot is another parameter. In our approach we use "action states" that act as an intermediate layer between core dialog and actions and are responsible for sending the current event to the correct robot according to the passed parameter. Dialog does not move into "action states", but instead these states are "called" [28] from the current dialog state, and upon finishing their execution dialog returns to the current state. Examples of "action states" are a "speak action state" with the text to be synthesized and the robotic agent to speak as parameters, or an "expression action state" with the name of the expression (e.g., smile) and the robotic agent to form it as parameters.

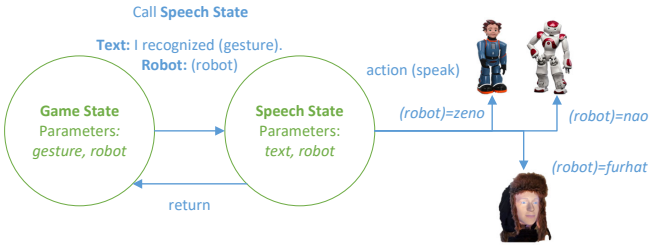

Fig. 4: Usage of "action states" and parameters in the dialog for announcing the recognized gesture in a gesture recognition scenario of a game.

Fig. 4 depicts the aforementioned examples. The state of a game that involves a gesture recognition scenario calls for an agent to speak, defined through a parameter chosen in some previous state called "robot". This parameter is not used by the game state but by the "speak action state" that sends a speech action event to the correct robot. The same figure also depicts that, when using parameters, only one state is needed for announcing the gesture.

Using this model, we can keep the core dialog flow decoupled from specific robot details, and adding a new robot to the system amounts to just adding its action events to the "action states" and handling the event on the robot side. Of course, there are still cases in which the capabilities of the robot (e.g., if it can move or not) cause variations in the scenario and have to be treated separately in the dialog.

B. Online System Integration

For the technical integration of sensors and robots we have designed a hybrid system that uses both the Windows and Linux operating systems and can handle multiple sensors in modular configurations. We use the IrisTK framework running in Windows (as mentioned in Section III-A) for real-time dialog management and event handling, while all required data processing takes place in three different Linux machines using ROS. Events from all modules are sent to a broker that handles their allocation to the suitable modules.

The three Kinect V2 sensors are connected to three different Linux machines and provide raw data of the sensors (i.e., color and depth images and raw audio). The touch screen is connected to a Windows machine and sends continuous feedback to the dialog module about the choices of the children, while also receiving and reacting to events sent by the dialog module that change its state automatically (e.g., the current game played).

IV. Child-Robot Interaction Use Case

In order to develop and evaluate the whole system, a specific use case that involves three games between children and robots has been designed. These three games, which focus on both verbal and non-verbal communication, are the following: a) "Show me the gesture", b) "Express the feeling", and c) "Pantomime". During these games, the child interacts with three different robotic agents (see Fig. 5). In

(a) Furhat (b) Nao (c) Zeno

Fig. 5: Robotic agents employed in the use case scenario.

(a) Show me the gesture (b) Express the feeling

(c) Pantomime

Fig. 6: Indicative frames from each game included in the use case scenario.

the following subsections, the robotic platforms and the CRI games for this specific scenario are described.

A. Robotic Platforms

The three robotic agents of this specific scenario are shown in Fig. 5: a Furhat head, a Nao, and a Zeno robot. We have enabled Furhat and Nao to speak in the Greek language through integration with a Greek text-to-speech engine [5]. The Furhat robot head has been created by Furhat Robotics, and it is a robotic head in which an animated face is back-projected on a three-dimensional mask. Furhat among other things is capable of speech, head movement with 2 degreees of freedom, and facial expressions. The Nao robot has been created by SoftBank Robotics and is a humanoid robot. It carries multiple sensors and is capable of speech. The Zeno robot has been developed by Robokind, has 21 degrees of freedom, with 7 degrees on the face. Zeno's ability to make a wide range of facial expressions improves CRI. Zeno is able of speech and is also equipped with multiple sensors.

All the above mentioned robots have been used for interaction with children. For instance, Nao has taken the role of kids' tutor in numerous applications, e.g., dance or quiz scenarios [26]. Zeno has been used in schools of Denmark as teachers' assistant [18]. Furhat has interacted with children and adults in public spaces, like museums [20].

Regarding the integration of the robots in our system, the Furhat robot head is already integrated within the IrisTK framework. For the Nao and Zeno robots, intermediate layers were developed, that receive events from the dialog module and convert them to the corresponding actions for each robot.

B. Child-Robot Games

As mentioned earlier, three games have been designed to prompt children to interact with robots using not only their voice, but also their facial expressions and body movements (see also Fig. 6). Thus, the kids perceive that they can play with robots without many limitations and act as they do while playing with their peers. These tasks are suitable for children users of age from six to ten years old, as they neither assume special knowledge nor restrict kids' spontaneity.

In the "Show me the gesture" task, the robotic agent prompts the child to form sequentially some gestures that denote a meaning. The gesture meanings are depicted on a screen and signify: an agreement, a greeting, a call to the robot for coming closer, the drawing of a circle in the air, to point towards something, to ask the robot to stop, and to ask the robot to sit down. Since the children are only provided with the meaning of the gesture and not the actual way to perform it, each child performs the gesture differently. When a gesture is performed, it is automatically recognized by the system, and the robotic agent asks the kid to confirm whether or not the recognition is correct. If the recognition is not correct, the child is asked to repeat the previous gesture, otherwise performs another gesture, and so on. Along with speaking, the agent often reacts to the gestures with a movement (if it is capable of moving), e.g., it walks to come closer to the child or waves back. With this task, apart from the child-robot interaction, we evaluate the gesture recognition module of our perception system. All of the robotic agents can take part in this game.

The "Express the feeling" game focuses on child-robot emotional interaction with emphasis on facial expressions. The child walks up to a touch screen that is centrally located and is presented with six hidden cards. When the kid chooses a card, an image that depicts a basic emotion appears. The robotic agent then asks the child to express this particular emotion, and subsequently the robot expresses the emotion as well. The six basic emotions used in this game are: happiness, anger, sadness, disgust, surprise, and fear. This task enriches the interplay between the child and the robot as it introduces the meaning of emotions. Through expression of emotions, the kid and the robot form a non-verbal way of communication and enhance their interaction. The Zeno robot participates in this game due to its capability of forming very realistic facial expressions. In this game we do not evaluate any of the perception components, but we intend to study the child-robot interaction while they both express facial emotions and validate the scalability of cross-platform robot architecture when more tasks and robots are included.

The game of "Pantomime" allows the child to interact with the robot in a different way, as the robot and child interchange roles. Twelve cards are presented on the screen, each one depicting a type of manual work. The robotic agent chooses one randomly and mimes it, while the kid must figure out and name which one is being performed. If the child's choice is correct, the robot compliments him/her, otherwise it allows the child to decide if he/she wants the

mimed task to be revealed or retry guessing. After the first round, the robot and the child interchange roles. The child this time chooses a card and tries to mime it while the robot tries to recognize it. After recognizing it, the robot asks the child to verify its recognition result. The twelve cards used in the "Pantomime" game are: painting a wall, cleaning a window, driving a bus, swimming, dancing, working out, playing the guitar, digging a hole, wiping the floor, ironing a shirt, hammering a nail, and reading a book. In this particular task, automatic speech recognition reinforces the naturalness of child-robot interaction since the child says his/her choice as he/she would act in a pantomime game with other children. In this challenging game both action and speech recognition subsystems are evaluated while the child interacts multi-modally with the Nao robot.

V. Evaluation

In order to evaluate the whole system we have employed two evaluation strategies: In the first we perform objective offline evaluation of the two recognition modules, namely the action/gesture and speech recognition ones, using the children dataset collected for the needs of system training and evaluation. In the second one, we subjectively evaluate the Multi3 system, by using a questionnaire that was filled out by children after their interaction with the system.

A. Children Data Collection and Objective Evaluation of Perception Modules

As noted in the previous sections, children data are required for a dedicated CRI system to perform robustly. Regarding visual data, children act, gesture, and express themselves in a different way than adults, as they are more imaginative and their behavior has not been standardised yet. Concerning audio data, children voice characteristics differ from adults, e.g., the pitch of the voice. Although several human action and speech databases exist with adult data, they are not suitable for training/adapting a perception system dedicated to children. At the same time, children data are hard and time-consuming to obtain.

Therefore, for each game we performed an extensive data collection featuring 28 children participants. During data collection, each child performed sequentially all seven gestures mentioned in the "Show me the gesture" game, expressed the six feelings from the "Express the feeling" task, mimed the twelve tasks of the "Pantomime" game, and also performed random movements for the background models. In addition, each child uttered 40 out of 120 phrases inspired by and adapted to the use case scenario. These consist of game-dependent utterances like pantomime answers with variations in pronunciation, e.g. "Are you dancing?" and "I think you are dancing", and some general purpose utterances, e.g. "yes","no".

To evaluate the gesture recognition system we employ the annotated multi-sensory data that was collected by the three Kinect sensors, depicting the 28 children of the data collection performing gestures in different positions, as well as background data with random movement. Using these data we have trained one individual model for each Kinect following a leave-one-out cross-validation approach.

	Single Camera			Fusion		
Feat.	Kin. #1	Kin. #2	Kin. #3	mean	min	max
Traj.	68.75	66.90	65.74	76.62	75.00	71.53
HOG	40.74	33.33	29.40	39.58	36.57	39.58
HOF	70.83	70.37	69.21	78.01	77.55	76.39
MBH	76.85	67.82	68.29	83.80	80.09	78.24
Comb.	77.78	73.84	73.61	81.94	**83.56**	77.55

TABLE I: Average classification accuracy (%) for the children gesture recognition task. Results for the five different features for both single and multi-stream cases are shown.

	Single Camera			Fusion		
Feat.	Kin. #1	Kin. #2	Kin. #3	mean	min	max
Traj.	63.08	48.62	45.54	64.00	61.23	62.15
HOG	39.69	32.00	27.69	43.38	35.38	41.85
HOF	68.31	56.31	48.62	68.31	65.54	68.92
MBH	70.77	60.92	61.85	74.46	73.54	72.31
Comb.	73.85	63.38	60.00	**74.46**	**74.46**	73.85

TABLE II: Average classification accuracy (%) for the children action recognition task. Results for the five different features for both single and multi-stream cases are shown.

	No-adapt		Adapt-all		Adapt-per-array	
	WCOR	SCOR	WCOR	SCOR	WCOR	SCOR
Kinect #1	79.30	70.53	98.41	95.95	98.30	95.95
Kinect #2	81.04	72.48	97.56	95.95	97.35	95.95
Kinect #3	76.85	66.83	97.45	94.60	97.56	94.60
Fusion	-	65.02	-	**97.05**	-	96.30

TABLE III: Average word (WCOR) and sentence accuracy (SCOR) (%) for the children DSR task. Results for each Kinect array separately and their decision fusion for all adaptation schemes are depicted.

Table I presents average accuracy results (%) for the 7 gestures. Results indicate that the combination of the proposed features (denoted as Comb. in the Table) performs slightly better for both single-sensor and multi-sensor cases. Additionally, the motion related descriptors (HOF, MBH) yield better results than the static (HOG) one, since all employed gestures include motion of the whole arm rather than formation of the handshape. Regarding the fusion of the different streams, recognition performance seems to improve significantly compared to the best single-stream result. Top performance is achieved using the min fusion scheme.

In order to verify the appropriateness of the proposed action and gesture recognition system in more challenging tasks, we evaluate the pantomime actions described previously. These actions are more complex than gestures, since children perform them in many different ways. In Table II we present the average accuracy results (%) for the 12 gestures as well as the background model. Despite the difficulty of the task, the system performs quite well with fusion yielding around 74% accuracy for a 13-class problem.

In order to objectively evaluate the DSR task offline, the collected data have been used both for adapting speech models and for testing. Results are presented in Table III in terms of word and sentence accuracies, denoted by WCOR and SCOR respectively. The first three rows depict WCORs and SCORs for each Kinect separately, and the last one presents the decision fusion results.

Three different adaptation schemes have been tested for comparison: In the "No-adapt" case, the employed models

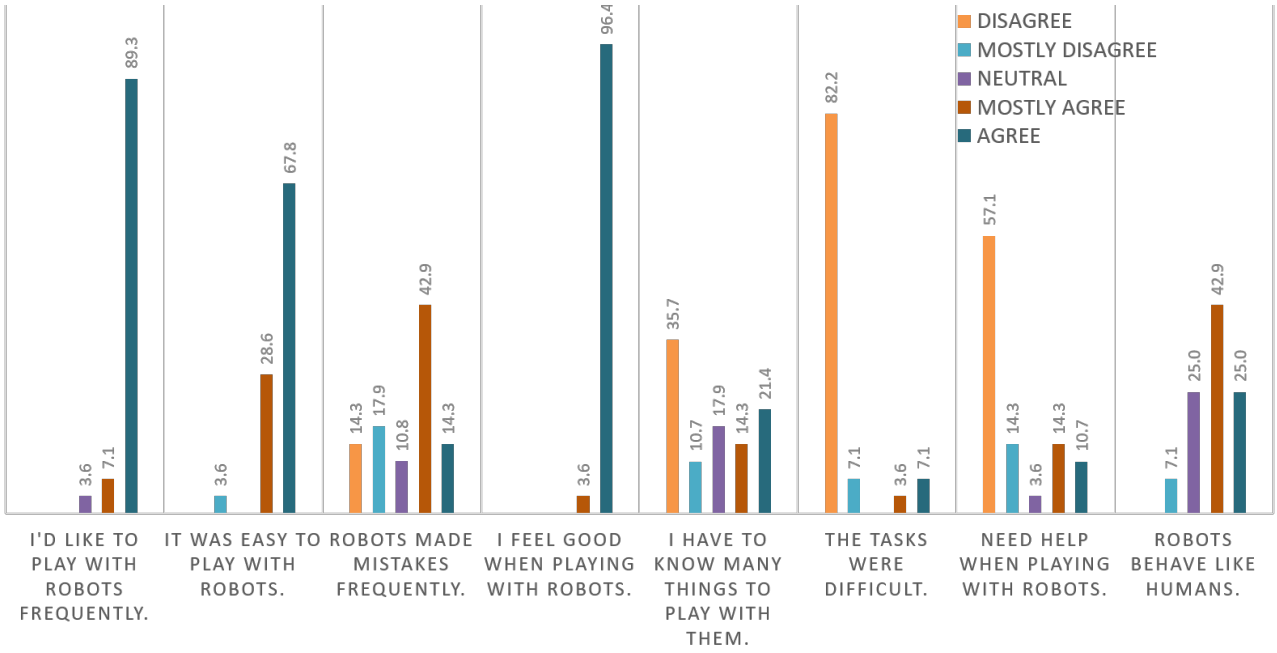

Fig. 7: Subjective evaluation by 28 children. The chart depicts the 8 different questions and the children responses (%) in a five-point ordinal scale.

have been trained on the Logotypografia database that contains adult data. The available children data have been used for testing. In the "Adapt-all" case, data from 20 out of 28 participants have been used in order to adapt speech models globally, i.e. data from all three Kinect arrays have been used to adapt a single model. The remaining 8 participants form the test set. The adaptation and testing has been 4-fold cross-validated. In addition we experimented with adaptation per array, by adapting one model per Kinect with data only from the specific Kinect array. The same data partitioning has been used, also 4-fold cross-validated.

Although speech recognition results are not very satisfactory for the "no-adapt" case, the system achieves good performance when children data are used to adapt the models, which underlines the importance of collecting children data. Adaptation per array appears unnecessary as it yields almost the same results with global adaptation. Performance is equally good for all Kinect sensors, with the third Kinect performing slightly worse. Fusion was also performed at decision level. For the no-adapt case, fusion performs worse than each Kinect individually, because it is based on unreliable speech recognition results, while for adapted models it achieves the best performance with an SCOR of 97.05%.

B. Subjective Evaluation of the CRI Use Case

Aiming to evaluate the use case scenario, we tested the Multi3 system real-time. For this purpose, 28 children (18 male, 10 female) from six to ten years old (average: eight years old), were invited to interact with the system. The children and their parents volunteered to participate in the experimental procedure when they met our team in dissemination events. The interaction took place in an appropriately designed friendly environment, decorated to remind a child's room. The experimental procedure involved one child each time and lasted for approximately fifteen minutes. The child entered the setup room accompanied by his/her parent(s) and a member of our team. Subsequently, the child was informed about the process, while getting familiar with the room and the robotic agents. The structure of the procedure and the rules of the games were explained to the child, and when he/she felt comfortable, the interplay started without any human intervention, with the robotic agents introducing themselves to the child. The interaction then continued with the scenario that involved the three games presented before.

Subjective evaluation of the Multi3 system took place after the completion of the child-robot interplay by asking the children to fill out an eight-question form. Each question depicted a statement, and the children were instructed to grade their agreement to it using a 5-point ordinal scale from "disagree" to "agree". Children were also asked to pick their favorite game and justify their answer. Results for all questions are depicted in Fig. 7. "Pantomime" was voted the most favorite game, because of the robot movement and speech. Also, 22 of 28 children stated that they enjoyed playing with the robots because the agents had cognition and perception of both their movements and their utterances, while the rest of them because robots perceived their movements only (no child expressed an "only utterances" preference).

As far as the graded questions are concerned, the majority of the children enjoyed interacting with the robots and would like to continue playing with them. Almost every child said that it wasn't difficult to play with the robots, while four out of five children found the tasks easy. More than half of them believed that they didn't need help to play with the robots. Regarding required prior knowledge, a large variance in their answers can be observed, which is related to the large variance in children ages. The majority of disagreement with the need for prior knowledge was observed in the age group of 9-10 years old. One quarter of the participants agreed that the robots behaved like humans, almost 45% of the

participants mostly agreed with this phrase, while another quarter of them were neutral. We should of course note the caveat of the children "ceiling effect" [23], when interpreting the above results.

Children responses during the subjective evaluation justify our choice to create a multi-modal perception system able to recognize utterances, gestures, and pantomime actions. The Multi3 system appears to be suitable for children of ages from six to ten years old, and the designed tasks achieved to both demonstrate the capabilities of the system, but also to be relatively easy and not boring for the children to accomplish.

Nevertheless, further improvements to the perception system and the use case scenario should be made in order to enable a more human-like and spontaneous interaction between children and robots. For example, the number of speech utterances and gestures should be increased, thus improving user experience and naturaleness of the interaction.

VI. CONCLUSIONS

We have proposed and presented Multi3, a multi-robot, multi-modal, and multi-sensory system for child-robot interaction. The contributions of this work are multi-faceted, spanning mainly the area of robotic perception where an action and speech recognition system specifically developed for child-robot interaction has been proposed. Moreover, a modular robot architecture able to handle multiple robots, as well as a carefully designed scenario targeted to children incorporating various engaging games, are additional contributions. System evaluation has been carried out using children data collected according to the proposed use case, with objective and subjective evaluation results confirming the success of our system both in terms of performance and user acceptability.

REFERENCES

[1] Robot Operating System (ROS). http://www.ros.org/about-ros/.

[2] I. Bayer and T. Silbermann. A multi modal approach to gesture recognition from audio and video data. In *Proc. ICMI*, 2013.

[3] 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. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[4] E. Broadbent, R. Stafford, and B. MacDonald. Acceptance of healthcare robots for the older population: review and future directions. *Intl. J. Social Robotics*, 1(4):319, 2009.

[5] A. Chalamandaris, P. Tsiakoulis, S. Karabetsos, and S. Raptis. The ILSP/INNOETICS text-to-speech system for the Blizzard challenge 2013. In *Proc. Blizzard Challenge Workshop*, 2013.

[6] C. Chang and C. Lin. LIBSVM: a library for support vector machines. *ACM Trans. Intelligent Systems and Technology*, 2(3):1–27, 2011.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

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

[9] 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, Journal of Behavioral Robotics*, 8(1):18–38, 2017.

[10] S. Fernando, R. Moore, D. Cameron, E. Collins, A. Millings, A. Sharkey, and T. Prescott. Automatic recognition of child speech for robotic applications in noisy environments. *CoRR arXiv:1611.02695*, 2016.

[11] M. Goodrich and A. Schultz. Human-robot interaction: a survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275, 2007.

[12] G. Gordon, C. Breazeal, and S. Engel. Can children catch curiosity from a social robot? In *Proc. HRI*, 2015.

[13] D. Harel. Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, 8(3):231–274, 1987.

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

[15] N. Kardaris, I. Rodomagoulakis, V. Pitsikalis, A. Arvanitakis, and P. Maragos. A platform for building new human-computer interface systems that support online automatic recognition of audio-gestural commands. In *Proc. ACMMM*, 2016.

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

[17] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. Child speech recognition in human-robot interaction: evaluations and recommendations. In *Proc. HRI*, pages 82–90, 2017.

[18] F. Kirstein and R. Risager. Experiences from long-term implementation of social robots in Danish educational institutions. In *What Social Robots Can and Should Do*, volume FAIA-290. 2016.

[19] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Proc. CVPR*, 2008.

[20] S. Moubayed, G. Skantze, and J. Beskow. The furhat back-projected humanoid head–lip reading, gaze and multi-party interaction. *Intl. J. Humanoid Robotics*, 10(01), 2013.

[21] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE*, 91(9):1306–1326, 2003.

[22] J. C. Pulido, J. C. González, C. Suárez-Mejías, A. Bandera, P. Bustos, and F. Fernández. Evaluating the child–robot interaction of the naotherapist platform in pediatric rehabilitation. *International Journal of Social Robotics*, 9(3):343–358, 2017.

[23] W. L. Richman, S. Kiesler, S. Weisband, and F. Drasgow. A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, 84(5):754, 1999.

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

[25] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos. Room-localized spoken command recognition in multi-room, multi-microphone environments. *Computer Speech & Language*, 46:419–443, 2017.

[26] R. Ros, M. Nalin, R. Wood, P. Baxter, R. Looije, Y. Demiris, T. Belpaeme, A. Giusti, and C. Pozzi. Child-robot interaction in the wild: advice to the aspiring experimenter. In *Proc. ICMI*, 2011.

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

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

[29] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *Proc. CVPR*, 2011.

[30] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proc. BMVC*, 2009.

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

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

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