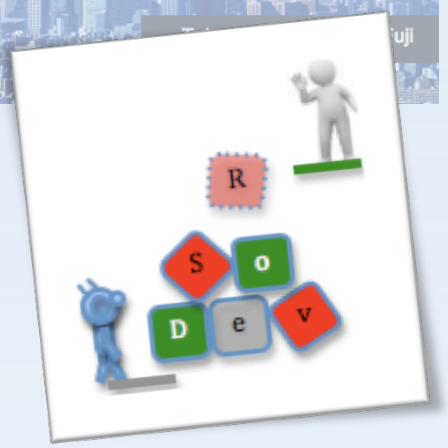
DevSoR 2013

IROS2013
New Horizon



PROCEEDINGS OF

International Workshop on Developmental Social Robotics (DevSoR): Reasoning about Human, Perspective, Affordances and Effort for Socially Situated Robots

IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2013
Tokyo, Japan
7 November, 2013

Editors

Amit Kumar Pandey
Rachid Alami
Alessandro Saffiotti
Peter Ford Dominey
Kazuhiko Kawamura

Sponsoring Projects

EU *SAPHARI* (*Safe and Autonomous Physical Human-Aware Robot Interaction*) project, funded by the European Community's 7th Framework Programme FP7-IST under Contract ICT-287513.

Romeo2 (*Humanoid Robot Assistant and Companion for Everyday Life*) project, funded by BPIFrance in the framework of the Structuring Projects of Competitiveness Clusters (PSPC).

PROJET
ROMEO2

Preface

Open-ended embodiment of robot's social intelligence will be the key of life-long development and learning capabilities of the robots in the human centered environment. This will facilitate the robot to adapt and enhance itself and exhibit socially accepted and socially expected behaviors. Therefore, the heart of the workshop lies in the concept of bottom up development of socially intelligent robots. For this, we need to identify the basic cognitive and behavioral blocks, which could facilitate the robot to develop more complex socio-cognitive intelligence. Hence, this first edition of the workshop focuses on some of such basic blocks: reasoning about human, perspective taking, affordance, effort, social signal and their applications.

The *International Workshop on Developmental Social Robotics (DevSoR): Reasoning about Human, Perspective, Affordances and Effort for Socially Situated Robots*, held during IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2013, on November 7, 2013 at Tokyo, Japan. The full day workshop started by an introduction by Dr. Amit Kumar Pandey, on the core idea behind the Developmental Social Robotics and the motivation behind the workshop. This was followed by 4 dynamic sessions, *Affordance based Reasoning* (chaired by Prof. *Kazuhiko Kawamura*), *Social Interaction based Reasoning and Learning I* (chaired by Dr. *Rachid Alami*), *Social Interaction based Reasoning and Learning II* (chaired by Prof. *Joachim Hertzberg*) and *Modeling Social Intelligence* (chaired by Prof. *Mohamed Chetouani*). The sessions, in addition to the paper presentations, featured five distinguished invited talks, from Prof. *Ashutosh Saxena* (Cornell University, USA), Prof. *Sinan Kalkan* (Middle East Technical University, Turkey), Prof. *Mohamed Chetouani* (Institute for Intelligent Systems and Robotics (ISIR), Paris, France), Prof. *Angelo Cangelosi* (University of Plymouth, UK) and Prof. *Kazuhiko Kawamura* (Vanderbilt University, USA).

At the end of the technical presentations, a stimulating panel discussion was organized, titled "*Developing Socially Intelligent Robot: What are the immediate challenges?*", which was moderated by Dr. *Rachid Alami and the* other members of the panel were Prof. *Angelo Cangelosi*, Prof. *Mohamed Chetouani* and Prof. *Kazuhiko Kawamura*. The panel discussion with active participation from the audience successfully identified some important aspects of immediate focus for developing socially intelligent robotics. It has been agreed that as now we have ways to analyze and compute various affordances, therefore we should now focus more on the use of affordance in various dimensions of social intelligence. Another aspect, which emerged was the need of design metrics for evaluating HRI, social interaction and development. Coming up with approaches to accelerate the 'developmental' process was another concern. To achieve this, it was suggested to equip the robot already with the necessary basic components of socio-cognitive development, instead of letting them grow as day one child, as well as to focus on sharing and transferring the learning and skills among robots (these in fact are the basic motivation behind this workshop). Addressing the complexity of lexical analysis for verbal communication and not forgetting the aspect of safety was among other interesting points.

Regarding the papers contributions, the workshop successfully attracted contributions from diverse domains on some of the basic aspects of social development. Affordance emerged as one of the most contributing aspects. Perspective taking, social interaction (verbal and non-verbal) based social learning, and modeling aspects of cognition and intention were other interesting domains of contributions. This proceeding includes all the accepted contributions presented at the workshop. The paper by *Awaad et al*. (presented by *Iman Awaad*) exploits the notion of functional affordances with conceptual similarity for substituting the right objects, during the planning to achieve the task. Whereas, the paper by *Ugur et al*. (presented by *Emre Ugur*) gets inspiration from infant development and uses the notion of *motionese* to identify the important steps and the boundaries for learning sub-goals of a task from demonstrations. The paper by *Tan et al*. (presented by *Kazuhiko Kawamura*) integrates the aspects of control, cognition and intention for the robot to adapt its behavior during social interaction. Whereas, the paper by *Grigore et al*. (presented by *Elena Corina Grigore*) shows the importance of socially assistive robotics (SAR) approach to teach children through social interaction. The paper by *Baddoura et al*. (presented by *Gentiane Venture*) presents a study about the people's first encounters with the robot, and how do those shape the sociability of the Human-Robot Interaction (HRI). The paper by *Rousseau et al*. (presented by *Salvatore Anzalone*) explores the notion of joint attention through perspective taking mechanism during social interaction to learn objects' names. Finally, the paper by *Daglarli et al*. (presented by *Gökhan İnce*) presents some of the basic constructs inspired by brain architecture to model socio-cognitive aspects of HRI.

We would like to thank all the participants, contributors, program committee members, the reviewers, the invited speakers and the panelists, to shape the workshop, by bringing an interesting program, excellent technical presentations and stimulating discussion. All these have paved the way for the successive editions of the workshop, by identifying the challenges we should focus in the development of socially intelligent robots. We would like to thank IROS 2013 organizing committee and IROS 2013 workshop chairs *Fumihito Arai*, Nagoya University (Japan), *George Lee*, Purdue University (USA), *Antonio Bicchi*, University of Pisa (EU) for providing us this great platform and assisting us in successful organization of the workshop. We would like also to express our thanks and gratitude to the local organizing team to facilitate smooth flow during the entire workshop.

We hope to see you all in the next edition of the DevSoR workshop, soon.

IROS-DevSoR 2013 Organizing Committee
November 2013

Amit Kumar Pandey, LAAS-CNRS, Toulouse, France
Rachid Alami, LAAS-CNRS, Toulouse, France
Alessandro Saffiotti, AASS, Örebro University, Sweden
Peter Ford Dominey, INSERM, France
Kazuhiko Kawamura, Vanderbilt University, USA

Committees

Organizing Committee

- *Amit Kumar Pandey*, *LAAS-CNRS, Toulouse, France*
- *Rachid Alami*, *LAAS-CNRS, Toulouse, France*
- *Alessandro Saffiotti*, *AASS, Örebro University, Sweden*
- *Peter Ford Dominey, INSERM, Stem Cell & Brain Research Institute, France*
- *Kazuhiko Kawamura, Vanderbilt University, USA*

Program Committee

- *Rachid Alami, LAAS-CNRS, Toulouse, France*
- *Ritta Baddoura,* CRISES, University of Montpellier, France
- *Mohamed Chetouani, ISIR, UPMC, Paris, France*
- Aurelie Clodic, *LAAS-CNRS, Toulouse, France*
- Lavindra de Silva, *LAAS-CNRS, Toulouse, France*
- *Peter Ford Dominey, INSERM, Stem Cell & Brain Research Institute, France*
- Shuzhi Sam Ge, *Social Robotics Lab, IDMI, National University of Singapore, Singapore*
- Rodolphe Gelin, *Aldebaran Robotics, France*
- *Joachim Hertzberg, Institute of Computer Science, Osnabrück University, Germany*
- *Anuj Kapuria, Hi-Tech Robotics Systemz, India*
- *Kazuhiko Kawamura, Vanderbilt University, USA*
- Alexandra Kirsch, *Tübingen University, Germany*
- Séverin Lemaignan, *EPFL, Lausanne, Switzerland*
- *Manuel Lopes, INRIA Bordeaux Sud-Ouest, France*
- *Amit Kumar Pandey, LAAS-CNRS, Toulouse, France*
- *Alessandro Saffiotti, AASS, Örebro University, Sweden*
- *Emrah Akin Sisbot, Toyota InfoTechnology Center, Mountain View, CA, USA*
- *Luis Felipe Marin Urias, Universidad Veracruzana, Mexico*

Table of Contents

Socializing Robots: The Role of Functional Affordances 1
Iman Awaad, Gerhard Kraetzschmar, Joachim Hertzberg

Affordance based imitation bootstrapping with motionese 9
Emre Ugur, Yukie Nagai, Erhan Oztop

Integrating Cognitive Control and Imitation Learning in a Socially Situated Robot 15
Huan Tan, Don Wilkes, Kazuhiko Kawamura

Feasibility of SAR Approaches - Helping Children with Learning Tasks 22
Elena Corina Grigore, Brian Scassellati

Human motion measures and adequacy of the response in relation to the robot's sociable character 25
Ritta Baddoura, Gentiane Venture

Learning object names through shared attention 31
Woody Rousseau, Salvatore Anzalone, Mohamed Chetouani, Olivier Sigaud, Serena Ivaldi

Computational Model of the Biologically Inspired Cognitive Architecture for Human Robot Interaction 37
Evren Daglarli, Hatice Kose, Gökhan İnce

Authors Index 44

Socializing Robots: The Role of Functional Affordances*

Iman Awaad[1], Gerhard K. Kraetzschmar[1] and Joachim Hertzberg[2]

***Abstract*— Just as humans behave according to the social norms of their groups, autonomous systems that become part of these groups also need to behave in socially-expected and accepted ways. For humans these social norms are learned through interaction with members of the group. In this work, we propose that the functional affordances of objects, what objects are meant to be used for, provide us with a starting point for the socialization of such agents. We model these functional affordances in Description Logics (DL) and show how this enables the socially-expected human behavior of substituting objects as needed to achieve a goal. In addition, we propose to combine these affordances with conceptual similarity and proximity in order to make more complex substitutions, which are socially acceptable in their given context. Finally, we describe how their use would allow the agent to take advantage of opportunities and how they are modified and extended through interaction with humans.**

I. INTRODUCTION

Domestic service robots are expected to carry out tasks around the house, such as cleaning, fetching objects, serving drinks and so on. Their success has traditionally been based on their ability to understand commands and accomplish the given tasks. Such agents are typically mobile manipulators with capabilities at varying levels of complexity, such as perception, mapping, manipulation, navigation, dialog management and task planning. The integration of these capabilities to form an architecture which enables a flexible and robust agent remains a focus of much research.

An agent which is also capable of learning, be it from demonstration, by experimentation, or by querying external knowledge bases, is no longer simply desirable, but necessary for life-long learning. We argue, that in addition to this, domestic service robots need the capability to acquire, assimilate and apply the social norms of the group with which they interact so that they can behave in socially-expected and accepted ways. Humans cohabiting the environment would gain from the more natural human-robot interaction which would result, and the agents would also gain from the flexibility that these norms provide humans when performing their everyday tasks.

*Iman Awaad gratefully acknowledges financial support provided by a PhD scholarship from the Graduate Institute of Bonn-Rhein-Sieg University. The work of Joachim Hertzberg reported here is supported by the RACE project, grant agreement no. 287752, funded by the EC Seventh Framework Programme theme FP7-ICT-2011-7.

[1]Iman Awaad and Gerhard K. Kraetzschmar are with the Department of Computer Science, Bonn-Rhein-Sieg University of Applied Sciences and B-IT Center, 53757 Sankt Augustin, Germany iman.awaad@h-brs.de, gerhard.kraetzschmar@brsu.de

[2]Joachim Hertzberg is with Osnabrück University and DFKI RIC Osnabrück Branch, 49076 Osnabrück, Germany joachim.hertzberg@uos.de

Children pick up social norms mainly through interaction with their families, and usually from schools, their own peers and (fortunately, or not) the media as well. They learn how to perform tasks, as well as how not to perform them (e.g. clothes should be hung or folded and placed in the closet; or that a glass of water should be placed on a coaster and not directly on the table). They learn manners (e.g. that they should use the word 'please' when asking for something), and they learn when and which substitutions are acceptable (e.g. that a mug may be substituted for a glass but not the other way around, and that such a substitution is not appropriate when the drink is for a guest). Such knowledge clearly goes beyond how to accomplish tasks.

Let us consider the task of serving water to someone. Humans *know* that a glass should be filled with water and served to the guest, and that glasses are in a particular cupboard in the kitchen. Robots should know this too. Should humans not find any glasses there, they *know* that there may be some glasses in a sink, or a dishwasher, but that they should be checked to ensure that they are clean. Robots should probably know this too. In some cases, humans may choose to simply use a mug instead of washing a glass. Robots may know that such substitutions are possible if this is specified explicitly. When serving the person, humans *know* that the glass should be placed on a coaster. Robots should know this too.

There are various types of knowledge described in the use case above. There is the knowledge of the goal (serving a glass of water) and matching this with the procedural knowledge of *how* to go about doing this. Usually, this knowledge includes the objects with which the tasks are to be accomplished (e.g. the glass, the coaster, the cupboard). There is also the knowledge of when it is socially acceptable to make substitutions and when it is not.

While the agent should be capable of learning much of the knowledge mentioned above, it is also expected to function sufficiently well, out-of-the-box. Knowing how to do many things, however, is not sufficient, nor can this be defined as intelligence. " ... The true test of intelligence is how we behave when we don't know what to do" [1].

What we want is an approach that allows the agent to determine when it has insufficient knowledge, to acquire it, when possible, or find an alternative, and then successfully carry out the task.

Our approach adopts the open world assumption (through the use of DL which by default assumes incomplete information), so unless our knowledge base contains a statement (or can infer one) to the effect that something is true or that it is false, our query would return 'do not know'.

In the work presented here, knowledge of how best to carry out basic tasks is encapsulated within Hierarchical Task Network (HTN) planning [2] methods and operators. Methods recursively decompose complex tasks into primitive ones which can be carried out through the execution of grounded operators. Together with the state of the world, these methods and operators constitute the planning domain.

In an environment shared by humans and artificial agents, this approach is beneficial, as it is more understandable for humans; and a good agent should be able to communicate its plan at all times [3]. In addition, it lets the human user to specify the way he/she wishes to have a task accomplished in an intuitive way.

Let us consider the task of watering a plant; the domain modeler would specify methods and operators which describe how the task is to be accomplished and specify that a watering can should be used. A planner queries for the initial state of the world and, given the goal, generates a task network. The plan is simply the sequence of actions found in the leaf nodes of the network from left to right.

If there is no watering can in the domain or, despite all of our methods and operators, no decomposition is found to accomplish the task, the plan generation process will fail. For example, when the watering can exists but is inaccessible and we have no means by which to make it accessible.

One would intuitively *expect* a human to ask for help or to simply use something else to accomplish the task (e.g. a tea kettle). This ability to effortlessly adapt our actions to unexpected situations, especially given the dynamic nature of our environment and the amount of uncertainty about it, is perhaps one of the most underestimated human abilities. Very often, changes in our plans have to do not so much with *how* we carry out a task, but *with what* we carry it out.

Similarly, it would be desirable for an autonomous agent to ask for help (instead of simply communicating that it cannot accomplish the task). It would be even better if the agent could itself reason about what a good substitution would be and ask for a user's approval before attempting to make the substitution.

This work argues for both the benefits that come from allowing agents to make substitutions; and demonstrates how the use of functional affordances, conceptual similarity and spatial proximity can allow agents to reason about and identify appropriate substitutions.

II. AFFORDANCES

The concept of affordances provides us with the necessary perspective with which to equip agents to behave with such flexibility. Affordances describe "*opportunities for action*" [4]. This work adopts the notion of affordances, although Gibson's action/perception coupling is not dealt with directly. Gibson's original definition has been refined by many researchers, but a generally agreed upon interpretation narrows the list of action choices to those of which an actor is aware. Using the refined definition, affordances are neither solely a property of the object nor of the actor, but of their relationship.

Under Gibson's original definition, the set of affordances for a given object may be quite large, and may include actions that are neither socially expected nor socially acceptable (e.g. throwing a chair). In this work, we adopt Norman's definition of *perceived affordances* which allude to "how an object may be interacted with based on *the actor's goals, plans, values, beliefs and past experience*" [5]. This is consistent with our ideal domestic service robot: a goal-based [6] agent that can also learn from experience, and adheres to the values and beliefs of its group.

A. Distributed Cognition

We need a starting point for our agent's socialization process – a kernel of norms, if you will, which represents these values and beliefs in a manner that permits the agent to make appropriate decisions. Where could knowledge of social norms come from, and what does it look like? To answer this question, a paradigm shift is necessary to view "knowledge" as facts that have been shaped by the values, beliefs and experiences of groups of people. For example, the fact that teacups are for drinking tea may not hold in Japan where tea is drunk from a bowl, or in Argentina where it is drunk from a hollow gourd. In fact, the word 'tea' itself would no doubt refer to different types of tea altogether. The English definition of a teacup is rooted in the English cultural tradition of drinking tea. We *know* this from experience (our own or, interestingly, that of *other individuals* of the group).

With this in mind, we note Hutchins's theory of distributed reasoning and cognition which states that knowledge lies not only within the individual, but in the individual's social and physical environment [7]. Others have further elaborated this idea [8]. This concept is appealing, as it acknowledges the impact that social groups have in shaping what we know. Moreover, it implies that it is no longer necessary for one to experience something him, her or itself in order to *know* something.

One could, therefore, argue that resources such as dictionaries, the Internet, WordNet, ConceptNet, OpenCyc [9], and the work of projects such as RoboEarth [10] are an example of distributed cognition, albeit for those groups whose native language is English since "language does not exist apart from culture" [11]. This paradigm allows us to reformulate the question: How can the agent acquire, reason and manipulate knowledge to behave in a socially compliant manner?

B. Functional Affordances

The answer lies in the simple notion that objects are made to be used for (or exist for use in) specific tasks, and that this knowledge has been shaped by the norms of the group. Such *functional affordances* [12] link the idea of "*purposeful actions*" to the objects, and account for *descriptive social norms* ("what is usually done in a given setting" [13]). They include within them the "values and beliefs" and provide us with a starting point for "past experience". They can then be manipulated and adapted based on further interaction with the social group that the agent is part of. The result is

behavior that is socially expected; we are using objects for what they were *meant* to be used for.

There are other benefits to using functional affordances. By considering functional affordances, and not *all* "opportunities for action", the action space is reduced. They also allow users to specify more general tasks. For example, when asked to "serve a drink", any object *meant* 'for drinking' could be served without the need to explicitly specify a particular drink. This is more important than it seems at first glance, since much of our interaction with each other involves a great deal of underspecification.

We propose to use dictionaries as a source from which an agent acquires these functional affordances. They provide concise and unambiguous definitions of objects that almost always include their function. For example, a teacup, is defined as "a cup from which tea is drunk" [14], and a cup is "a small, bowl-shaped container for drinking from, typically having a handle" [14]. Therefore, dictionaries make ideal sources to mine the functional affordances of objects from.

Objects may have more than one functional affordance (e.g. a bottle has the primary functional affordance of storing liquids but a secondary functional affordance can be *learned through interaction*: it is also for drinking from). These affordances are included within the domain model and are represented compactly in DL. This allows us to use the reasoning powers of existing tools to bring about the robust and flexible behavior described above. The functional affordances of parts of objects are also modeled. This has two main benefits. First, it acts as a causal link, explaining why an object has a given affordance, and second, it helps the agent to recognize affordance cues or stimuli [15], [16] at execution time and respond to them.

Our HTN planning domain already provides us with the 'best way' of carrying out a task (e.g. it would specify that tea should be served in a teacup). Knowing the function of an object allows us to behave flexibly in case of plan generation failure (e.g. we know that all tea cups are dirty, and we do not know how to make them clean, so a plan cannot be generated) or execution failure (e.g. we did not know they were all dirty at planning time but found out during the course of execution). Choosing to use another object with the same functional affordance is the socially-expected and generally accepted course of action.

III. SOCIALIZING AGENTS

In the following sections, we demonstrate how our affordance-based approach leads to flexibility, makes for compact representations, and allows the social norms to be refined, to those of the group, through interaction. For example, humans cohabiting the environment might ask the robot to clean the bathrooms only with the blue cleaning cloths, or to serve them tea only in their favorite cup.

A. Socially-expected Behavior

In Section I, we saw how the act of making substitutions is a socially-expected behavior in itself. We expect that people are able to find ways to accomplish their tasks under all but the most extreme cases. In this section we demonstrate how the combination of procedural knowledge (*how to* accomplish a task) and the functional affordances of objects (*what* objects are meant to be used for) together provide us with the socially-expected *choice* of the substituted objects (e.g. glasses and mugs are both used to drink from).

Simply querying the knowledge base (KB) for objects with the given functional affordance provides us with an appropriate substitute. This is accomplished without the need for cumbersome, ad hoc and often subjective categorization of objects. For example, ontologies of domestic objects (such as those presented in Section II-A) may contain categories such as 'furniture' or 'perishable objects'. The problem with this is, first, that it simply refers to qualities that a group of items may have (in the case of perishable goods, they will eventually perish); and that the decision of whether an object belongs to such a category or not may be subjective (is a spoon or a chandelier considered furniture?).

World models tend to describe the form of the world: objects, their shapes, colors or locations and their spatial relationship to one another. The same world can be described by the functions it is meant to afford. Studies in child psychology have found that children use functional affordances to generalize the name of newly-learned artifact categories and otherwise rely on global similarity when they could not interact with the objects [17].

There are cases, however, when using functional affordances alone will not be enough. Some objects are used for a very specific task (e.g. watering cans are used for watering plants). The only other object which is used for the same task would be a 'hose', and this is only for watering plants outdoors. In this case, both share the same functional affordance of watering plants; but whereas it may be desirable to substitute the watering can for the hose, the opposite is not true, and so a substitution using only functional affordances may fail.

Here, the agent would need to look for objects which are conceptually similar to the watering can. The similarity measures which are often used may not yield the results we have in mind (we may not care about the color of an object, but rather the presence of a handle for example).

For describing similarity, we propose the use of *Conceptual Spaces* [18]. They provide a multidimensional feature space where each axis represents a quality dimension (e.g. brightness, intensity, and hue). Points in a conceptual space represent objects, while regions represent concepts.

Let us take the example given in [18]: the three quality dimensions in our example above can together be used to describe the 'color' domain. A region on the red axis could be described as having the property 'red'. A point in this region could represent the concept 'apple' in conjunction with other domains such as 'taste' or 'shape'. We could even relate the property 'red' to the taste 'sweet'.

Conceptual spaces are built up by the various quality dimensions. The agent should learn the relation between these quality dimensions and given tasks. For example, for lifting an object, the most important quality dimension is its

weight – its color would be irrelevant. These relations could then be used as weighting factors to determine how well an object would substitute for another in achieving a given task (similarity would be measured as the weighted Euclidean distance).

Conceptual spaces can also represent shape (e.g. handles and spouts). The detection of these quality dimensions obviously requires more processing by the perception components than for example, the simple detection of hue. To substitute a watering can to water plants, the capacity to hold water is the most important affordance, followed by the presence of a handle and a spout. Using conceptual spaces, the agent might find that the tea kettle is the most appropriate substitution. The combination of active perception at execution time and task-oriented perception would allow the agent to actively search for those features (e.g. spouts and handles) which are relevant to the task at hand, as opposed to passively picking up any and all cues. [19] has shown that the time complexity for such a search is far better when compared to a data-driven search.

B. Socially-accepted Behavior

Having shown how functional affordances provide us with the ability to make basic socially-expected substitutions, we now demonstrate how socially-acceptable substitutions can be made by combining them with conceptual similarity and proximity in various ways to create a hierarchy of constraints.

Using functional affordances and conceptual similarity, an artificial agent can start by attempting to satisfy the constraints specified in the methods and operators (e.g only use a unique instance, such as *my* teacup – if this was specified in the goal – or an instance of a given object). If it fails to find the suitable object, it would iteratively attempt to find objects which satisfy fewer and fewer constraints.

The first level above that of using an instance of a given object (e.g. a teacup) is to use any object with the same functional affordance and high conceptual similarity (e.g. a mug). The next higher level would remove the constraint that the substitute should be conceptually similar, relying only on a shared functional affordance (e.g. a drinking flask). Should the agent not find such objects and given the old adage that "form follows function" (the form of objects is based on their function), conceptual similarity is then used to identify those objects which do not share the same functional affordance and yet are conceptually similar (e.g. a measuring cup). The top level attempts to infer the function-relevant attributes and identify objects matching these properties (e.g. a jar).

It is important to note here that *injunctive social norms* ("what is typically approved in society" [13]) are highly dependent on context and may differ from person to person. For example, it may be acceptable for me to have my tea served in a mug, but may not be acceptable in the presence of guests, or for another user.

The ability of the agent to acquire and manage user preferences in their proper context is a must. Both of these topics are currently being investigated.

Humans prefer to take advantage of objects within their immediate spatial surroundings when making substitutions (e.g. using magazines which may be on the table instead of a coaster). Agents should also exploit spatial proximity. The importance of proximity can be altered to make it either easier to move from one level of constraints to another by increasing its importance (prefer objects which are close, even if they are in a less constrained category of objects), or more difficult to move up by decreasing its importance.

The work presented in [20] and [21], while addressing a different focus, is perhaps the closest to ours in that they also use functional affordance and conceptual spaces to measure similarity. Their work is based on an adapted version of the HIPE theory of action and so they have included additional types of affordances based on both physical and socio-institutional constraints.

C. Acting

Making substitutions also makes it possible for agents to take advantage of opportunities, e.g. using a magazine that happens to be on the table as a coaster. In order to accomplish this, we need to combine both the execution of plans which have been generated through the deliberation process and reactive behaviors which may be triggered by affordance cues.

We propose a simple blackboard architecture where affordance cues, in the form of conceptual space quality dimensions, are being posted by all artificial agents as they move through the environment. These cues might be of varying complexity, from simple color hues which would cost very little in terms of perceptual processing to more complex concepts such as shape. They might have been picked up as part of the plan's execution, and would be kept in the system for a given duration. Upon execution failure, the cues which are in close proximity can be used to identify viable candidates for substitutions. As Steedman points out in [22], "...it is probably better to look at those plans the situation affords, rather than backward chaining to conditions that there may be no way for you to satisfy...". By having all agents post cues, information about the state of the world can be shared.

The same behavior can guide plan execution when things are going as planned, allowing the agent to take advantage of opportunities before failures occur. For example, cues associated with a drink bottle may have been picked up on the way to the location specified in a plan. This 'short cut' could be taken advantage of, again depending on the flexibility that the human user has allowed. A cupboard full of glasses would guide the agent to grasp any of them. In the case of execution failure, an agent might take the more 'resourceful route' of making a substitution or attempt to use the same object by finding other instances, or of using objects with the same functional affordance.

Taking advantage of opportunities by reacting to affordance cues has the added benefit of injecting that bit of randomness that often leads to improvements. Although our approach will use a plan library to avoid having to generate

plans for the same routine tasks from scratch, reliance on these plans could lead to stagnation. New events or objects may provide a better way for the task to be accomplished, but would never be considered if the same saved plan is always reused. For example, buying an autonomous vacuum cleaner may render the continued use of the conventional one undesirable. The use of a strategy that would occasionally resort to generating a plan, instead of using one from the library, should be investigated.

Whether through interaction or through observation, an agent's view of objects' functional affordances will change. If we recall the possibility of learning an additional functional affordance for a bottle as mentioned in Section II-B, we see that it is just as possible for the functional affordance to be unlearned or marked undesirable if the agent is so instructed. This developmental dimension that arises by considering the function of objects fits nicely with the profile of our desired domestic service robot.

One open question is when the agent should stop climbing the "flexibility ladder", that is, when it should concede that it cannot achieve the task.

IV. DISCUSSION

As mentioned above, the focus of our work is to enable agents to handle unexpected situations more robustly by substituting objects as humans do. Traditionally, researchers have focused on learning the link between objects and actions, for example, by analyzing the link between object attributes and the actions they afford (as in [23], [24], [25]), through experimentation (e.g. [26], [27], [28], [29]), or by imitation/demonstration/action recognition (e.g. [30], [31]).

These approaches involve time-consuming processes, and a number of them may yield affordances which lead to the suboptimal use of objects. Moreover, they may neither be goal-oriented nor socially acceptable, as they ignore context/situation. Simply linking object attributes with the actions that they afford does not provide us with socially-expected or acceptable behavior. For example, the affordance *rollable* is often given to cylindrical parts, although rolling bottles is not a socially-expected behavior. Knowing that concave objects are fillable is obviously useful, but simply knowing that a spoon may be filled, does not make the task of watering a plant with one socially expected (or acceptable). Learning affordances by experimenting can confirm or repudiate the existence of an affordance, but says nothing about the affordance leading to socially-compliant behavior. That is not to say that experimentation is not an important part of development. It lets agents understand their own body's movements and by so doing, facilitates learning by imitation [32].

Learning from demonstration is more suitable for socializing robots. The ongoing research in the field of action recognition is highly relevant. It lays the foundation, not just for equipping agents to learn how to perform new tasks by watching humans perform them, but for the more socially-complex ability of anticipating the actions of humans (as in [33], [34], [35]). Proactive behavior, in response to anticipating other agents' actions, can be seen as opportunistic behavior. This research also facilitates the socialization process directly as the substitutes that humans themselves make in different situations can be learned by the agent.

The work carried out in the natural language understanding field is also of great importance as it would make available to the agent the vast quantity of written, as well as audio and video resources, in addition to facilitating communication between the lay user and artificial agents. Moreover, the study of language is inherently intertwined with our own understanding of 'planned action' [22]. Language describing human-object interactions [36] is already a focus.

Language is but part of the human-robot interaction process. Another vital component of social intelligence [37] is having agents pick up social cues, e.g. to recognize humans' emotional states in order to respond effectively. This requires multimodal interaction such as gesture, gaze, head movements, vocal features, posture, proxemics and touch [37]. Together, such building blocks would allow more sophisticated behavior to emerge, such as perspective taking [38].

Systems that learn users' personal preferences through repeated interactions with them (such as [39]) and manage various profiles are necessary. Much work already exists – our own online profiles and preferences are tracked and managed. We even actively facilitate this process via social media. Investigating the application of these methods to our human-robot interaction scenarios, would be beneficial.

Given the stated goal of placing service robots in domestic environments, it is surprising that marketing professionals, whose job it often is to place consumer goods in domestic environments, have not played a larger role. It is their job to know the consumer well, to be able to recognize the various target groups and to know what each group *expects* and would *accept*. It would seem that they are best placed to identify which out-of-the-box capabilities/features these groups would expect.

The importance of context has been emphasized throughout this work; however, the means by which to represent situations and context remain a focus of research. For the moment, our approach simply represents context implicitly with the plan library as the initial state of the world. The work of projects such as RACE [40] is therefore of vital importance as they tackle the complex issues of representing whole experiences (including context) and learning from them.

Researchers in the inter-disciplinary field of normative multiagent systems (nMAS) have reaffirmed the importance of norms being contextual [41] and have investigated both computational models of norms, and architectures which support their use, among others. The violation of norms, and the consequences of doing so, are a major theme within the field. In this work, this topic is not dealt with, nor do we take into consideration the concepts of obligation, prohibition, deadline, or role for example. The scope of our use of norms remains limited to the substitution behavior.

V. CURRENT STATUS AND FUTURE WORK

In Section I, we described many types of knowledge involved in simply fetching a glass of water. Within our system, some of this knowledge is stored in an OWL-DL KB. This allows us to leverage the efficient reasoning powers of the available tools, such as the Pellet reasoner used here. In contrast to OWL-Full, which is used in research work such as [42] and [43], OWL-DL is decidable.

HTN methods and operators encapsulate the procedural knowledge of how to accomplish a task. The use of HTN planning, and of its hierarchical approach in particular, is a design choice which fits nicely with the overall approach presented here. The choice of the SHOP family of planners is a pragmatic one. The arguments presented here for the use of functional affordances, conceptual similarity, and proximity, could in essence be used to make substitutions using other planning approaches, given that objects' functional affordances are included in the domain, and that conceptual similarity and proximity can be measured.

To demonstrate how substitutions can be made in our approach, let us consider the task of serving tea. Intuitively, this involves going to the kitchen, making the tea and serving it to the user. Let us assume that we have, stored in the KB, the contents of the house including: two teacups (`teacup1`, `teacup2`), and two mugs (`mug1`,`mug2`).

In order to generate a plan, we use the approach presented in [44] to first create a constrained problem – one that only includes the relevant methods, operators and states of the objects that could be relevant. This is accomplished by also representing the methods and operators in the OWL-DL KB. Having identified the given task to be accomplished: `serve`(`robot`, `embodiedAgent`, `drink`), the relevant methods and operators are instantiated in the ABox. Some examples are depicted in Figure 1. The approach uses the information within these methods and operators to extract the relevant literals describing the initial state. The preconditions which determine applicability include e.g. teacups, and so this information is included within our planning problem (no mugs, flasks, measuring cups, or jars for example, are included in the constrained domain at this stage). The planner proceeds to generate the plan, grounding the variables appropriately, e.g. `kettle1` for ?`kettle`. Let us assume, that `clean`(?`teacup`) could not be instantiated because both teacups were dirty. Climbing the "flexibility ladder" involves detecting this failed precondition and passing it on to the Control module (see Fig. 2). It then triggers the expansion of the domain to include a possible substitute based on the current "rung" of the "ladder" and the importance of proximity. In this case, the module would query for other objects used for drinking from, and then choose instances of those which are most similar to a teacup. For example, mugs are more similar than bottles. Simply substituting one object for another can be insufficient. The object may have specified methods and operators that accomplish the same task differently. For example, filling a tea kettle would involve removing the lid and replacing it again. These steps

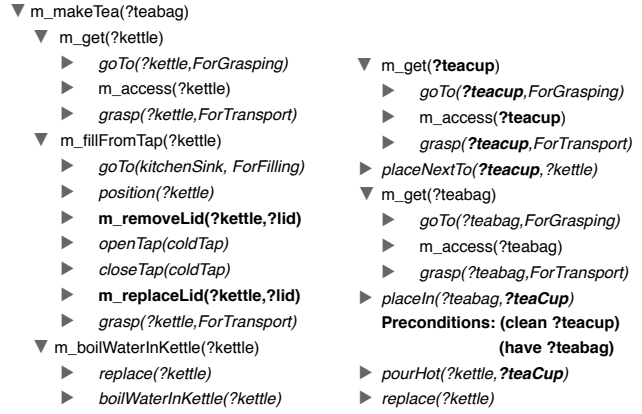


Fig. 1. An example decomposition of the `makeTea` task. In bold are examples, from the text, of the differences in method decompositions for `m_fillFromTap`(?`kettle`) and `m_fillFromTap`(?`wateringCan`), variable types (?`teacup`), as well as an example of the preconditions used to generate the initial state.

are absent when filling a watering can. Our approach takes these situations into account and can therefore be seen as transforming both the goal and the plan, as in [46]. The Control module expands the domain to include the instances of the substitutes as well as all methods and operators that include its variable type in their arguments, thus ensuring that e.g. the method `fill`(?`kettle`) is included in a newly expanded domain to water plants.

To accomplish the task with a substituted object, the original decomposition should be preserved as much as possible. Thus, a placeholder instance of a clean `teacup` is given to the planner and a plan is successfully generated. At this point, each method and operator (original plan) is swapped with a corresponding method/operator for the new object (if one exists), while the remaining task network is preserved. The variable bindings in the original methods and operators for ?`teacup` are replaced with ?`mug` instead. The preconditions and effects for the actions of the new plan are then verified to ensure it is executable.

In the KB, members of the *HumanScaleObject* class are related to a functional affordance via the *usedFor* property. Functional affordances, such as *DrinkingFrom*, *Boiling*, or *Watering*, are subclasses of the *ActionOnObject* class. Fig. 3 shows part of the ontology.

In our ongoing research, we are investigating the representation of objects in conceptual spaces and the specification, then learning, of the objects' quality dimensions' weights for a given task. The possibility of autonomously acquiring the functional affordances from online sources, as well as developing a means to manage profiles and preferences in the form of the plan library, will also be investigated.

ACKNOWLEDGMENT

The authors would like to thank Elizaveta Shpieva for her help in implementing some of the ideas presented here. The authors also thank the reviewers for their valuable feedback which has helped to improve this manuscript.

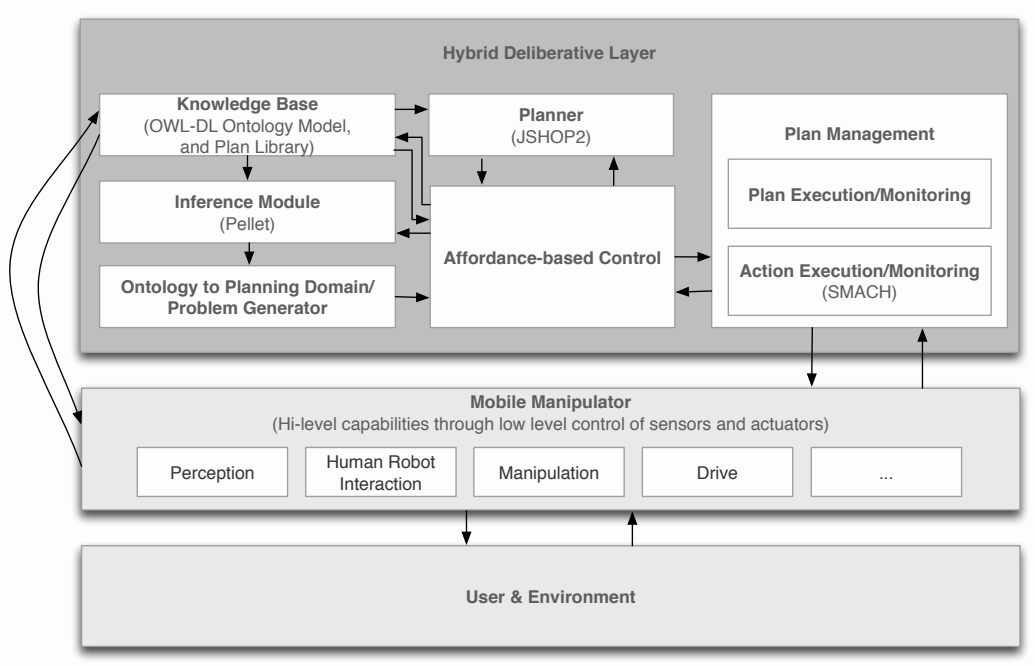


Fig. 2. Software architecture of the system extending the hybrid deliberative layer to use affordance-based reasoning in a domestic environment [45]

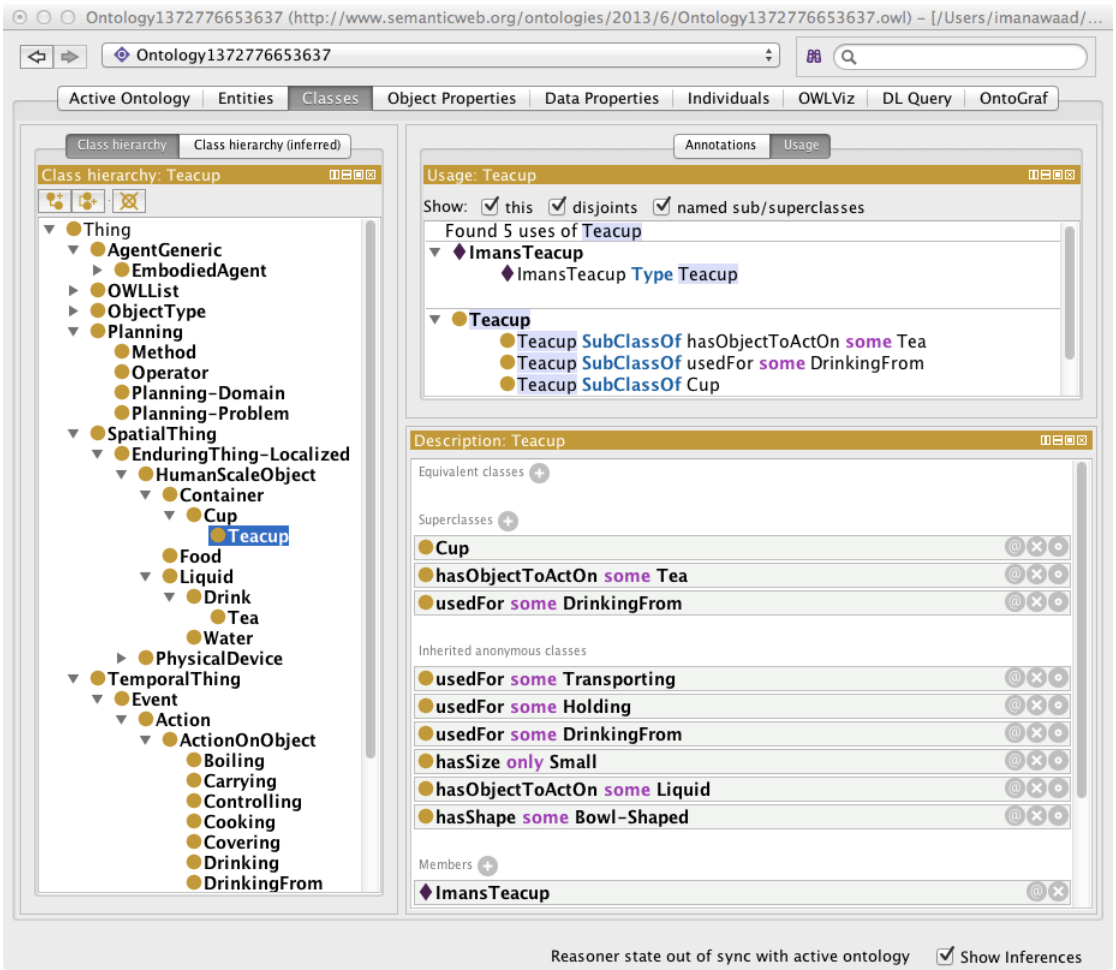


Fig. 3. The representation of the Teacup class in DL. An instance, ImansTeacup, and the class's properties, including inherited ones are seen on the right

REFERENCES

[1] J. C. Holt, *How Children Fail*. Pitman, 1964.

[2] K. Erol, J. Hendler, and D. S. Nau, "HTN planning: Complexity and expressivity," in *In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*. AAAI Press, 1994, pp. 1123–1128.

[3] J. M. Bradshaw, P. J. Feltovich, and M. Johnson, *The handbook of human-machine interaction: a human-centered design approach*. Farnham, Surrey, England; Burlington, VT: Ashgate, 2011, ch. 13, pp. 283–300.

[4] J. J. Gibson, *The ecological approach to visual perception*. Houghton Mifflin (Boston), 1979.

[5] D. Norman, *The psychology of everyday things*. Basic Books (New York), September 2002.

[6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Prentice Hall, December 2003.

[7] E. Hutchins, *Cognition in the wild*. MIT Press, 1995.

[8] J. Zhang and V. L. Patel, "Distributed cognition, representation, and affordance," *Cognition and Pragmatics*, vol. 14, no. 2, pp. 333–341, 2006.

[9] Cycorp, "OpenCyc," Online at http://www.opencyc.org/, http://www.opencyc.org/.

[10] N. Hubel, G. Mohanarajah, R. van de Molengraft, M. Waibel, and R. D'Andrea, "RoboEarth Project," Online at http://www.RoboEarth.org, 2010.

[11] E. Sapir, *Language: An introduction to the study of speech*. New York: Harcourt, Brace and company, 1921.

[12] H. R. Hartson, "Cognitive, physical, sensory, and functional affordances in interaction design," *Behaviour & IT*, vol. 22, no. 5, pp. 315–338, 2003.

[13] R. R. Reno, R. B. Cialdini, and C. A. Kallgren, "The transsituational influence of social norms," *Journal of Personality and Social Psychology*, vol. 64, 1993.

[14] E. McKean, Ed., *The New Oxford American Dictionary*. Oxford University Press, May 2005.

[15] G. Fritz, L. Paletta, G. Dorffner, R. Breithaupt, and E. Rome, "Learning predictive features in affordance based robotic perception systems," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, 2006, pp. 3642–3647.

[16] J. Ortmann and W. Kuhn, "Afforded actions," *Applied Ontology*, 2012.

[17] D. G. K. Nelson, "Attention to functional properties in toddlers' naming and problem-solving," *Cognitive Development*, vol. 14, no. 1, pp. 77 – 100, 1999. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885201499800197

[18] P. Gärdenfors, "How to Make the Semantic Web More Semantic," in *Proceedings of the Third International Conference (FOIS 2004)*, 2004, pp. 17–34. [Online]. Available: http://www.lucs.lu.se/People/Peter.Gardenfors/Articles/Fois2004PG.doc

[19] R. C. Arkin, *Behavior-Based Robotics*, ser. Intelligent Robots and Autonomous Agents. Cambridge, MA, USA: MIT-Press, 1998.

[20] K. Janowicz and M. Raubal, "Affordance-based similarity measurement for entity types," in *COSIT'07: Proceedings of the 8th international conference on Spatial information theory*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 133–151.

[21] M. Raubal and R. Moratz, "A functional model for affordance-based agents," in *Proceedings of the 2006 international conference on Towards affordance-based robot control*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 91–105.

[22] M. Steedman, "Plans, affordances, and combinatory grammar," *Linguistics and Philosophy*, vol. 25, 2002.

[23] K. Varadarajan and M. Vincze, "Object part segmentation and classification in range images for grasping," in *15th International Conference on Advanced Robotics (ICAR)*, June 2011, pp. 21 –27.

[24] T. Hermans, J. M. Rehg, and A. Bobick, "Affordance prediction via learned object attributes," in *IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration*, 2011.

[25] M. Stark, P. Lies, M. Zillich, J. Wyatt, and B. Schiele, "Functional object class detection based on learned affordance cues," in *6th International Conference on Computer Vision Systems (ICVS)*, vol. 5008. Santorini, Greece: Springer Berlin / Heidelberg, 2008, pp. 435–444.

[26] E. Ugur, E. Sahin, and E. Oztop, "Unsupervised learning of object affordances for planning in a mobile manipulation platform," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 4312–4317.

[27] B. Ridge, D. Skocaj, and A. Leonardis, "Self-supervised cross-modal online learning of basic object affordances for developmental robotic systems." in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2010, pp. 5047–5054.

[28] D. Kraft, R. Detry, N. Pugeault, E. Baseski, J. H. Piater, and N. Krüger, "Learning objects and grasp affordances through autonomous exploration," in *Computer Vision Systems, 7th International Conference on Computer Vision Systems, ICVS 2009, Liège, Belgium, October 13-15, 2009, Proceedings*, ser. Lecture Notes in Computer Science, M. Fritz, B. Schiele, and J. H. Piater, Eds., vol. 5815. Springer, 2009, pp. 235–244.

[29] J. Sun, "Object categorization for affordance prediction," Ph.D. dissertation, Georgia Institute of Technology, August 2008.

[30] B. Moldovan, M. V. Otterlo, P. M. Lopez, J. Santos-Victor, and L. D. Raedt, "Statistical relational learning of object affordances for robotic manipulation," in *ILP*, 2011.

[31] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[32] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning about objects through action – initial steps towards artificial cognition," in *Robotics and Automation (ICRA), 2003 IEEE International Conference on*, 2003, pp. 3140–3145.

[33] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," in *Proceedings of Robotics: Science and Systems (RSS)*, 2013.

[34] A. K. Pandey, "Towards socially intelligent robots in human centered environment," Ph.D. dissertation, University of Toulouse, June 2012.

[35] E. Ugur, E. Sahin, and E. Oztop, "Predicting future object states using learned affordances," in *ISCIS*, 2009, pp. 415–419.

[36] M. Patel, C. H. Ek, N. Kyriazis, A. Argyros, J. Valls Miro, and D. Kragic, "Language for learning complex human-object interactions," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013.

[37] I. Poggi and F. D'Errico, "Social signals: A psychological perspective," in *Computer Analysis of Human Behavior*, A. A. Salah and T. Gevers, Eds. Springer, 2011, pp. 185–225.

[38] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which one? grounding the referent based on efficient human-robot interaction," in *19th IEEE International Symposium in Robot and Human Interactive Communication*, 2010.

[39] M. Mason and M. C. Lopes, "Robot self-initiative and personalization by learning through repeated interactions," in *Proceedings of the 6th international conference on Human-robot interaction*, ser. HRI '11. New York, NY, USA: ACM, 2011, pp. 433–440. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957814

[40] S. Rockel, B. Neumann, J. Zhang, K. Dubba, A. Cohn, S. Konecny, M. Mansouri, F. Pecora, A. Saffiotti, M. Günther, S. Stock, J. Hertzberg, A. Tome, A. Pinho, L. S. Lopes, S. von Riegen, and L. Hotz, "An ontology-based multi-level robot architecture for learning from experiences," in *Designing intelligent robots: reintegrating AI II*, 2013.

[41] G. Andrighetto, G. Governatori, P. Noriega, and L. W. N. van der Torre, Eds., *Normative Multi-Agent Systems. Dagstuhl Follow-Ups*, ser. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013, vol. 4.

[42] M. Tenorth and M. Beetz, "KnowRob - knowledge processing for autonomous personal robots," in *Intelligent Robots and Systems (IROS), 2009 IEEE/RSJ International Conference on*, 2009, pp. 4261–4266.

[43] M. Tenorth, A. Perzylo, R. Lafrenz, and M. Beetz, "The RoboEarth language: Representing and exchanging knowledge about actions, objects, and environments," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, 2012, pp. 1284–1289.

[44] R. Hartanto, "Fusing DL reasoning with HTN planning as a deliberative layer in mobile robots," Ph.D. dissertation, University of Osnabrück, August 2009.

[45] I. Awaad, G. K. Kraetzschmar, and J. Hertzberg, "Affordance-based reasoning in robot task planning," in *Planning and Robotics (PlanRob) Workshop at 23rd Intlernational Conference on Automated Planning and Scheduling (ICAPS)*, 2013.

[46] M. T. Cox and M. M. Veloso, "Goal transformations in continuous planning," in *1998 AAAI Fall Symposium on Distributed Continual Planning*. Menlo Park, CA: AAAI Press, 1998, pp. 23–3.

Affordance based imitation bootstrapping with motionese

Emre Ugur[1,2,3], Yukie Nagai[4], and Erhan Oztop[2,5]

[1] Institute of Computer Science, Univesity of Innsbruck, Innsbruck, Austria
[2] ATR, Dynamic Brain Imaging, Kyoto, Japan
[3] NICT, Advanced ICT, Kyoto, Japan
[4] Graduate School of Engineering, Osaka University, Osaka, Japan
[5] Ozyegin University, Istanbul, Turkey

Emails: emre.ugur@uibk.ac.at, yukie@ams.eng.osaka-u.ac.jp, erhan.oztop@ozyegin.edu.tr

***Abstract*— Learning through self-exploration and imitation are crucial mechanisms in developing sensorimotor skills for human infants. In our previous work, we showed that a robot can self-discover behavior primitives and learn object affordances similar to infants. Then building predictive mechanisms at the affordance level allowed movement planning and simple goal-level imitation on our robot. The work we describe in this paper builds upon this system; it describes how our robot, by using the learned behaviors and prediction mechanisms, can go beyond simple goal-level imitation and become a better imitator. For this, we develop mechanisms to enable the robot to recognize and segment, with the help of the demonstrator, an ongoing action in terms of its affordance based goal satisfaction. Extracting sub-goals or important features from a demonstration is not straightforward as demonstrated action trajectory may not correspond to any robot behavior developed so far. Inspired from infant development, we use motionese to enable the robot to identify the important steps and the boundaries in the otherwise complex stream of motion involving multi-objects. Once the sub-goals are obtained, the robot imitates the observed action by chaining these sub-goals and satisfying them sequentially. In the experiments, a tutor used his own action repertoire to move an object around, and the robot was able to detect and achieve the sub-goals with its affordance prediction mechanisms and behavior primitives.**

I. INTRODUCTION

Infants develop cognitive abilities in a progressive and staged way. In the initial months of their life, they learn discriminating and discovering new behavior primitives such as grasp, shake, and hit from simple actions and reflexes, e.g, reaching and palmar-grasp [1]. Next, they use these behaviors to explore the environment and interact with the objects [2]. It is plausible to think that while interacting with the environment, babies monitor the consequences of their actions and relate the consequences to the visual properties of the objects they interact with. In other words, they learn object affordances, i.e. the action possibilities offered by their environment [3]; and learn to differentiate ends from means. Goal-emulation, a form of imitation characterized by the replication of the observed end effect, starts after this period, and infants become skilled at imitating unseen movements after 12 months of age [4]. Infants' means of imitation changes over time; while younger infants are more inclined in achieving the goal of a demonstrated action, older infants tend to exactly imitate (and in later stages over-imitate) the observed target action sequence even if those actions are not physically related to the goal [5].

In our previous work, we showed that a robot can self-discover behavior primitives and learn object affordances similar to infants. This was achieved by first differentiating the actions based on the differences in tactile perception [6], and then by learning the relations between objects, behaviors and effects created in the environment through physical interaction [7]. After learning, the robot was able to make plans to achieve desired goals, emulate end states of demonstrated actions, monitor the plan execution and take corrective actions using the perceptual structures employed or discovered during learning. However obtaining more complex sensorimotor skills such as learning multi-object affordances and learning certain delicate tasks may require support from outside, especially when not only 'ends' but also the 'means' is important.[1]

In this paper, the next stage of developmental progression is studied in the form of imitation learning where the affordance prediction capability for single-objects does not suffice to reproduce complex movements. These tasks are taught to a robot through imitation, where the robot observes the demonstration, extracts important steps from the movement trajectory, encodes those steps as sub-goals and find the behaviors to achieve these goals. Learning higher level skills based on previously learned simpler ones is more economical and usually easier for building a complex sensorimotor system[8], [9]. Therefore, in our study learned affordance perception and behavior primitives are used as basic elements in understanding and achieving sub-goals.

[1]Imitation refers to finding the behavior sequence that enables the robot to follow a similar trajectory with the demonstration. Goal emulation on the other hand refers to computing the behavior sequence to achieve the goal independent of the demonstration. Even young infants are good in goal-emulation as probably goal-emulation does not require understanding of demonstrator's intent or exact action sequence. It is sufficient to extract only the final situation to map it to a goal state and to find the behavior sequence from its own repertoire to achieve this goal. Similar to younger infants, goal emulation is easier in the earlier stage of our robot's development. On the other hand, imitation requires representing and mapping important features of the complete action trajectory, which is accomplished in the later phases of development.

Extracting sub-goals or important features from a demonstration is not straightforward as demonstrated action trajectory may not correspond to any robot behavior developed so far. For example when the robot is asked to achieve a task, the observed trajectory may not be represented in robot's sensorimotor space, and executing the behavior that seemingly achieves the goal would not satisfy the imitation criteria. Infants also have similar difficulties in mapping observed actions within their own repertoire and in imitating these actions successfully.

To overcome this difficulty, parents are known to make modifications in infant-directed actions, i.e. use "motionese" [10], [11]. Motionese is characterized by higher range and simplicity of motion, more pauses between motion segments, higher repetitiveness of demonstration, and more frequent social signals to an infant [10], [12]. Fine-grained analysis using a computational attention model further reveals the role of motionese in action learning [13]. Longer pauses before and after the action demonstration underline the initial and final states of the action (i.e. the goal of the action) whereas shorter but more frequent pauses between movements highlight the sub-goals of the action [14]. Of particular interest is that such modifications are elicited by the responses of an action learner [15]. Not only the age of a learner but also the ability to recognize the demonstrated action (i.e. visual attention) influences the task demonstration.

Inspired from infant development, in this paper we also use 'motionese' to enable the robot to identify the important steps and the boundaries in the otherwise complex stream of motion involving multi-objects. A human tutor can exaggerate the relevant features in his demonstration and enable the robot to map the exaggerated sub-steps into its own behavior repertoire and imitate the action sequence successfully.

The idea of using previously developed capability for affordance prediction in imitation learning is not new. [16] also used affordances that are modeled in Bayesian Networks to interpret demonstration and imitate using the robot's own behaviors. While they were able to recognize one-step actions and make one-step predictions, our robot can make multi-step plans for goal emulation as our affordance framework supports multi-step prediction. However more importantly, our system, with the reported work in the current paper, can extract multi-step behaviors from the demonstration that may include multi-objects. Moreover different from other studies [17], as we follow a developmental approach, our system would support naive tutors who can naturally adapt demonstrations based on the robot's imitation performance and use some 'motionese' features enabling the robot to imitate complex actions.

The rest of this paper is structured as follows. In the next section, we give the physical properties of the experimental platform, and detail the robot's perceptual and motor abilities. Next, we describe the affordance framework that enables next state prediction, plan generation, goal emulation and imitation. In Section IV, the experiment with an experienced tutor is given where 'motionese' based demonstration is shown to be effective in imitating complex action sequences.

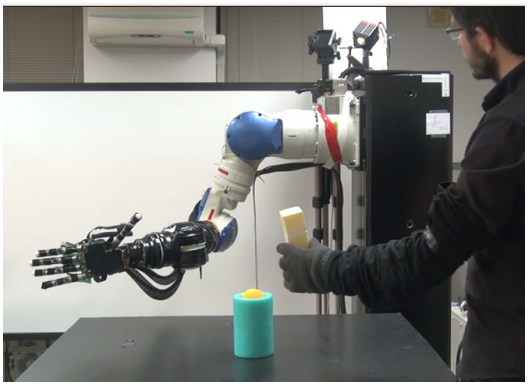


Fig. 1. The robot hand, arm, range camera (top-right), some objects and a tutor constitutes the experimental setup.

II. EXPERIMENTAL FRAMEWORK

A. Robot System

An anthropomorphic robotic system equipped with a range camera is used as the experimental platform. This system employs a 7 DOF Motoman robot arm, that is placed on a vertical bar similar to human arm as shown in Fig. 1. A five fingered 16 DOF Gifu robot hand is mounted on the arm to enable manipulation. The maximum length of Motoman arm and Gifu hand is 123 cm and 23 cm, respectively. For environment perception, an infrared range camera (SR-4000), with 176x144 pixel array, 0.23° angular resolution and 1 cm distance accuracy is used.

B. Perception

a) Object Detection: The robot's workspace consists of a black table, a human demonstrator's arm and hand, the robot's own actuator and colored objects. The demonstrator's and robot's actuators are also covered with black material to distinguish the objects easily. The region of interest is defined as the volume over the table, and black pixels are filtered out as the range readings from black surfaces are noisy. As a result, the remaining pixels of the range image are taken as belonging to one or more objects. Because the range camera does not provide reliable color information, the objects are segmented using Connected Component Labeling algorithm [18] based on the depth information.. In order to reduce the effect of camera noise, the pixels at the boundary of the object are removed, and median and Gaussian filters with 5x5 window sizes are applied. Finally, a feature vector for each remaining object is computed using the 3D positions obtained from the depth values of the corresponding object pixels as detailed in the next paragraph.

b) Object feature vector computation: The perception of the robot at time t is denoted as $[\boldsymbol{f}_{o_0}^{t,()}, \boldsymbol{f}_{o_1}^{t,()}..]$ [2] where $\boldsymbol{f}$ is a feature vector of size 43, and the superscript $^{()}$ denotes that no behavior has been executed on the object yet. Four channels of information are gathered and encoded in a feature vector for each object o_i. The first channel consists of *object*

[2]Note that t and o_i are sometimes omitted in the rest of the text in order to ensure easy readability of the notation.

visibility feature which encodes the knowledge regarding the existence of the object. The second and third channels correspond to the *object position* and *object dimensions* in the coordinate frame shown in Fig. 1. The fourth channel encodes the shape related features, where the distribution of the local surface normal vectors are used. Specifically histograms of normal vector angles along the latitude and longitude, 18 bins each, are computed and used. The final feature vector is:

$$\boldsymbol{f}^{()} = [vis, \boldsymbol{pos}, \boldsymbol{dim}, \boldsymbol{hist}_{\text{latitude}}, \boldsymbol{hist}_{\text{longitute}}]$$

C. Behaviors

The robot interacts with the objects using three different behaviors, namely *grasp*, *release*, and *push* where the object center position is used to guide the execution. For any behavior, *initial*, *target* and *final* hand positions are computed with an offset to the object center, and a trajectory that passes through these three points are executed as follows:

- *Initial* position is the offset from the object where the robot places its hand prior to interaction. This parameter is fixed and same for all behaviors, and takes the robot hand to the back-right diagonal of the object from the robot's perspective. If the object to be interacted is already in robot's hand, the initial position is set as the current position of the robot hand as there is no need to re-position the hand.
- *Target* is the offset from the object-center that determines which part of the robot's hand makes contact with the object. Using this parameter which is unique for each behavior, the robot touches to the object with its palm in *grasp* and *release* behaviors, and with its fingers in *push* behavior.
- *Final* position is the offset from the object where robot brings its hand at the end of the behavior execution. Final position can be set to any arbitrary point in the robot's workspace except very close to the table to avoid any collision.
- *Hand-close* and *hand-open* positions are again offsets from the object. The hand clenches into a fist with *grasp* and *release* behaviors when it is close to the object center, and wide-opens with *release* behavior at the end of action execution. *Push* behavior does not change hand-state unless the object is already in the robot's hand. In this case the hand wide-opens in the beginning.

With appropriate ranges of the parameters, target objects can be grasped, released or pushed to different locations depending on the *final* position. While this set sounds like a minimal behavioral repertoire that is manually coded, it is not. On the contrary, these behaviors along with their parameters were transfered from a previous developmental stage, namely behavior formation phase, where they were discovered through exploration using palmar-grasp reflex and a simple *reach* action that modeled a one-two months old infant's seemingly non-purposeful hand babbling [6]. Object-palm contact which precedes hand flexion, and object-finger contact without any hand-closing during push behavior were direct consequences of this behavior formation phase.

In the rest of the paper, the behaviors are represented as $b_{\{target,open,close\}_j}(pos_{final})$ or $b_j(\boldsymbol{\rho}_f)$ in short, where j denotes the behavior type and $\boldsymbol{\rho}_f$ denotes the relative *final* position of hand.

III. MOVEMENT GENERATION BASED ON AFFORDANCE PERCEPTION

Imitation and goal emulation are achieved by finding behavior sequences that will bring the initial state ($\mathcal{S}_{\text{init}}$) to the goal state ($\mathcal{S}_{\text{goal}}$) depending on or independent of demonstration, respectively. For this purpose, the robot should have the ability to predict the effects of its behaviors on the objects, i.e. it should be able to predict the next state ($\mathcal{S}_t$) for any behavior executed in a given state ($\mathcal{S}_t$). In this section, we present the structures and methods that enable imitation and goal emulation.

A. Affordances and Effect Prediction

In our previous work [7], [6], the affordances were defined as (object, behavior, effect) relations, and with this we have shown that affordance relations can be learned through interaction without any supervision. While visibility and position features could be accurately predicted for any behavior, the change in other features such as dimension or shape could not be predicted reliably. As learning the prediction ability is not focus of this paper, we skip the details and shortly present how prediction operator works. This prediction operator can predict the effect on position and visibility features given object feature vector, behavior type and behavior parameters:

$$(\boldsymbol{f}^{()}, b_j, \boldsymbol{\rho}_f) \rightarrow \boldsymbol{f'}[v, \boldsymbol{p}]^{b_j}_{\text{effect}} \quad (1)$$

where $\boldsymbol{f'}[v, \boldsymbol{p}]^{b_j}_{\text{effect}}$ denotes the predicted ($'$) effect in visibility and position features.

B. State Transition

The state corresponds to the list of feature vectors obtained from the objects in the environment:

$$\mathcal{S}_0 = [f^{()}_{o_0}, f^{()}_{o_1}, .., f^{()}_{o_m}]$$

where $()$ denotes the zero length behavior sequence executed on the objects, and m is the maximum number of objects. If the actual number of objects is less than m, the *visibility* features of non-existing objects are set to 0.

State transition occurs when the robot executes one of its behaviors on an object. Only one object is assumed to be affected at a time during the execution of a single behavior, i.e. only the features of the corresponding object is changed during a state transition. The object monitoring module always makes sure that the object indexes are set correctly during interactions. Furthermore as shape and dimension features cannot be predicted reliably, they are assumed to remain fixed. Thus, the next state can be predicted for any behavior using the prediction scheme given in Eq. (1) as follows:

$$\mathcal{S}'_{t+1} = \mathcal{S}_t + [...0, \boldsymbol{f'}_o[v, \boldsymbol{p}]^{b_j}_{\text{effect}}, 0, ..] \quad (2)$$

where b_j behavior is executed on object o and only *visibility* and *position* components of this object change by the summation operator.

Using an iterative search in behavior parameter space, the robot can also find the best behavior and its parameters that is predicted to generate a desired (des) effect given any object:

$$bb(\boldsymbol{f}^{()}, \boldsymbol{f}[v,\boldsymbol{p}]^{\text{des}}_{\text{effect}}) = \arg\min_{b_j,\boldsymbol{\rho}_f}(\boldsymbol{f}[v,\boldsymbol{p}]^{\text{des}}_{\text{effect}} - \boldsymbol{f'}[v,\boldsymbol{p}]^{b_j}_{\text{effect}}) \quad (3)$$

where bb denotes "best behavior" operator.

C. Goal-emulation and plan generation

In the previous section, how the robot can (1) predict the effect given object-behavior pair and (2) find the best behavior to acquire a desired effect were explained. Because prediction is based on vector summation, the robot can estimate the total effect that a sequence of behaviors will create by simply summing up all effect vectors, and thus can use this for multi-step prediction.

Goal-emulation is achieved by generating a plan, i.e. finding the behavior sequence required to transform the given state into the goal state. In this study, forward chaining is used to search the state space and find a sequence (see Fig. 2 right). Forward chaining uses a tree structure with nodes holding the perceptual states and edges corresponding to (behavior-object) pairs. The execution of each behavior on each different object can transfer the state to a different state based on Eq. (2). Starting from the initial state encoded in the root node, the next states for different behavior-object pairs are predicted for each state. In order to reduce the search time, the states with minimal distance to the goal state are expanded first.

The goals are represented as desired world states, however as only the position and the visibility can be predicted, the goal representation only includes *position* and *visibility* features:

$$G = [pos_{o_1}, vis_{o_1}, pos_{o_2}, vis_{o_2}..]$$

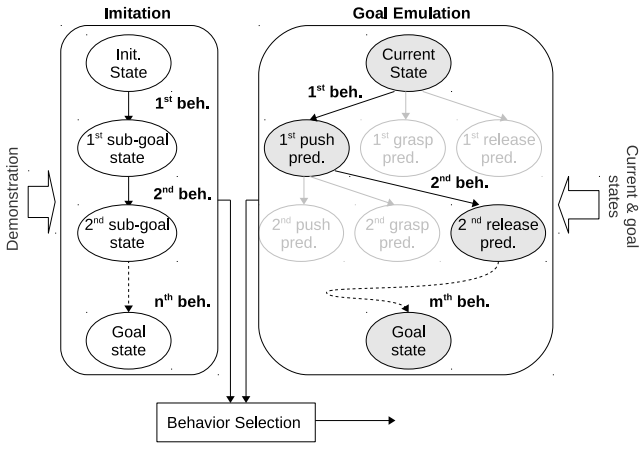


Fig. 2. The robot can choose to follow the demonstration by executing the behaviors from Imitation module or can find the sequence of behaviors using Goal Emulation module in any step to reach to the desired state.

D. Imitation through scaffolding

The robot observes the demonstration and extracts the initial and goal states, as well as the intermediate states (encoded as sub-goals) by detecting pauses which may be introduced by a motionese engaged tutor. If no pause can be detected, then a random intermediate state would be picked up as the sub-goal state.

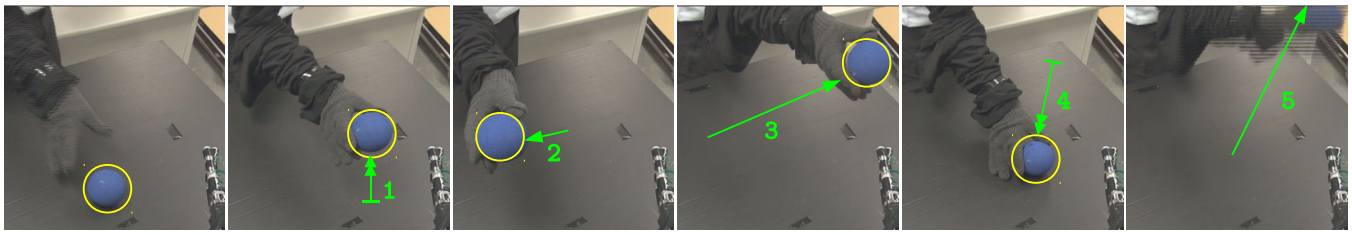
Imitation module (Fig. 2, left panel) finds the behavior sequence that brings the initial state to the goal state following the detected state sequence. Finding the behavior that transfers one observed state to the next observed state corresponds to one-step goal-emulation, which the robot can perform as described in Section III-C and illustrated in Fig. 2, right panel. Thus, imitating the behavior sequence practically corresponds to applying goal-emulation for each successive sub-goal extracted from the observed demonstration. Therefore, the arrows that connect successive sub-goal states in Imitation Module (Fig. 2, left panel) uses goal-emulation mechanism as a subroutine. In the experiments reported in this article, each arrow happened to correspond to a single-affordance perception, i.e. each state transition was achieved by one behavior. However, our framework is not limited with this, and in fact it can find multi-step plans for reaching individual sub-goals.

Selecting behaviors based on Imitation Module results in following the exact trajectory of the demonstrator, i.e. achieving all detected sub-goals using robot's behavior repertoire, to the extent that as decimated by the pauses inserted by the tutor. On the other hand, when Goal Emulation Module (Fig. 2, right panel) is selected, then it finds a behavior sequence that brings the current state to the goal state using forward chaining independent of the intermediate states.

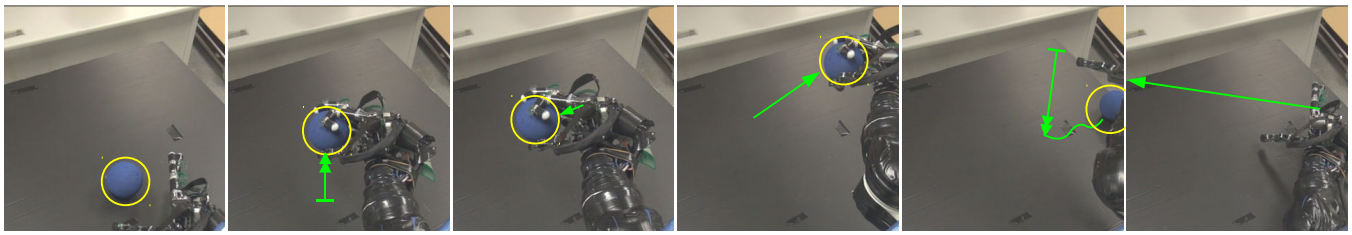
Using different approaches (modules) to imitate an observed action has its own advantages and disadvantages. If the tutor has engaged in a motionese based interaction with the robot, and provides sufficient cues to the robot, Imitation Module would make complex imitation possible. However this requires keeping all sub-goals in the memory of the robot cognitive system, and executing all corresponding actions, which might not be practical if the tutor makes a large number of pauses. Furthermore, the Imitation Module needs additional mechanisms to deal with failures during execution, and to take corrective actions. On the other hand, Goal Emulation needs no additional mechanisms as it can automatically recover by simply reassessing the current state and re-planning. However it may fail in multi-object environments as predictions are made based on single-object affordances. In the current implementation, we followed a simple approach where the Behavior Selection Module simply reflects the choice of the experimenter that is conveyed through a Graphical User Interface. Work is underway to make the system autonomously choose the appropriate execution mechanisms depending on the demonstration.

IV. EXPERIMENTS

In this experiment, we verify our imitation system by showing an action trajectory to the robot with several pauses



(a) Demonstration of the experimenter



(b) Imitation of the robot

Fig. 3. Snapshots where the robot imitates the demonstration performed by an experienced tutor. In (b) from left to right, grasp-up, grasp-left, grasp-right, release, and push actions are planned and executed. Please refer to Section IV for details.

and letting the robot to imitate our demonstration. The demonstration was performed by an experienced tutor who knows the working principles of the system. The tutor used his own action repertoire to move the object around, while the robot monitored the object, detected the pauses, and encoded these pauses as intermediate states. Following the demonstration, the robot found the sequence of behaviors to obtain the intermediate states in sequence using the imitation procedure summarized in Fig. 2 and behavior selection procedure given in Eq.(3).

The humans demonstration and imitation of the robot are shown in Fig. 3. From the robot's point of view, the object was (1) lifted up, (2) moved to left, (3) moved to right in the air, (4) put on the table, and (5) removed from the table. The robot, after detecting the intermediate states, computed a behavior sequence with the following behaviors:

1) *grasp* $(-3, +21, -1)$
2) *grasp* $(+9, +13, -3)$
3) *grasp* $(-21, +5, +4)$
4) *release* $(+4, 0, +4)$
5) *push* $(+10, +4, 0)$

Execution of these behaviors resulted in a (sub-goal) trajectory that is similar to the demonstration with the following exceptions. First of all, the object was not moved exactly to the same positions in intermediate steps because of the noise in the perception and due to kinematic constraints of the robot. Second, after the object was released, it rolled over the table (to the right) and did not end up exactly below the hand as the robot predicted. Third, push behavior could not roll the object off the table without (the experimenter) bending the table as the push was not strong enough to cause a high-speed roll. Still, the robot was able to accomplish the task by achieving the observed sub-goal changes using its own behaviors. For example, the tutor's 'removal' of the object from the table was mapped to *push* behavior of the robot as the object was rollable, and *push* applied to a rollable object was predicted to make the object disappear. As another example, 'putting on table' action performed by the tutor was mapped to robot's *release* behavior which has similar effect.

As we presented in Section III.C, our goal-emulation system supports multi-step planning and execution of a behavior sequence to achieve the desired goal. However, in this particular experiment, achieving each subgoal was possible with execution of one behavior primitive. While the the full power of multi-step affordance prediction was not explicitly demonstrated in this experiment, the focus of this paper was to show the feasibility of imitation through goal emulation rather than providing an integrated and complicated scenario with complex imitation and planning. Readers interested in goal-emulation through plannning based on multi-step affordance prediction can refer to [7] where a 7-step plan was generated and executed to bring an object to an observed goal position.

V. CONCLUSION

This study addressed how a robot following a development learning approach for its motor repertoire formation and affordance based planning can benefit a tutor who naturally adapts his demonstration to teach the robot. The limited capacity of the robot to imitate triggers a change in the tutor to modify his demonstration. In infant development this is often called motionese which captures the general notion that when caregivers teach new skills to infants through demonstration, they often demarcate important parts of their action by using pauses, sharp movements, repeats and attention grabbing signals.

In this study we aimed to realize such a learning scenario on a physical robot. For this, we built upon the development framework we have developed, which equipped our robot - via learning- with motor behaviors [6] and affordance based prediction capabilities [7]. We have shown that our robot could perform goal level imitation, which is called as 'goal

emulation'. In this mode of operation, only the final goal of the action matters; not how it is achieved. For example a demonstration of grasping and taking away a ball from the view of the robot can be emulated by our robot with a push behavior (as this will roll off the ball from the table). In the current study, we augmented this with a general interactive imitation learning system that naturally engages the tutor in motionese, in which the tutor demarcation was used to chop a complex action into simpler pieces that can be handled by goal emulation. This allows the robot to perform tasks that otherwise would be impossible to be executed via simple goal emulation.

To enable such functionality, first an interactive system is created, where the robot tracks the movement of objects and tries to segment parts of the dynamic scene in terms of changes in its affordance space. Once the robot can perceive and hence represent some part of the demonstrated action in terms of changes in its percept, it can reproduce that part with goal emulation. Chaining of such segmentation and goal level imitation then enables the robot to imitate the observed action. This may not always be the exact copy of the demonstrators action, but may get closer to it if the tutor is not satisfied with the imitation of the robot and introduces more motionese markers to help the robot.

As a next step, we plan to recruit naive subjects without any prior knowledge about this research as tutors. We postulate that the developmental system can naturally engage the naive tutors to modify their movements so as to make the robot understand their actions as the motionese theory predicts. In order to achieve such performance, we can extend the current system by including other motionese cues such as repeating behavior, and allowing the robot directly signal the tutor that he understood a specific portion of the demonstrated action.

This framework should be extended in different directions for better and more natural skill acquisition. For a truly developmental system, the imitated action sequences should be integrated to behavior primitive space of the robot in a seemless way. The robot also needs to give some feedback during probably initially failing demonstration attempts of naive tutors so that the tutors can easily learn how the robot learns to imitate. One other limitation is that our current system sees the world as changes happening on the object percept. So it cannot really do imitation of an action which does not include objects. The relation between object-based and object-free imitation mechanisms is still an open question and promising direction for the future.

ACKNOWLEDGMENT

This research was partially supported by European Communitys Seventh Framework Programme FP7/2007-2013 (Specific Programme Cooperation, Theme 3, Information and Communication Technologies) under grant agreement no. 270273, Xperience. This research was partially supported by a contract in H23 with the Ministry of Internal Affairs and Communications, Japan, entitled 'Novel and innovative R&D making use of brain structures'.

REFERENCES

[1] J. Piaget and B. Inhelder, *The Psychology of the Child*. New York, USA: Basic Books, 1966.

[2] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: a survey," *IEEE Tran. Auton. Mental Dev.*, vol. 1-1, 2009.

[3] J. J. Gibson, *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, 1986.

[4] B. Elsner, "Infants' imitation of goal-directed actions: the role of movements and action effects," *Acta psychologica*, vol. 124, no. 1, pp. 44–59, 2007.

[5] C. Huang and T. Charman, "Gradations of emulation learning in infants: imitation of actions on objects," *Journal of Experimental Psychology*, vol. 92, no. 3, pp. 276–302, 2005.

[6] E. Ugur, E. Sahin, and E. Oztop, "Self-discovery of motor primitives and learning grasp affordances," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 3260–3267.

[7] E. Ugur, E. Oztop, and E. Sahin, "Goal emulation and planning in perceptual space using learned affordances," *Robotics and Autonomous Systems*, vol. 59, no. 7–8, pp. 580–595, 2011.

[8] M. Haruno, D. Wolpert, and M. Kawato, "Hierarchical mosaic for movement generation," in *Excepta Medica International Coungress Series*. Amsterdam, The Netherlands: Elsevier Science B.V., 2003, pp. 575–590.

[9] M. Kawato and K. Samejima, "Efficient reinforcement learning: computational theories, neuroscience and robotics," *Current Opinion in Neurobiology*, vol. 17, pp. 205–212, 2007.

[10] R. Brand, D. Baldwin, and L. Ashburn, "Evidence for 'motionese': Modifications in mothers infant-directed action," *Developmental Science*, vol. 5, pp. 72–83, 2002.

[11] R. J. Brand, W. L. Shallcross, M. G. Sabatos, and K. P. Massie, "Fine-Grained Analysis of Motionese: Eye Gaze, Object Exchanges, and Action Units in Infant-Versus Adult-Directed Action," *Infancy*, vol. 11, no. 2, pp. 203–214, 2007.

[12] K. J. Rohlfing, J. Fritsch, B. Wrede, and T. Jungmann, "How can multimodal cues from child-directed interaction reduce learning complexity in robots?" *Advanced Robotics*, vol. 20, no. 10, pp. 1183–1199, 2006.

[13] Y. Nagai and K. J. Rohlfing, "Computational Analysis of Motionese Toward Scaffolding Robot Action Learning," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 44–54, 2009.

[14] Y. Nagai, "From Bottom-Up Visual Attention to Robot Action Learning," in *Proc. of the 8th IEEE International Conference on Development and Learning*, 2009.

[15] Y. Nagai, A. Nakatani, and M. Asada, "How a robots attention shapes the way people teach," in *Proc. of the 10th International Conference on Epigenetic Robotics*, 2010, pp. 81–88.

[16] M. Lopes, F. Melo, and L. Montesano, "Affordance-based imitation learning in robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 1015–1021.

[17] K. Mochizuki, H. Nobuta, S. Nishide, H. G. Okuno, and T. Ogata, "Developmental human-robot imitation learning with phased structuring in neuro dynamical system," in *IROS2012 Workshop on Cognitive Neuroscience Robotics*, 2012.

[18] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision, Volume I*. Addison-Wesley, 1992.

Integrating Cognitive Control and Imitation Learning in a Socially Situated Robot

Huan Tan, D. Mitchell Wilkes, *Member, IEEE,* and Kazuhiko Kawamura, *Fellow, IEEE*

***Abstract*— This paper proposes a control architecture for a socially situated robot by combining motion recognition and generalization, intention recognition and cognitive control. This architecture provides a method for a humanoid robot to generalize common features of human motions through observation and store them in the long-term memory, to recognize a particular human motion to estimate his/her intention and modify its own arm movement. Several experiments were carried out on a humanoid robot to demonstrate the proof of concept.**

I. Introduction

In the future, it is expected that humanoid robots will be deployed increasingly at work, in homes and public spaces to assist humans in a variety of tasks. For socially situated robots, it will be necessary to be able to recognize and imitate human behaviors by understanding human intentions and modifying their own behaviors accordingly. Recent advances in the cognitive neuroscience have considerably enlarged our understanding of social cognition of human actions [1]. The actions we perform are usually driven by prior intention. For example, a person grasping a cup may grasp it in order to drink from it, or to hand it to another person. Is it then possible to anticipate what he/she is going to do next from the way he/she reaches and grasps the cup? It is argued by some that it is possible to understand the intentions of others by simply observing their movements [2] [3], while others are more skeptical [1]. In this paper, we will focus on robotic behavior modifications based on social cognition using intention recognition, robotic imitation learning, and cognitive control.

The study of robotic imitation learning meantime attempts to enable robots to learn low-level behaviors from demonstration and to apply them in different situations [3] [4]. Current imitation learning methods mimic human motion trajectories [5] and couple them into sequences [6] for robots to perform simple tasks such as reaching or grasping. Recently, we proposed a framework for imitation learning that uses a number of techniques to enable a robot to perform learned tasks in situations that differ, to a small extent, from those in which it was learned [7].

This paper describes a robotic behavior modification technique based on human intention recognition, imitation learning, and cognitive control. The rest of the paper is organized as follows: Section II summarizes related work on imitation learning, cognitive control, and intention recognition. Section III proposes a robotic behavior modification framework based on situation awareness. Section IV describes experiments performed. Section V analyzes the experimental results and summarizes the contributions of this paper.

Tan was with Electrical Engineering and Computer Science Department, Vanderbilt University Nashville, TN 37212 USA (e-mail: huan.tan@ieee.org).

Wilkes, and Kawamura are with Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN 37212 USA (e-mail: {huan.tan, mitch.wilkes, kaz.kawamura}@vanderbilt.edu).

II. Related Work

A. Imitation Learning

Current research on imitation learning can be divided into two categories [8]: One tries to train robots to replicate motion dynamics [5], and the other is to train robots to learn action primitives and higher-level behaviors [6] [9].

Most researchers in the imitation learning community have worked on how to adaptively generate motions which are similar to demonstrations. The Dynamic Movement Primitives (DMP)-based method is well accepted as a general behavior generation method [5]. In the DMP-based imitation learning, the robot learns a non-linear model through a regression process or by trials. Optimal control [10] [11], and Reinforcement Learning [12] are typically applied to train the robots. Others include the "Lagrange method" that tries to minimize the error between the demonstration and the generated trajectory in both the original data space and the low-dimensional latent space [8] [13] [14].

Recent applications include grasping [11] [15] [16] [17], and manipulation [18].

B. Cognitive Control

Cognitive (or executive) control is a term borrowed from cognitive neuroscience, and refers to processes "that allow information processing and behavior to vary adaptively from moment to moment depending on current goals, rather than remaining rigid and inflexible." [19]. Cognitive control processes include a broad class of mental operations including context understanding and goal representation, and strategic processes such as attention allocation and high-level stimulus-response mapping [19]. Application of cognitive control to robotics was first proposed by Kawamura and Gordon in 2006 [20]. In the subsequent years, a variety of robotic architectures have been proposed [21] [22] [23]. Cognitive control regulates robot behaviors through analysis of situational information such as attention [21] and intention [24] [25].

C. Intention Recognition

The actions we perform in daily life are usually driven by either an explicit or a hidden intention. In [2], Becchio et al argued that understanding others' intentions by observing their arm movements is possible. In [24], Sartori et al studied

how recognition of the intentions of a human observer can influence and modulate the actions of the human performing a task. Researchers in human-robot cooperation used a number of statistical techniques to recognize human intention. For example, Aaro and Kragic [25] and Kelley et al [26] used Hidden Markov Models to recognize human intention by a robotic agent. Schrempf et al. used dynamic Bayesian Networks in so-called proactive human-robot cooperation [27].

In our cognitive control system, we intend to use a social cue such as extending an open hand to recognize the intention of a human observer to modulate the behavior of a robot. In our experiments we will consider three basic possibilities: no detectable action by the human, an action that is detected but is not understood to convey any type of social cue (extending a closed hand) and a detected action that does convey a social cue (extending an open hand). Clearly, for sophisticated activity the robot would need to detect and understand a potentially large number of actions and social cues, however, for the purposes of illustrating and demonstrating the combination of intention recognition and cognitive control we only require a basic set of actions and cues.

III. METHODOLOGY

The modular approach adapted in this paper is based on the cognitive control architecture developed in our lab [21].

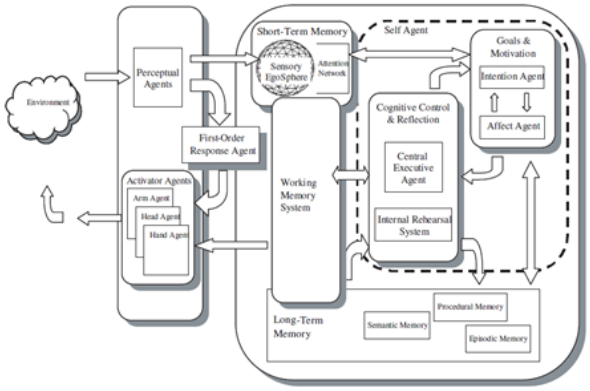


Fig.1 Cognitive Robot Control Architecture

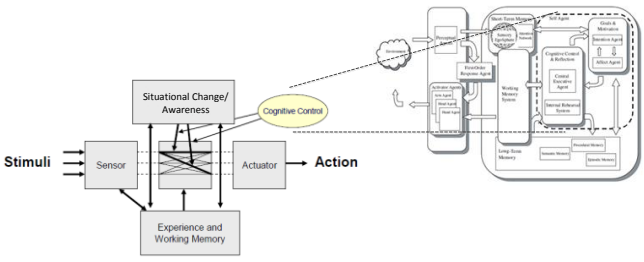


Fig.2 Cognitive Control for Social Robots

It is shown in Fig.1. The cognitive control portion as applied to social robots is shown in Fig. 2. Key modules for robot behavior learning and modification are shown in Fig. 3.

In the rest of this section, key modules are described in detail.

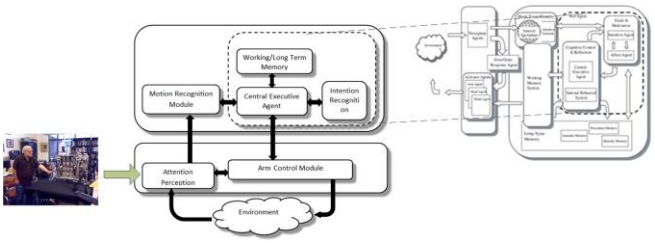


Fig.3 Robot Behavior Learning and Modification System

A. Perception/Attention Module

The system uses observed social behaviors as the basis for generating motions. First the robot observes human motions in the situated social setting. Observed motion trajectories and task-relevant information (e.g., the distances between the end-effector and a target object in a manipulation task) are recorded for modeling and analysis. Observed motion data is recorded with time-stamped vectors.

In our system, a Kinect sensor was used to record the arm trajectory and the position and orientation of the hand of a human. In the learning stage, various motions of the right arm/hand of a human is recorded and used by the robot to generalize and learn the observed behaviors.

For the purposes of the concept of proof, it was decided to use a relatively basic perception and attention model to illustrate the integration of the intention (i.e. opening of a hand) recognition into a cognitive control framework. The robotic agent monitors, perceives and attends to three different situations. In the first, the human observer is not moving his arm, thus no intentional gesture is present. In the second, the observer is moving his arm with his hand closed, so the motion does not convey any intention. Finally, in the third situation, the observer is moving his arm and opens his hand signaling about the intention of the observer.

B. Motion and Activity Recognition

Tasks humans perform in social settings comprise a set of low-level motions. Different tasks require different sets of motions. Due to measurement errors, noise in the environment and inconsistencies in human in performance of the same task, the obtained motion trajectories will vary. There may be, however, common features within the demonstrations. An appropriate analysis and comparison of observed tasks may find common features hidden in observed motion trajectories and thereby learn motions and behaviors from demonstrations. In Section IV reaching and pushing behaviors are learned in order to perform the experiments.

Our proposed imitation learning method considers motions to be attribute-based. That is, common internal features found for observed behaviors are represented as a set of attributes. A labeled or named motion or behavior can be described in terms of three attributes: (1) the requisite preconditions or task-specific environmental conditions for execution, (2) internal constraints which confine the behavior during execution, and (3) post results that characterize the outcome of a behavior:

$$Behavior = \{name, Precondition, Internal\ Constraints, Post\ Results\} \quad (1)$$

At the behavior generalization stage, the goal is to find the most common features for the pre-conditions, internal

constraints and post results respectively. The design of the required common features is flexible, and researchers can define their own features.

A group of features are predefined and stored in the memory system. We define three groups of features for pre-conditions, internal constraints and post results respectively:

$$X = \{X_{pre}, X_{interna}, X_{post}\} \quad (2)$$

where X_{pre} contains a list of l pre-conditions, $X_{internal}$ contains m internal constraints, and X_{post} contains n post results.

A sample definition table of categories of pre-conditions, internal constraints and post results are pre-defined as shown in Table 1. The goal of the behavior generalization stage is to find suitable conditions, constraints and results for each behavior from demonstrations.

Table 1 Pre-Definition of Pre-Conditions, Internal Constraints and Post Results

Pre-Conditions	Internal Constraints	Post Results
0.Unnecessary	0.Unnecessary	0.Unnecessary
1.Minimize the distance between the hand and the target position	1.Keep similar dynamics	1.Minimize the distance between the hand and the object-related position
2.Keep the distance between the hand and the target position	2. Generate the same trajectory	2.Keep the distance between the hand and the target position
3.Object in hand	3.Keep the distance between the hand and the obstacle larger than a predefined value	3.Grasp the object
4. Object not in hand		4.Release the object

From the demonstrations, we used the criterions in Table 2 and Table 3 to find the most common feature for pre-conditions, post results and internal constraints.

Table 2 Criterions of Pre-Constraints and Post Results

Pre-Constraints/Post Results	Criterions
0.Unnecessary	None
1.Minimize the distance between the hand and the target position	The distance between the hand and the target position is smaller than a threshold value
2.Keep the distance between the hand and the target position	The distance between the hand and the target position is larger than a threshold value
3.Object in hand	The closed signal from the gripper and the distance between the hand and the target position is smaller than a threshold value
4. Object not in hand	The opened signal from the gripper

In Table 2, for feature 1 and 2, the distance d between the hand and the target position is computed directly using the Euclidean distance and then normalized to probability values.

$$F_1 = e^{-\frac{(d-d_0)^2}{k_1}}, l = 1,2 \quad (3)$$

$$F_2 = e^{\frac{(d-d_0)^2}{k_2}} \quad (4)$$

where k_1 and k_2 are normalization parameters.

Features 3 and 4 are determined by the measurements of control signal from the grippers.

Table 3 Criterions of Internal Constraints

Internal Constraints	Criterions
1.Keep similar dynamics	DTW similarities between normalized motion trajectories are larger than a threshold value
2.Generate the same trajectories	DTW similarities between motion trajectories are larger than a threshold value

For feature 1 in Table 3, Dynamic Time Warping (DTW) [28] distances between two demonstrations are computed first. Then we have a matrix to describe the distances.

$$d^{DTW} = \begin{bmatrix} 1 & d^{DTW}{}_{1,2} & \cdots & d^{DTW}{}_{1,D-1} & d^{DTW}{}_{1,D} \\ 0 & 1 & & d^{DTW}{}_{2,D-1} & d^{DTW}{}_{2,D} \\ \vdots & & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & d^{DTW}{}_{D-1,D} \\ 0 & 0 & & 0 & 1 \end{bmatrix} \quad (5)$$

where D is the number of demonstrations. The elements of the first row are normalized using the following equation:

$$\bar{d}^{DTW}{}_{i,j} = e^{-\frac{d^{DTW}{}_{i,j}{}^2}{k_v}} \quad (6)$$

Then the normalized variance of the first row of $\bar{d}^{DTW}$ is computed as the probability score of this feature.

After the behavior generalization stage, a behavior graph can be constructed based on the generalized features to describe the relationships among these generalized behaviors (please see the Appendix). All learned behaviors are represented as vertexes in a behavior graph, and the edges are defined by matching the pre-condition of a behavior and the post result of another behavior. If the robot finds that the pre-condition of a behavior and the post result of another behavior match, an edge is added.

C. Intention Recognition Module

The Intention Recognition Module recognizes the intention of the human in the environment and sends the recognition results to the Cognitive Control Module.

Using the hand gesturing images stored in the LTM, if the hand is closed, the robot ignores it and continues to reach the target object; if the hand is opened, the robot determines that the human wants to greet the robot and modify its motion to greet the human.

The recognized positions of the fingers are represented as:

$$P = \{p_1, p_2, \dots, p_5\} \quad (7)$$

where $p_i (1 \le i \le 5)$ is the position value of the i_{th} finger. The position of the center of the palm is recorded as p_a.

The distances between the fingers and the palm are computed as:

$$d_i = \|p_i - p_a\| \quad (8)$$

where $\|\ \ \|$ is the Euclid distance.

A Gaussian model is constructed from the vector of distances $D = \{d_1, d_2, \ldots, d_5\}$. If the mean of the Gaussian model μ is larger than a chosen threshold value, the hand is considered to be opened; if it is smaller than this threshold value, the hand is considered to be closed. Cerezo's method [29] is used to recognize the hand gesture of the human. For specific technical details, please refer to Cerezo's paper [29].

D. Cognitive Control Module

The robot switches goals or tasks according to the current situation and the results of human intention recognition. By incorporating the sensory information, recognition results, and the stored knowledge in the LTM, robot makes decisions using situational affordances [30]. The following rules illustrate how the CEA will choose appropriate social behavior:

Given a social setting the robot is in, the CEA looks for a social cue in terms of human behavior and related parameters. Then the CEA searches the LTM to find whether the behavior has been learned. The input of the searching is the name of the required behavior, and the output is a returned Boolean value. Based on the Boolean value, the CEA uses the following rules:

If the search result is false, then switch to the learning stage;

If the search result is true, then switch to behavior generation;

If the behavior sequence generation is completed, then switch to motion trajectory generation;

If the motion trajectory generation is completed, then switch to execution;

If the recognized intention is stored in the LTM, then switch to the intention related task.

The first two rules are used to route the current task to behavior learning or generation based on the searching results in the LTM. Through behavior learning, robot learns the motions and the semantic names of the behaviors, and it generalizes the common features of the observed behaviors. The third and fourth rules are related to generating motion trajectories to complete the required task. The fifth rule is used to switch tasks when the human intention has been recognized and found in the LTM.

IV. EXPERIMENTS

In this section, we use three experiments to validate our proposed system. Basically, we want to demonstrate that our robot control system using perception/attention, intention recognition and cognitive control modules are capable of replicating results similar to Sartori's action modification experiments [24] using three scenarios: In the first, the human bystander is simply standing still, and thus no potential intentional gesture motion is present. In the second, the bystander moves his/her arm but the robot does not recognize any intention stored in the long-term memory. Finally, in the third scenario, the bystander is opening his/her hand and this particular arm motion conveys information about the intention of the bystander from the past experience of the robot. Thus the robot changes its behavior and reaches to the bystander.

A humanoid robot, named ISAC, is asked to reach a box on a table to its left or right. Reaching and pushing behaviors are taught by human teachers for ISAC to generalize and are stored in the Long-Term Memory (LTM).

In the first experiment, a human simply stands in front of ISAC. Then ISAC sticks to its original goal of reaching a box on the table.

In the second experiment, the human puts a closed hand in front of ISAC when ISAC tries to reach the box. ISAC however ignore this signal that is not stored in the LTM and sticks to its original goal of reaching a box on the table.

In the third experiment, the human puts an open hand in front of ISAC when ISAC tries to reach the box. Then ISAC recognizes the intention of the human from its past experience stored in the LTM, i.e., he/she wants to shake hands with the robot. Then ISAC switches the task goal and extends the hand to the human.

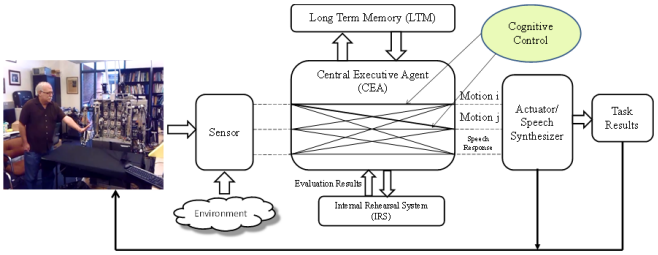


Fig. 4 Cognitive Control Framework

Fig.4 displays the cognitive control framework used in our method.

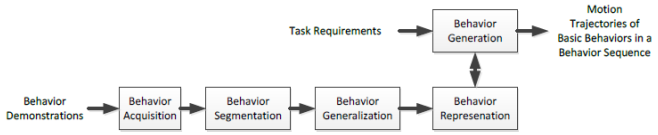


Fig.5 Behavior Learning and Generation Framework

Fig.5 displays the behavior learning, generation, and modification system in our method. It is divided into the following main parts: behavior acquisition, behavior segmentation, behavior generalization, behavior representation, and behavior generation. The "Demonstration Acquisition" block records the motion trajectories of the hands of human teachers. The "Segmentation" block segments the observed behavior sequences into several basic behaviors. The "Goal-Oriented Behavior Generalization" block generalizes the common features of demonstrated motions. The "Behavior Representation" stores the learned knowledge. In the generation stage, given a new human command, the robot generates behavior sequences composed of basic behaviors. Behaviors are modified according to current situation. In this paper, we aim to integrate the two frameworks shown in Fig.4 and Fig.5.

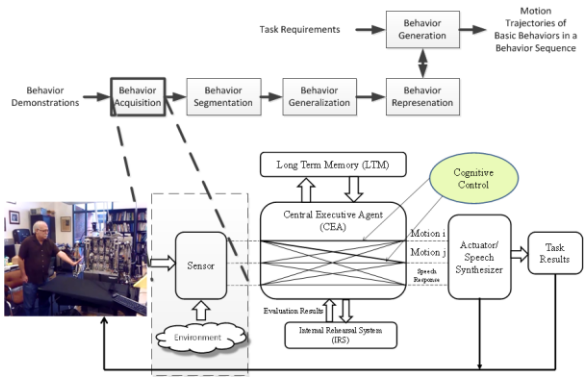


Fig.6 Integration of Behavior Acquisition

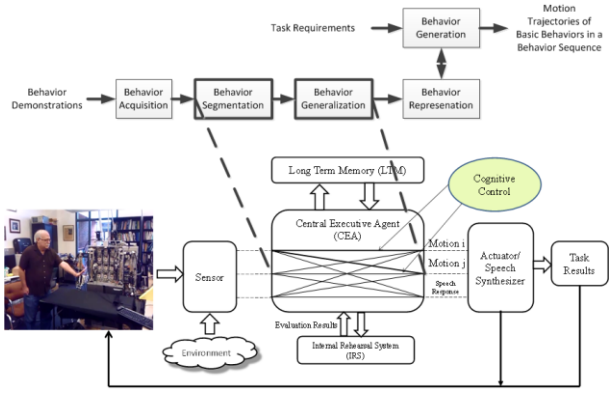


Fig.7 Integration of Behavior Segmentation and Generalization

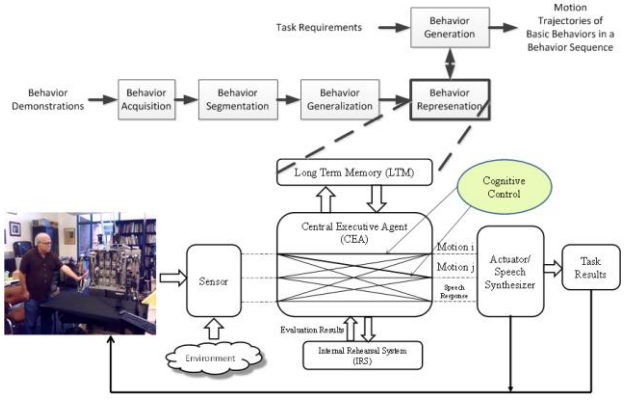


Fig.8 Integration of Behavior Representation

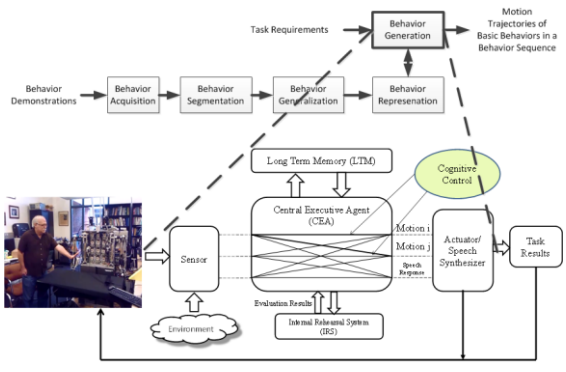


Fig.9 Integration of Behavior Generation

In Fig.6 the Behavior Acquisition of the imitation learning framework is integrated with the Sensor. In Fig.7, the Behavior Segmentation and the Behavior Generalization are integrated with the CEA. In Fig.8, the Behavior Representation is integrated with the LTM. In Fig.9, the Behavior Generation is integrated with the Perception/Attention, the STM, the CEA, the IRS and the Executor.

A. *Experiment 1*

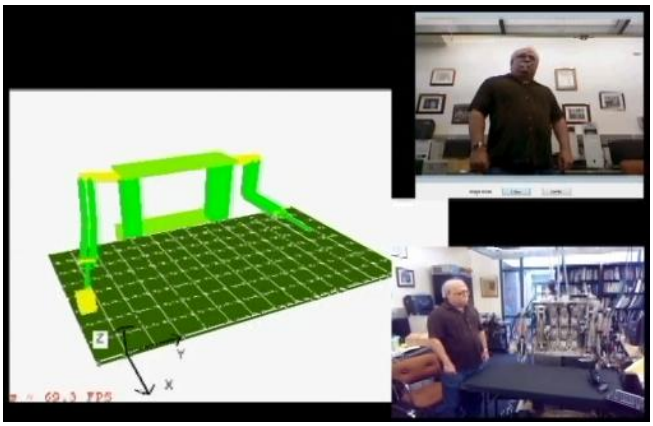


Fig. 10 Results of Experiment 1

Fig. 10 displays the simulation results of experiment 1. In this experiment, the human does not interrupt the action of ISAC and ISAC sticks to its original task goal.

B. *Experiment 2*

Fig. 11 displays the simulation results of experiment 2.

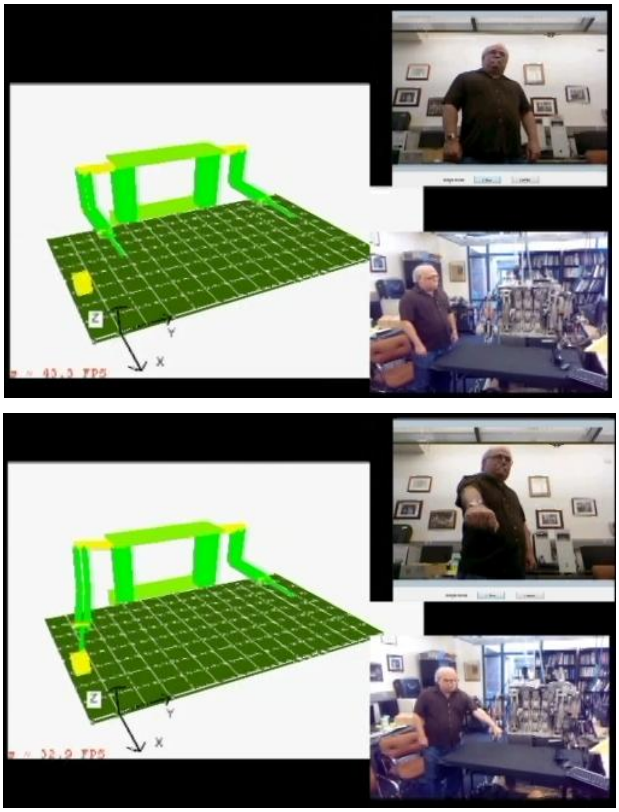


Fig.11 Results of Experiment 2

In this experiment, the human puts his hand in front of ISAC when ISAC tries to reach the box. In Fig.5, the hand is closed which means that the human does not want to have a handshake with ISAC. Then ISAC sticks to its original task goal.

B. Experiment 3

Fig.12-a and Fig.12-b display the simulation results of experiment 3.

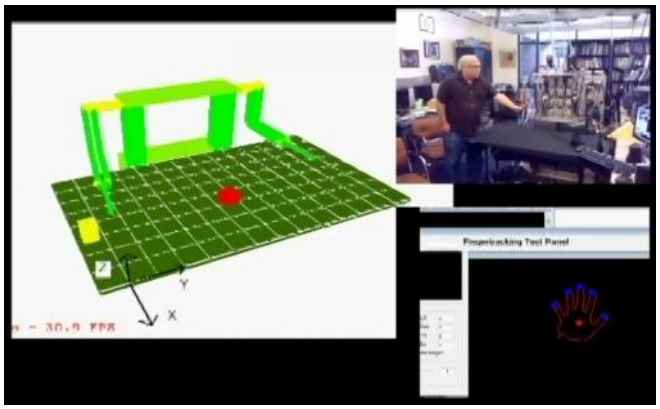


Fig.12-a Results of Experiment 3 (Initial Task)

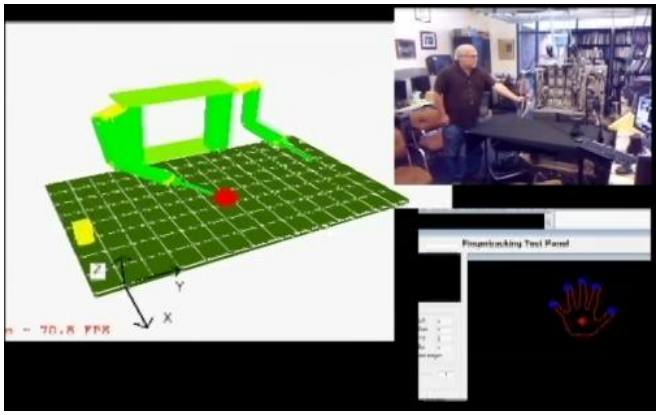


Fig.12-b Results of Experiment 3 (Switched Task)

In this experiment, the human puts his hand in front of ISAC when ISAC tries to reach the box. In the upper picture, the position of the hand of the human is labeled as a red sphere. The contour of the hand of the human has been recognized very well. Since the hand is opened, ISAC considers that the human wants to have a handshake. Then ISAC switches the task goal to reach the hand of the human.

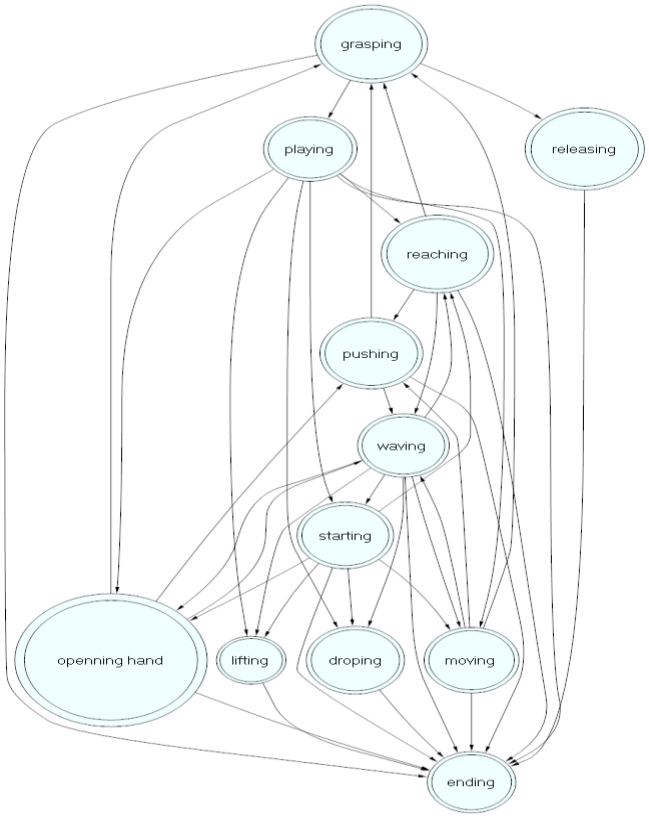
From the results of the three experiments, we have demonstrated that our cognitive control architecture can be used by a robot to modify its behavior in social settings if the robot can recognize the intention of the human by observing motions.

V. Conclusion and Future Work

This paper proposed to integrate imitation learning and cognitive control for a robot to recognize the intention of human actions in social settings and make appropriate decisions to modify its behaviors. In [24], Sartori et. al. showed that an unexpected social request by another human being sometime overrides preplanned actions and modulates his/her the action control system. In this paper, we have sought to demonstrate this type of behavior modulation using imitation learning [7] and cognitive control framework based on our previous work [21]. The experimental results demonstrate that using our integrated cognitive robot architecture, a robot can modulate preplanned tasks based on experience deterministically.

In this paper, ISAC only recognizes one simple intention of the human, i.e. opening a hand and extending the arm. In the future, we need to design several intention recognition methods and store the recognized intention in a hierarchical way in the long-term memory. Another important aspect of the intention recognition though movements is the method in which this intentional information is encoded in the reaching movement and how it may be recognized by the robot. While experiments described in [2] demonstrate that human observers are able to recognize this information, it is not clear exactly how the humans are performing this recognition. Additionally, simple rule-based deterministic models of intention and action are not sufficient for more complex tasks and social situations. Probabilistic models are needed such as models involving Bayesian networks, partially observable Markov decision processes or even hidden Markov models. Such models support the evolution, over time, of the probability of the subject's intention given the observations up to the present.

Appendix



An Example of Behavior Graph

Some behaviors, being some kind of default or self-motivated behavior that the robot interrupts if it is given a task, are connected to "Starting". At that time it starts another task.

REFERENCES

[1] P. Jacob and M. Jeannerod, "The motor theory of social cognition: a critique," *Trends in cognitive sciences*, vol. 9, pp. 21-25, 2005.

[2] C. Becchio, *et al.*, "Grasping intentions: from thought experiments to empirical evidence," *Frontiers in human neuroscience*, vol. 6, 2012.

[3] S.-J. Blakemore and J. Decety, "From the perception of action to the understanding of intention," *Nature Reviews Neuroscience*, vol. 2, pp. 561-567, 2001.

[4] A. Billard, *et al.*, "Robot programming by demonstration," in *Handbook of robotics*, B. Siciliano and O. Khatib, Eds., ed. New York, NY, USA: Springer, 2007.

[5] A. Ijspeert, *et al.*, "Learning attractor landscapes for learning motor primitives," *Advances in Neural Information Processing Systems*, vol. 15, pp. 1523-1530, 2003.

[6] R. Dillmann, *et al.*, "Acquisition of elementary robot skills from human demonstration," in *1995 International Symposium on Intelligent Robotic System*, Pisa, Italy, 1995, pp. 185-192.

[7] H. Tan and K. Kawamura, "A Computational Framework for Integrating Robotic Exploration and Human Demonstration in Imitation Learning," in *the 2011 IEEE International Conference on System, Man and Cybernetics*, Anchorage, AK, USA, 2011, pp. 2501-2506.

[8] S. Calinon, *et al.*, "On learning, representing, and generalizing a task in a humanoid robot," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 37, pp. 286-298, 2007.

[9] M. Mataric, *et al.*, "Behavior-based primitives for articulated control," in *1998 International Conference on Simulation of Adaptive Behavior*, Cambridge, Massachusetts, USA, 1998, pp. 165–170.

[10] E. Theodorou, *et al.*, "A generalized path integral control approach to reinforcement learning," *The Journal of Machine Learning Research*, vol. 11, pp. 3137-3181, 2010.

[11] F. Stulp, *et al.*, "Learning to grasp under uncertainty," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 5703-5708.

[12] J. Kober and J. Peters, "Reinforcement Learning in Robotics: A Survey," *Reinforcement Learning*, pp. 579-610, 2012.

[13] A. Billard, *et al.*, "Discriminative and adaptive imitation in uni-manual and bi-manual tasks," *Robotics and Autonomous Systems*, vol. 54, pp. 370-384, 2006.

[14] S. Calinon, *et al.*, "Learning and reproduction of gestures by imitation," *Robotics & Automation Magazine, IEEE*, vol. 17, pp. 44-54, 2010.

[15] M. Do, *et al.*, "Towards a unifying grasp representation for imitation learning on humanoid robots," in *the 2011 IEEE International Conference onRobotics and Automation,*, 2011, pp. 482-488.

[16] J. Aleotti and S. Caselli, "Part-based robot grasp planning from human demonstration," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 4554-4560.

[17] P. Vinayavekhin, *et al.*, "Towards an automatic robot regrasping movement based on human demonstration using tangle topology," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3332-3339.

[18] P. Pastor, *et al.*, "Skill learning and task outcome prediction for manipulation," in *2011 IEEE International Conference on Robotics and Automation*, 2011, pp. 3828-3834.

[19] J. Ragland, *et al.*, "Neuroimaging of cognitive disability in schizophrenia: search for a pathophysiological mechanism," *International Review of Psychiatry*, vol. 19, pp. 417-427, 2007.

[20] K. Kawamura and S. Gordon, "From intelligent control to cognitive control," in *11th International Symposim on Robotics and Applications (ISORA)*, Budapest, Hungary, 2006, pp. 24-27.

[21] K. Kawamura, *et al.*, "Implementation of Cognitive Control for a Humanoid Robot," *International Journal of Humanoid Robotics*, vol. 5, pp. 547-586, 2008.

[22] M. Malfaz, *et al.*, "A Biologically Inspired Architecture for an Autonomous and Social Robot," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, pp. 232-246, 2011.

[23] P. Munoz, *et al.*, "A Cognitive Architecture and Simulation Environment for the Ptinto Robot," in *2011 IEEE Fourth International Conference on Space Mission Challenges for Information Technology*, Palo Alto, California, USA, 2011, pp. 129-136.

[24] L. Sartori, *et al.*, "Modulation of the action control system by social intention: unexpected social requests override preplanned action," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, p. 1490, 2009.

[25] D. Aarno and D. Kragic, "Motion intention recognition in robot assisted applications," *Robotics and Autonomous Systems*, vol. 56, pp. 692-705, 2008.

[26] R. Kelley, *et al.*, "Understanding human intentions via hidden markov models in autonomous mobile robots," in *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*, 2008, pp. 367-374.

[27] O. C. Schrempf, *et al.*, "A novel approach to proactive human-robot cooperation," in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, 2005, pp. 555-560.

[28] D. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *1994 AAAI Workshop on Knowledge Discovery in Databases*, Seattle, Washington, USA, 1994.

[29] F. T. Cerezo. Available: http://frantracerkinectft.codeplex.com/

[30] S. M. Gordon, *et al.*, "Neuromorphically inspired appraisal-based decision making in a cognitive robot," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, pp. 17-39, 2010.

Feasibility of SAR Approaches - Helping Children with Learning Tasks

Elena Corina Grigore[1] and Brian Scassellati[2]

Abstract— **This paper gives an overview of two projects using a socially assistive robotics (SAR) approach to teach children about nutrition and problem solving, respectively. The first project describes a three-week long study designed for children aged 5 - 8 to interact with a robot that teaches them about making healthy food choices. The study is presented in detail in [1]. This paper gives an overview of the project, describing the results indicating that the children were engaged in a six-session interaction with a robot, that they took less time to respond to the robot towards the end of the study, and that the children had a very positive impression of the robot both before, and after the study. The results show promise for the feasibility of using SAR in one-on-one long-term interactions with children. The second project describes the feasibility of using a robot as part of a problem-solving activity for children. The project is ongoing, and so the paper describes the protocol, and some expected results.**

I. INTRODUCTION

Learning through an interactive and innovative way can be achieved by allowing children to engage in one-on-one interactions with socially expressive robots. Socially assistive robotics (SAR) has been identified as a research area that focuses on helping people through social interaction, rather than any other type of interaction [2], [3]. The two projects we present explore how well a SAR approach is capable of keeping children engaged with the task they are given. The first study explores longer-term exposure to a robot, while the second study focuses more on integrating the robot as part of an already well-established process of teaching.

II. BACKGROUND

Techniques that use widely available technology can be used for the purposes of helping children with various learning tasks. The question, however, is how engaged are the children throughout the learning task and how much better can we do with a system that employs human-robot interaction (HRI) in order to constantly keep children focused on the task? Literature on the topic shows that HRI within a SAR approach can be beneficial for learning. Leyzberg et al. showed that the time it takes people to solve a puzzle decreases when they receive the same hints in the embodiment, i. e. the robot is physically present, versus the on-screen condition [4]. Kidd and Breazeal show that people maintain their diet and exercise habits for longer when guided through the process by a socially assistive robot than when employing other types of intervention, namely a standalone computer method and a paper log method [5].

In our first project, we employ SAR to teach children about nutrition and to inform them about how they can make healthy food choices. Our topic choice stems from the importance of mitigating childhood obesity: "Obesity among children and adolescents has been shown not only to lead to a higher risk of being overweight in adulthood [6], but also of numerous diseases later in life, including high cholesterol and triglycerides, hypertension, and type 2 diabetes [7]. Educating children about healthy food and beverage choices, and motivating them to make healthier choices, can help to lower rates of obesity [8]." [1, p. 1]

Our second project is geared towards teaching children how to problem-solve. The data to be obtained in this study will be used to later inform the development of a study involving children clinically referred for behavioral difficulties. Statistics show that 50% of the American population meets criteria for a diagnosable mental disorder at some point in their lives, while in a given year, one in four Americans meets criteria for a such a condition [9], [10]. Most people in this category, however, do not receive any type of treatment [11]. As a consequence, various treatments have been suggested, including some novel models (e.g., task shifting, best-buy interventions; see [12]). Social robots can be used within various such types of therapy sessions to keep patients engaged in the interaction and focused on the task.

III. NUTRITION STUDY

A. Overview and Platform

The study focused on nutrition investigated the use of a robotic platform to teach children about healthy food choices. The study is part of a multi-site collaboration between Yale University, University of Southern California, the Massachusetts Institute of Technology, and Standford University under the umbrella of an extensive project with the overarching goal of developing robots that interact autonomously with children, helping them with various learning tasks. [1] We ran the study at two different sites, Yale and USC, using a Wizard-Of-Oz system, with a teleoperator choosing the relevant dialogue item for the children's answers. The interaction flow, however, was autonomous, and not controlled by the teleoperator. Our goal for the future is to move towards a fully autonomous platform.

The platform we used during this study is a robot called DragonBot, a socially expressive dragon-like robot with five degrees of freedom, developed at MIT [13], which can be seen in fig. 1. USC designed the skin in collaboration with an expert puppeteer to make the creature as appealing as

[1]elena.corina.grigore@yale.edu
[2]scaz@cs.yale.edu, http://scazlab.yale.edu

[1]This study was ran under the Expedition on Making Socially Assistive Robots funded by the NSF. We thank everyone from Yale University, University of Southern California, MIT, and Standford University involved in developing hardware and software for this study, and in running it.

Fig. 1. Participants' view of DragonBot.

possible to children. The robot has four different sized wings to allow it to "grow" from week to week, as it gets stronger and stronger, as an effect of the children choosing healthy food items for it to eat.

In order to keep the children engaged in the interaction, we created a backstory that spans across the entire course of the interaction. The story is that the robot is about to take part in a dragon race, which he very much desires to win. In order for the robot to win the race, it must become strong and fast, and that can only be accomplished by eating healthy. Children can choose items from a series of fake foods to feed the robot every week. They are thus drawn into this game of helping the robot become stronger each week so that it can ultimately win the big race.

B. Design of the Study

The study spans across three weeks, each week covering a different food topic (e. g. lunchbox, snacks, meals). Each week consists of two one-on-one 10-15 minute sessions. During the first session, the robot acts as an expert, conveying information to the children about the foods presented that week (we call this the Expert Session or ES). During the second session, the robot acts more as a peer, asking the child to make food selections to help it become strong and fast (we call this the Cooperative Session or CS).

C. Data Collected

We collected several types of electronic data, including information about the teleoperator's dialogue choices, as well as video and audio data. In order to measure the level of engagement of the children with the robot, we administered three different questionnaires to the children. The first two types were interaction questionnaires, one that included questions about the perceived value or usefulness of the interaction, and one that included questions about how the children perceived the social presence of the robot (used to quantify the effectiveness of the robot's social capabilities). These questionnaires were administered twice, once after the first interaction, and once after the final one. The third questionnaire asked the children to rate the robot's features, such as bad/good, cuddly/not cuddly, etc. It was administered before the intervention, but after a brief group introduction to the robot, and after the intervention. We also collected information about child temperament by asking the parents to fill out a Child Behavior Questionnaire. This questionnaire contains a 4-point Likert-type scale asking the parents to rate their children's behavior and personality.

D. Results and Conclusions

This section gives an overview of the multiple aspects of the interaction we considered when analyzing the data. These results are presented in detail in [1].

Based on the questionnaire asking children to rate the robot's features, we found that children in the study had an extremely positive perception of the robot, both before the intervention, and after the final session. Our next significant result was that children engaged with the robot and immersed themselves in the story. This is suggested by the decrease in the mean response time (time it took a child to respond to the robot's prompts) from day 1 (4.3 seconds on average) to day 6 (3.5 seconds on average). Due to the short period of the intervention, we found limited evidence showing that children learned about nutrition over the course of the three weeks. Children do show more nutritional knowledge, but this might also be due to the increase in cognitive demands related to making food choices over the weeks. In fact, children took longer to choose food items as the intervention progressed, suggesting they become more thoughtful and thorough in giving their answers over time.

More results indicated that the children engaged more and more with the robot over time since their type of responses changed from week to week: they started off with simple answers (e. g. "Yes", "No", "Hmm"), and continued to use expanded answers ("This is what I fix for dinner..."), and even relational answers (suggesting the children were beginning to relate to the robot, e. g. "You said you didn't like it!"). We also did not find a link between child temperament and social interaction with the robot, meaning that children with diverse temperaments could develop a relationship with a robot.

IV. PROBLEM SOLVING SKILLS STUDY

A. Overview and Platform

The problem-solving skills study is an undergoing project in collaboration with the Yale Parenting Center [14]. The ultimate goal of this project is to integrate a robotic platform into a problem-solving skills training process for children clinically referred for behavioral difficulties. The initial study aims to use the same robotic platform described in the previous study (DragonBot) for a single session within the problem-solving skills training method with children who are not clinically referred.

B. Design of the Study

Children who participate in the Problem-Solving Skills Training (PSST) program at the Yale Parenting Center go through a 12-session process to learn ways to cope with real-world situations that may prove difficult for them. These sessions focus on a 5-step method of how to appropriately deal with an everyday problem a child may encounter. The five steps are designed to help children come up with different potential behavior options, evaluate the consequences of each, and make a decision based on this.

Children become highly engaged and motivated while interacting with robots. Based on this observation, we are

integrating DragonBot as part of this process, so that children can learn the steps through interacting with a robot. We change the above-described design to be able to control for the multitude of variables associated with a 12-session process and to evaluate the feasibility of using such a technique.

All participants will initially take part in a 30-minute PSST session with a member of the study staff, focusing on teaching the child a series of three problem-solving steps, a subset of the steps used for children with disruptive behavior problems [15]. After having completed this training session, the participants will either be assigned to the practice-with-DragonBot or the practice-alone condition. Children in the former condition will be introduced to DragonBot and told to "teach" the robot the same problem-solving steps they just learned. Similarly to the previous study, DragonBot has a backstory of being a baby dragon that needs help with problem-solving. Through this teaching task, the children will be able to practice working with the steps alongside a robot by creating a peer-to-peer relationship with it. Children in the latter condition will be instructed to review the steps that they just learned on their own.

C. Data to be Collected

We are interested in assessing how well-suited the use of a robotic platform in such a context is, and in assessing the acceptability of this kind of treatment.

Children will complete different questionnaires, based on the condition they are assigned to. Children assigned to the robot condition will complete the Child Reaction to the Robot Interview, to assess their reaction to the robot (including questions on likability, animacy, physical appearance, and utility), and a Child-Robot Alliance Interview (to assess the child's relationship with the robot). Children assigned to the practice-alone condition will complete the Child Reaction to the Practice Task Interview, containing questions designed to obtain the children's feedback about the task. All children will complete the Child Version of the Treatment Evaluation Inventory, containing questions to assess how acceptable the treatment is from the perspective of the child.

Parents will be given the opportunity of watching the sessions their children are participating in, through a video monitor system. After having observed the sessions, the parents will be asked to complete the Child Behavior and Temperament Questionnaire. This data will help in later analyzing whether temperament is linked to the type of interaction we will observe between the child and the robot. Parents will also complete the Parent Version of the Treatment Evaluation Inventory to assess whether they view the treatment as acceptable.

D. Predicted Results

We predict that children in the robot condition will be more engaged in the task than children in the practice-alone condition. We also predict that the children interacting with DragonBot will be highly engaged in the task and in teaching the robot the steps, leading to their better understanding of what they had previously learned.

V. CONCLUSIONS

This paper gave the overview of two projects using HRI within a SAR approach to help children with learning tasks. The projects show the feasibility of using such approaches given the high level of engagement children demonstrated throughout one of the presented studies and the expected level of interaction as part of the ongoing study. This encourages us to continue using such techniques and to continue exploring the benefits of using interactive ways of helping children gain educational knowledge on different topics.

REFERENCES

[1] E. Short, K. Swift-Spong, J. Greczek, A. Ramachandran, A. Litoiu, and E. C. Grigore, "How to train your dragonbot - socially assistive robots for teaching children about nutrition through play," 2013.

[2] D. Feil-Seifer and M. J. Mataric, "Defining socially assistive robotics," in *Rehabilitation Robotics, 2005. ICORR 2005. 9th International Conference on*, 2005, pp. 465–468.

[3] A. Tapus, M. Mataric, and B. Scasselati, "Socially assistive robotics [grand challenges of robotics]," *Robotics Automation Magazine, IEEE*, vol. 14, no. 1, pp. 35–42, 2007.

[4] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proc. of the Annual Meeting of the Cognitive Science Society (CogSci)*, no. 1, 2012, pp. 1882–1887.

[5] C. Kidd and C. Breazeal, "Robots at home: Understanding long-term human-robot interaction," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, 2008, pp. 3230–3235.

[6] A. S. Singh, C. Mulder, J. W. R. Twisk, W. V. Mechelen, and M. J. M. Chinapaw, "racking of childhood overweight into adulthood: a systematic review of the literature." *Obesity Reviews*, no. 9, pp. 474–488, 2008.

[7] D. S. Freedman, W. H. Dietz, S. R. Srinivasan, and G. S. Berenson, "The relation of overweight to cardiovascular risk factors among children and adolescents: The bogalusa heart study," *Pediatrics*, vol. 103, no. 6, pp. 1175–1182, 1999.

[8] D. Spruijt-Metz, "Etiology, treatment, and prevention of obesity in childhood and adolescence: A decade in review," *Journal of Research on Adolescence*, vol. 21, no. 1, pp. 129–152, 2011.

[9] R. C. Kessler and P. S. Wang, "The descriptive epidemiology of commonly occurring mental disorders in the united states*," *Annu. Rev. Public Health*, vol. 29, pp. 115–129, 2008.

[10] R. Kessler, S. Aguilar-Gaxiola, J. Alonso, S. Chatterji, S. Lee, J. Ormel, T. Üstün, and P. Wang, "Special articles. the global burden of mental disorders: an update from the who world mental health (wmh) surveys," *Epidemiologia e psichiatria sociale*, vol. 18, no. 1, p. 23, 2009.

[11] R. C. Kessler, O. Demler, R. G. Frank, M. Olfson, H. A. Pincus, E. E. Walters, P. Wang, K. B. Wells, and A. M. Zaslavsky, "Prevalence and treatment of mental disorders, 1990 to 2003," *New England Journal of Medicine*, vol. 352, no. 24, pp. 2515–2523, 2005.

[12] A. E. Kazdin and S. M. Rabbitt, "Novel models for delivering mental health services and reducing the burdens of mental illness," *Clinical Psychological Science*, vol. 1, no. 2, pp. 170–191, 2013.

[13] A. Setapen, Master's thesis, Massachusetts Institute of Technology, Department of Architecture, Program in Media Arts and Sciences, 2012.

[14] [Online]. Available: http://childconductclinic.yale.edu/

[15] A. E. Kazdin, T. C. Siegel, and D. Bass, "Cognitive problem-solving skills training and parent management training in the treatment of antisocial behavior in children." *Journal of consulting and clinical psychology*, vol. 60, no. 5, p. 733, 1992.

Human motion measures and adequacy of the response in relation to the robot's sociable character.

Ritta Baddoura, Gentiane Venture

***Abstract*— During an unannounced encounter, we study the appreciation of the robot's sociable character (measured via the answers to a questionnaire) and the motion (arm and head movement frequency and smoothness using IMU sensors) of two humans interacting with a proactive humanoid (NAO). We also investigate the dependencies between the participants' response to the robot's non-verbal actions and their perception of its sociability. Our results show positive correlations between finding the robot sociable and responding adequately to its engaging gestures on one hand, and between finding the robot sociable and the frequency and jerk of the human hand and head movements. Therefore, the social dimension attributed to the robot seems to be a major component of the human willingness to interact, as well as of the general success of the interaction. Furthermore, these results suggest that it might be possible to infer human appreciation of a robot and human involvement in an interaction, using their measurable physical movements, thus opening up to interesting applications in HRI from the robot control design point of view.**

I. INTRODUCTION

Robots are gradually appearing in our society and their presence in public and private spaces is still a new experience for humans. The variety in robots, in humans and in possible encounters makes the study of human-robot interaction (HRI) more complex and at the same time more expressly needed. Robots are expected to interact with humans in an efficient and natural way: intuitive and easy exchanges are the two main characteristics defining the sociability of artificial agents [1]. Social interaction and robot acceptability are amongst the most important concepts explored in social robotics nowadays. Although frequently used by the scientific community, these expressions refer in reality to complex psychological and social concepts difficult to precisely define [2], [3]. In addition the understanding of these concepts is inter-cultural and inter-personal [4], [5]. Acceptability or user acceptance is usually defined as the "demonstrable willingness within a user group to employ information technology for the task it is designed to support" [6]. Rogers [7] describes an innovation diffusion model and theory, based on a technology's five characteristics, which determine its diffusion and its acceptance. Compatibility, which refers to consistency with social practices and norms among users, is one of these five characteristics.

The authors acknowledge the support provided by the Montpellier 3 University, France (L'Ecole Doctorale 59 and C.R.I.S.E.S) and the "Women's Future Development Organization".

Ritta Baddoura is with the C.R.I.S.E.S (Center for Cross-disciplinary Research in Humanities and Social Sciences), University of Montpellier 3, Montpellier, France (phone: +33-6-15211446, fax: +81-4-2385-7204, e-mail: rita.baddoura@etu.univ-montp3.fr).

Gentiane Venture is with the Dpt of Mechanical Systems Engineering, Tokyo University of Agriculture and Technology, Tokyo, Japan (e-mail: venture@cc.tuat.ac.jp).

The way a "*socially adapted*" interaction builds up and the way social acceptance and social well-being occur are difficult to comprehend in a human-human interaction, knowing that what might be acceptable or satisfying for one individual is likely to be differently perceived by another. Few social experiences happen without experiencing some ambiguity or ambivalence, especially in a first encounter. When it comes to human-robot interaction, questions of social acceptance and of "intuitive" and "successful" interactions seem more crucial since the difference between humans and robots is fundamental and ontological. The quality of such interactions depends strongly on the robot: not only on its appearance, but also on its abilities, features and autonomy degree. The interaction quality depends also on the humans' perception and appreciation of the robot and their readiness to adapt to it, thus taking its abilities and limitations into account and compensating for them in order for the interaction to happen [8]. Additionally, the degree to which a human-like nature for a robot is needed is not yet understood to a sufficient degree and studies that focus on such a human-like nature are very rare [8]. Most studies [9], [10] agree on the fact that further research is needed to better understand and determine which aspects and degrees of similarity and likeability are required in order to enable more empathic and intuitive HRI.

In the present paper, we aim at understanding how a first encounter builds up to a successful or unsuccessful HRI, particularly in relation to the participant's perception and evaluation of the humanoid robot partner's sociability during the encounter. We are also interested in the way the evaluation and appreciation of the robot's sociable character might be associated to the participant's measurable arm and head motions. In a previous study [11] we showed that the more the robot was found sociable by the human partner, the more interacting with it was perceived as familiar, secure and comfortable. Our results also proved politeness and sociability of the robot to be highly and positively correlated, as well as showed non-verbal social gestures (greetings) performed by the robot to be efficient in promoting its sociable character [11].

Non-verbal communication has the ability to replace verbal language to a large extent, especially when it is about communicating simple information, giving social cues or conveying emotions and intentions [12]. The importance of building communicative robots that are able to generate social cues through gesture has been showed by some recent studies such as [13]. In [13] authors were also able to underline the positive effects of gestures during HRI by showing that people will have a more meaningful social interaction with a robot and enjoy it more when the robot shows gestures than when it does not. Also, people will report a greater level of

engagement with a robot when the robot shows gestures during an interaction than when it does not [13], [14]. [15] showed meaningful correlation between certain body movements (such as eye contact and synchronized arm movement) of humans interacting with a robot and their subjective evaluations of this robot, thus suggesting that humans make evaluations based on their body movements. [15] also found that when a human highly evaluates the robot, the human behaves cooperatively with it, which will further improve its evaluation.

In [16] we showed that the more humans found the robot to be sociable, the higher was the intensity of their arm motion when greeting it back goodbye. In the present study, we hypothesize that the participants' movement frequency and smoothness are directly and strongly associated with their appreciation of the robot's sociable character and that the more sociable the robot seems to them, the more they tend to successfully respond to its gestures: (H1) The more humans find the robot sociable, the more they are prone to adequately react to its engaging actions. (H2) The human evaluation of the robot partner's sociability and their arm and head measurable motions (frequency and smoothness) are dependant.

II. METHOD AND EXPERIMENTS

The experiment involves a triad: a robot and 2 participants (X and Y) at a time. The participants are only invited to answer a questionnaire on the perception of robots. They are informed that the set is filmed and that sensors are placed around their head and wrist for motion capture. They do not know about the robot's intervention and their possible interaction with it. The only instruction given to them is to answer a questionnaire.

Participants were randomly assigned to one of the two sitting positions that resulted from a 1(X) x 1(Y) between-subjects design (NAO's behavior when handing the envelope: Smooth (with X) vs. Resisting (with Y)). Resisting behavior refers to the fact that NAO stands slightly farther from participant Y than it did from participant X when handing him/her the envelope. It also refers to NAO keeping the envelope for four seconds in its fingers before releasing it to Y, whereas the release to X was immediate. Once the experience starts, there is no further intervention from the staff. Participants are not instructed about what they ought to do, it is all upon their own judgment. The scenario's duration involving the robot is about 1 minute. Then, the questionnaire requires 5 to 10 minutes to be filled.

A. The robot

NAO (Aldebaran Robotics) is a 57-cm tall commercial humanoid robot. Its body has 25 degrees of freedom (DOF) whose key elements are electric motors and actuators. We used the programming software delivered with the robot to control it. We deliberately chose feed-forward control of the robot for repeatability. Of course, the substitution of NAO with any other robot can change the impression felt during the interaction, yet it would not change the association of certain physical behaviors (motion) with the mental and psychological states of the participants.

B. The participants

The 20 pairs of students, 40 students in total (14 women, 26 men), were recruited on the campus of Tokyo Univ. of Agriculture and Technology, and volunteered to participate in a study on the perception of robots. Participants range in age from 19 to 35 years (X: M= 23.75, SD= 3.53; Y: M= 22.7, SD= 1.68). Though previous exposure to robots was not controlled when recruiting the participants, candidates were mainly students from agriculture, biology and chemistry departments. We considered that having seen a robot in videos or having been exposed to a robot does not necessarily mean exposure to a humanoid robot or to the same robot used in the experiment. The interactive and relational (and possibly reciprocal) dimensions involved in HRI are more subjective than rational and even a person who might be used to manipulating robots might, once the robot manifests as an interaction proactive partner, not behave with the same comfort or detachment or obviousness than the one possibly expected.

C. Experimental choices and set-up

(1) The experimental set-up consists of a rectangular area limited by colored screens. It is furnished with a carpet, a low table equipped with pens, and 2 cushions put directly on the floor on each side of the table, providing therefore a comfortable Japanese-style ambiance, closer to a cozy space rather than to an anonymous lab. Also when seated on the cushions, participants are positioned on a low level which, given NAO's small height, enables face-to-face contact.

(2) The experiment starts with NAO entering the room, facing the table and holding in each hand an envelope with the word "Questionnaire" obviously written down on it. NAO walks towards the participants, then stops a few centimeters away from the table and greets them by bowing (his head bends with a slight forward bending of the upper torso). NAO turns towards participant X sitting to its left and extends its left arm holding the envelope in their direction. After a few seconds, its fingers release tension and the envelope is then ready to fall down in the participant's hand or on the floor, depending on the participant's reaction (Fig. 1).

Then NAO turns towards participant Y, extends its right arm holding the second envelope in their direction. NAO is slightly more distant from participant Y than it was from participant X; so in order for the envelope exchange to happen, Y has not only to extend his/her arm, but also to lean forward and reduce the distance from NAO (Fig. 1). Another difference from the interaction with X, is that NAO will now keep the envelope 4 seconds between its fingers before releasing it. Having delivered both envelopes, NAO waves goodbye with its right hand, turns around and walks back towards the door.

Participants are free to start filling the questionnaire any time after receiving the envelope. We chose to ask them to answer the questionnaire at the end of the encounter and not after each key-moment in order to enable the interaction to be uninterrupted, and to enable the candidates to remain as natural and spontaneous as possible without any disturbance. The whole situation lasts for one minute only and the participants' memory and impressions about their encounter with the robot are likely to be still fresh and vivid.

(3) Having two participants at a time might probably bring some uncontrolled variable, but it also contributes to limit the artificial dimension of the experiment and enhance the real dimension of the encounter. This choice allows NAO to manifest different -possibly perceived as "subjective"- behaviors regarding the same action: delivering the questionnaire. Furthermore, we felt that the stress that might generate from the unpredictable factors proper to the situation as well as from a close encounter with a robot might be counter-balanced, or at least eased, by being two persons facing the robot (all the pairs were recruited together and consequently knew each other).

(4) The robot shows a slightly different behavior with each participant, which allows limiting its repeatability and predictability (and mechanical functioning as a machine). It also allows seeing how this difference of behavior would be interpreted by both participants. This situation is a particular illustration of what could happen in the future in public (or even domestic) spaces where the robot has precise tasks to accomplish and is prone to interact with different users that are not inevitably aware of its intervention and are relatively free to interact with it.

(5) NAO is not presented here as an experimental object but rather has a proactive role and a real essential task to accomplish which makes it a clear potential interaction partner. Furthermore, it punctuates the encounter's beginning & end with non-verbal greetings (NAO bows in the beginning according to the Japanese way of greeting and greets goodbye by waving its hand in a more international style this time).

D. Data collection

We used in this study distinct but complementary tools in order to have a more accurate and faithful access on what was really experienced by the participants as well as to limit ambiguity in the results and explore the possibility of combining variables of different kinds (e.g. answers to the questionnaire and reactions to the robots) to analyze the data available and have a different perspective on the participants' experience of the encounter with the robot.

(1) The questionnaire proposed to the participants consists of three parts/methods addressing different topics but also sometimes the same topic considered from different perspectives. The questionnaire is written in Japanese to avoid possible confusions in the nuances that an insufficient level of English could bring. It consists of a first part using a 7-point Likert scale, a second part with Multiple Choice Questions (one of the topic addressed here is earlier exposure to robots), and a third very short part consisting of two open-ended questions enabling the participants to describe NAO and the interaction with it in their own words. In the present study we focus on the participants' ratings of NAO's sociability which are obtained using the 7-point Likert scale: 1 meaning "not sociable at all", 7 "Highly sociable". We added 0 for "Irrelevant statement" to allow a more precise expression.

(2) Each experimental session is video recorded using two stable cameras: One is filming the set from behind and gives images of the robot entering the set and of its interaction with the participants. The other is facing the participants and providing images of their movements and facial expressions. This tool is particularly used to collect data on the participants' non-verbal behavior and on their reactions (answer back or not) to NAO's gestures. The recorded data is reinforced with observation notes taken by the psychologist of our team.

(3) Two IMU (Inertial Measurement Unit) sensors are used for each participant. One is fixed on the forehead to capture the head and upper torso movements; the other on the arm -the right arm for X participants and the left for Y participants, each being respectively the closest arm to the robot's position and the one to be most likely used (from our observations on a pilot study of 20 candidates) by the participants to fetch the envelope. The IMU sensors measure the longitudinal accelerations and the rotational velocities around 3-axes. Thus, more discrete micro-movement data is recorded giving us another level of information regarding the participants' experience and reactions to NAO. Data for two pairs of candidates are unavailable.

E. Data analysis

(1) The motion data are analyzed from the IMU and from the video. From the IMU data, the 3 components of the rotational velocity and the 3 components of the acceleration are post-processed separately to obtain two types of information. Frequency analysis: First a simple frequency analysis on the hand motion data (angular velocity) is performed during the grasping motions and when answering the robot's goodbye by waving the hand; the frequency (*Hz*) of the first pick is used. Second, the smoothness of the motion during the overall interaction is computed.

Motion smoothness: There are several methods to assess the motion smoothness [17], we chose the jerk metrics ($1/s^2$). For that, the accelerations are used to compute the jerk magnitude averaged over overall motion and normalized with respect to the peak speed. The smaller the jerk metric is the smoother the movement is.

(2) We calculated the descriptive statistics (95% Confidence Interval) related to the participants' responses to the robot's engaging actions to interact as well as the descriptive statistics based on their answers to all the parts of the Questionnaire except for the open-ended questions part from which answers were used when clarification was needed [11]. We also calculated the Chronbach's α reliability for certain items in the questionnaire and found the questionnaire to be valid and to have a good internal reliability (Table I).

TABLE I. CHRONBACH'S α RELIABILITY TEST FOR SELECTED ITEMS IN THE QUESTIONNAIRE RELATED TO THE ENCOUNTER WITH NAO AND NAO'S SOCIABILITY

Questionnaire items	**Participants' evaluation of the encounter (About NAO & Interacting with it)** **Sections A & B**	**NAO's Sociability** **A6 A9**
Chronbach's α reliability	0.83	0.84

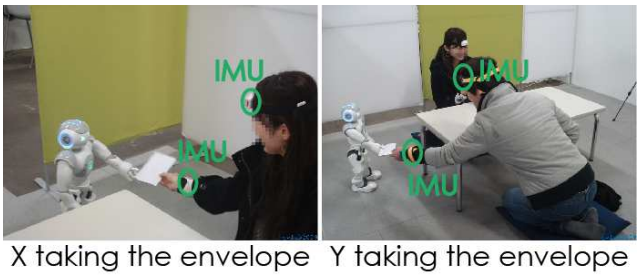


Figure 1. The experimental setup when X and Y take the envelope from NAO

III. MAIN RESULTS

Evaluating possible dependencies (Spearman's rank correlation coefficient) between different pairs of variables (Table II, Table III), particularly in relation to the robot's sociable character, to the participants' reactions to its gestures and to the measured motion, showed interesting and strong associations between many criteria, and also revealed the absence of any correlation between some others. Only the most significant results related to this paper's hypotheses are presented here (Fig. 2, Table II and Table III). In Table II and Table III, the p-value is obtained using the Student's t-distribution.

A. Earlier exposure to robots

Earlier exposure was low for X (M= 2.55, SD= 2.06, SEM= 0.42) and medium for Y (M= 3.85, SD= 2.43, SEM= 0.54). 40% of X reported having never been exposed to real robots, 45% said to be familiar with robots from movies and literature. 40% of Y reported to have been exposed at least once to a real robot, 55% said to be familiar with robots from movies and literature. Generally, most of the participants (70% of the 40 participants) had never been exposed to a real robot before.

B. Reactions to NAO's gestures

Most participants found rather easy/clear to understand NAO's actions (X: M= 5.25, SD= 1.71, SEM= 0.38; Y: M= 4.8, SD= 1.73, SEM= 0.38) and NAO's behavior (X: M= 5.00, SD= 1.97, SEM= 0.44; Y: M= 4.55, SD= 1.79, SEM= 0.4). Most participants (80% X; 75% Y) understood NAO's intention of giving them the envelope and found it easy to react to it. Only 35% of X and 25% of Y found it easy to decide on how/whether to react to its greetings. Participants were mostly confused about taking decisions regarding: reacting or not to NAO's greetings and actions since they were not told too (35% X), opening or not the envelope (55% X, 35% Y) and taking or not the envelope when NAO resisted (55% Y).

Though they have not been previously informed about the interaction or instructed about what they ought to do, 80% X and 85% Y were proactive towards NAO's arm movement and took the envelope (Fig. 2). Of course, Y had seen NAO performing the same movement with X which might have facilitated their reaction, knowing that this possible effect was not addressed in our study. Nevertheless, as the novelty of NAO resisting before handing the envelope is introduced with Y, the large number of participants who adequately reacted is to be noted. Reacting to NAO's greetings was less effective as less than half of the participants answered to it (hello: 45% X; 35% Y; goodbye: 30% X; 35% Y). More generally, the more participants were able to make sense of the interaction, the more they found it easy to react to NAO's engaging actions (X: P=0.02, R=0.50; Y: P=0.04, R=0.46) [18].

We ran a T-Test to compare X and Y participants' reactions to the robot's engaging actions. When comparing X participants' and Y participants' respective reactions to NAO's greetings, before (greeting hello) and after (greeting goodbye) exchanging the envelope, we found the difference in their response to be not statistically significant, thus likely to be the result of random chance alone [11]. The results of the T-Test showed similar lack of statistical significance when comparing between X participants' reactions and Y participants' reactions to NAO handing them the envelope. No statistical proof was found to assert that X and Y reacted differently to NAO, nor to assume that the difference of behavior showed by NAO when handing the envelope respectively to X and Y participants had a relevant impact on their respective reactions (to greetings and to the envelope exchange).

C. Evaluation of NAO's sociable character

To describe NAO's character and its perception by the participants, several adjectives such as sociable, polite, caring, interesting, funny, seductive or hostile and unpredictable, are rated by them in the first part of the questionnaire. We made the choice of including a large variety of adjectives, not necessarily related to NAO's role or to its displayed characteristics during the experiment, in order to enable the participants to express as widely and personally as possible their perception of NAO. Indeed, as showed by [19], [20], humans tend to draw inferences about a robot's abilities and personality in a way going beyond its observable actions. In [11], results showed that the participants' perception of NAO and their perception of interacting with it were generally highly positive. Even NAO's difference of behavior from participant X to participant Y, during the envelope exchange, though unpredictable, is not negatively interpreted neither by X or Y.

Most participants found NAO to be medium-to-high sociable (X: M= 4.95, SD= 1.82, SEM= 0.40; Y: M= 4.85, SD= 2.08, SEM= 0.46). T-Test results showed a lack of statistical significance when comparing X participants' and Y participants' respective evaluation of NAO's sociability. We therefore considered X and Y participants as one group of 40 candidates when calculating the correlations between their perception of NAO's sociable character and their response to its different gestures (greeting hello, handing the envelope, greeting goodbye).

D. Finding NAO sociable and reacting to its gestures

High positive correlations (Spearman) were validated between the participants' evaluation of NAO's sociability and their reactions to its different gestures (Table II). More particularly, results showed that the more participants found NAO sociable, the more they reacted adequately to its gestures, i.e. the more they greeted it back hello and goodbye, and the more they took the envelope from it.

E. Appreciating the robot's sociable character in relation to the motion frequency and smoothness

There is a significant positive correlation between finding the robot sociable and the frequency of the arm motion when taking the envelope (Table III). The more sociable the robot is perceived, the higher the frequency of the movement is. Similarly, there is a strong positive correlation between the frequency of waving back goodbye and the sociability of NAO (Table III).

No significant correlation was validated regarding the robot's sociable character and the frequency of waving back hello. This is likely to be a consequence of the fact that the robot's hello gesture and the participants' reaction to it (waving back hello) occur in the very first beginning (the first ten seconds more exactly) of the unexpected encounter: at this moment, the whole encounter (meaning the robot, its intervention and the interaction with it) is new to the participants who probably have not had yet the time to evaluate the robot's character and have not yet a clear opinion/impression about it.

When considering the smoothness of the motion, the jerk metrics of the head motion appear to be positively correlated with the evaluation of NAO's sociability (Table III). The head movement tends to be more jerky (less smooth) when the robot is perceived as being sociable, which means that on the contrary the head movements have a tendency to freeze when the robot is perceived as having little/limited sociability which translates into a less jerky (smoother) motion.

TABLE II. SIGNIFICANT SPEARMAN CORRELATIONS BETWEEN TWO PARAMETERS: EVALUATING NAO'S SOCIABLE CHARACTER AND REACTING TO NAO'S PROACTIVE ACTIONS

Variables	p-value	Corr.
NAO is sociable / Reacting to NAO greeting hello	0.01	0.47
NAO is sociable / Reacting to NAO greeting goodbye	0.01	0.54
NAO is sociable / Reacting to NAO handing the envelope	0.01	0.65

TABLE III. SPEARMAN CORRELATIONS BETWEEN EVALUATING NAO'S SOCIABLE CHARACTER AND TWO MOTION CHARACTERISTICS: THE ARM FREQUENCY WHEN REACTING TO NAO'S GESTURES AND THE JERK METRICS OF THE OVERALL MOVEMENTS OF THE HEAD AND TORSO

Variables	p-value	Corr.
NAO is sociable / Frequency of the human arm when greeting back NAO goodbye	0.001	0.54
NAO is sociable / Frequency of the human arm when reacting to NAO handing the envelope	0.01	0.31
NAO is sociable / Jerk metrics of the human overall movements of the head and torso	0.05	0.43

IV. DISCUSSION AND CONCLUSION

When comparing X participants' and Y participants' respective reactions to NAO's actions (greeting hello, handing the envelope, greeting goodbye), as well as when comparing their respective perception of NAO's sociability, no proof was found to assert that the difference of behavior displayed by NAO had an impact on X or Y. This might be explained by the fact that Y participants saw NAO handing X participants the envelope, and that X and Y participants' respective perception of it and of the interaction with it were based on the global performance of the robot and not strictly on their one-to-one interaction with it [11]. Also, a recent study [21] provided the evidence for the fact that the unpredictability of a robot's actions does not necessarily lead to less acceptance or less liking of it. Therefore, it is possible to consider that the unpredictability of the encounter with the robot in our experience, as well as the unpredictability of its change of behaviour, did not make neither Y participants, nor X participants, less (or differently) appreciate NAO, or less (or differently) respond to it.

The results in Table III show that the arm and head movements of a human partner interacting with a humanoid robot are directly and strongly associated with their appreciation of the robot's sociable character (H2 validated). The results in Table II also prove that the human partner's appreciation of the robot's sociability and the adequacy of the human partner's response to the robot's gestures, are strongly dependent (H1 validated). The sociability of a humanoid robot appears to be essential to successfully and adequately engage humans in an interaction (e.g. a joint-task such as exchanging an envelope).

It is also possible to suggest that the success of a human-robot interaction is strongly readable from the human partner's motion analysis, particularly from the head and arm movements. When the robot is perceived as being sociable or is likened to a partner with sociable traits, the participants move their arm more often (Table III) and respond adequately to the robot's actions more often (Table II), whereas their head moves in a more brisk manner (Table III) which might indicate a potential feel of confidence experienced by them as well as reflect their engagement in the interaction. It is important to keep in mind that the analysis (jerk metrics) of the smoothness of the head and torso movements was computed during the whole interaction with the robot. This means that the motion capture of the participants' head and torso is carried out during the complete duration of the encounter. These head and torso movements are not only the ones possibly displayed when answering NAO's hello (some participants greeting NAO back by bowing, whereas others just reacted by waving their hand or their arm), but are the overall movements displayed during the interaction.

Though participants gave medium to high scores to "NAO is sociable", most of them engaged more frequently in a task-oriented (or useful) interaction (taking the envelope from NAO as they needed the questionnaire it contained), than in a pure social interaction (such as greeting) [16]. Participants appreciated NAO's sociable character but did not feel obliged to act towards it with reciprocity and equivalent sociability [11]. Results also suggest that even though enabling robots to perform socially-engaging gestures such as greetings, has a

positive impact on appreciating the robot's sociability and responding adequately to its actions, these social features of the robots are not necessarily sufficient to efficiently engage humans in what would be a strictly social exchange/interaction. In [22], the authors have shown that humans tend to respond socially to robots, whereas in [23] the authors have shown that it is probable that they won't exactly react to them as they would to other humans. In the particular context of the present study, it is important to keep in mind that participants were not instructed to interact on demand or to fulfil experimental directives regarding the response to the robot's actions.

We are aware that this paper's results are limited to a specific interaction with a specific robot: NAO, and would gain in being tested with different types of interaction scenarios and with different types of robots. The study being conducted in Japan, our findings will gain to be compared to results from a future experiment conducted in a western culture, in order to get an insight on intercultural variations. In that case, some experimental choices such as the Japanese design of the set-up (low table, cushions on the floor), and the use of bowing by way of greeting, would be changed and adapted to the experiment's new context, in order to avoid cultural bias. Furthermore, conducting the same study in another culture would enable us to investigate, in relation to this culture's social codes, the interest of measuring the motion of certain specific body parts (such as the arm, or the head and torso) as well as of investigating their possible dependencies with the human partner's inner experience.

The dependencies existing between body motion, appreciation of the robot's sociability, and adequacy of the human response to the robot proactive partner seem crucial for the quality and the success of HRI. The trends highlighted in this paper can provide a support to design intelligent systems able to instantly measure and possibly 'live' assess and interpret human internal experience and evaluation of the robot's characteristics (e.g. its sociability), thus adjusting their behavior accordingly. All these findings as well as the interpretations and the control design developments they open up to, ought to be investigated in future studies.

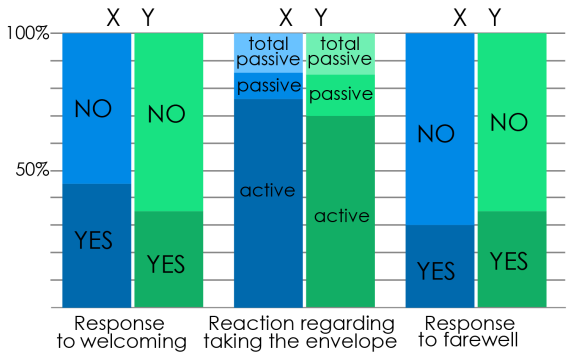


Figure 2. Participants' reactions to the robot's gestures

REFERENCES

[1] L. Brayda and R. Chellali, "Measuring Human-Robots Interactions", *Int. J. Soc. Robot.*, no 4, pp. 219–221, 2012.

[2] N. Lee, H. Shin, and S. Sundar, "Utilitarian vs. hedonic robots, role of parasocial tendency and anthropomorphism in shaping user attitudes", *Proc. Int. Conf. Human-Robot Interaction*, pp. 183–184, 2011.

[3] S. Turkle, "A nascent robotics culture: New complicities for companionship", *AAAI Technical Report Series*, 2006.

[4] K. Fischer, "Interpersonal variation in understanding robots as social actors", *Proc. Int. Conf. Human-Robot Interaction*, pp. 53–60, 2011.

[5] M.Walters, D. Syrdal, K. Koay, K. Dautenhahn, and R. Boekhorst, "Human approach distances to a mechanical-looking robot with different robot voice styles", *Proc. IEEE Int. Symp. Robot and Human Interactive Communication*, pp. 707–712, 2008.

[6] A. Dillon, "User acceptance of information technology", in *W. Karwowski (ed), Encyclopaedia of human factors and ergonomics*, Taylor and Francis, London, 2001.

[7] EM. Rogers, *Diffusion of innovation*. 5th ed. Free Press, New York, 1995.

[8] E. Takano, T. Chikaraishi, Y. Matsumoto, Y. Nakamura, H. Ishiguro, and K. Sugamoto, "Psychological effects on interpersonal communication by bystander android using motions based on human-like needs", *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, pp. 3721– 3726, 2009.

[9] C. Bartneck, T. Kanda, H. Ishiguro, and N. Hagita, "My robotic doppelganger - a critical look at the uncanny valley theory", in *18th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 269–276, 2009.

[10] L. Canamero, "Playing the emotion game with feelix: What can a lego robot tell us about emotion?", *Socially intelligent agents: Creating relationships with computers and robots*, pp. 69–76, 2002.

[11] R. Baddoura, R. Matuskata, and G. Venture, "The familiar as a key-concept in regulating the social and affective dimensions of HRI", in *Proc IEEE/RAS Int. Conf on Humanoid Robots*, pp. 234–241, 2012.

[12] T. Kanda, H. Ishiguro, M. Imai, T. Ono, "Development and evaluation of interactive humanoid robots" *in Proc. IEEE 92*, vol. 92, no. 11, 2004, pp. 1839–1850.

[13] A. Kim, J. Han, Y. Jung, K. Lee, "The Effects of Familiarity and Robot Gesture on User Acceptance of Information", in *8th ACM/IEEE Int. Conf. On Human-Robot Interaction (HRI)*, 2013, pp. 159-160.

[14] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, F. Joublin, "Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot", in *Proc. 3rd Int. Conf. on Social Robotics*, Springer-Verlag Berlin, Heidelberg, 2011, pp. 31-41.

[15] T. Kanda, H. Ishiguro, M. Imai , T. Ono, "Body Movement Analysis of Human-Robot Interaction", in *Proc. 18th Int. Joint Conf. on Artificial Intelligence*, 2003, pp. 177-182.

[16] R. Baddoura and G. Venture, "Social vs. Useful HRI: experiencing the familiar, perceiving the robot as a sociable partner and responding to its actions", *Int. J. Social Robotics*, 2013 (in press).

[17] B. Rohrer, S. Fasoli, H. Krebs, R. Hughes, B. Volpe, W. Frontera, J. Stein, and N. Hogan, "Movement smoothness changes during stroke recovery", *J. of Neuroscience*, vol. 22, no. 18, pp. 8297–8304, 2002.

[18] R. Baddoura, T. Zhang, and G. Venture, "Experiencing the familiar, understanding the interaction and responding to a robot proactive partner", In *ACM/IEEE Int. Conf. on Human-Robot Interaction*, pp. 247–248, 2013.

[19] A. Powers, and S. Kiesler, "The advisor robot: Tracing people's mental model from a robot's physical attributes", Int. *Conf. on Human-Robot Interaction 2006*, pp. 218-225, 2006.

[20] F. Eyssel, F. Hegel, G. Horstmann and C. Wagner, "Anthropomorphic inferences from emotional nonverbal cues: A case study", In *Proc. of the 19th IEEE Int. Symposium in Robot and Human Interactive Communication (RO-MAN 2010)*, pp. 681-686, 2010.

[21] F. Eyssel, D. Kuchenbrandt and S. Bobinger, "Effects of Anticipated Human-Robot Interaction and Predictability of Robot Behavior on Perceptions of Anthropomorphism", *Int. Conf. on Human- Robot Interaction (HRI 2011)*, pp. 61–67, 2011.

[22] P.J. Hinds, T.L. Roberts and H. Jones, "Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task", *Human Computer Interaction*, pp. 151-181, 2004.

[23] S. Cowley and T. Kanda, *Friendly machines: interaction-oriented robots today and tomorrow*, Alternation, 2005.

Learning object names through shared attention

Woody Rousseau, Salvatore M. Anzalone, Mohamed Chetouani, Olivier Sigaud, Serena Ivaldi

***Abstract*— In typical developmental social robotics scenarios, a robot learns new skills or high(er)-level representations of the objects and its environment through a rich interaction with a human tutor. In this paper, we focus on the importance of joint attention and mutual engagement for a natural interaction between the robot and its caregiver, studying the case where the iCub humanoid robot learns the name of objects. We briefly describe our software and control architecture, particularly the way we realize a seamless joint attention mechanism on the robot that allows the human and the robot to interact in a natural way, such as humans would do. We investigate the interaction performance and the response of the human caregiver during the interaction. Overall, our experimental results confirm the importance of joint attention for a more natural human-robot interaction.**

I. INTRODUCTION

Developmental robotics aims at endowing robots with learning capabilities similar to that of infants so that they can build their own representations of the surrounding world and their own skills [2]. In human infants, interaction with human caregivers is a key component of this learning process. To take this important aspect into account, developmental social robotics has emerged recently as a new field at the intersection between developmental robotics and Human Robot Interaction (HRI), which is intended to build social robots able to behave adequately in the presence of humans.

The focus of this new field is on the social signals that play a key role in driving the interaction between the robot and its human caregiver in two directions [1]. From the human to the robot, these signals are mainly intended to facilitate the robot's learning process. From the robot to the human, they are rather intended to induce in the human the feeling to have a natural interaction with an intelligent agent, such as a human partner. Therefore, interaction must be real-time and multimodal.

Two hallmarks of human-human interaction are mutual engagement and joint attention [9]: mutual engagement is the act of gazing to the face of the other, to establish and maintain a visual contact; joint attention is the act of gazing to a target and to drive the attention of the other such that the two partners gaze at the same target or look in the same direction. Through a combination of mutual engagement and joint attention, it is possible to capture and maintain the attention of the partner, focused on a specified object, for instance to share knowledge about it [10]. Finally, the combination of the two is such that that a natural connection (engagement) is established between the partners, facilitating communication [12].

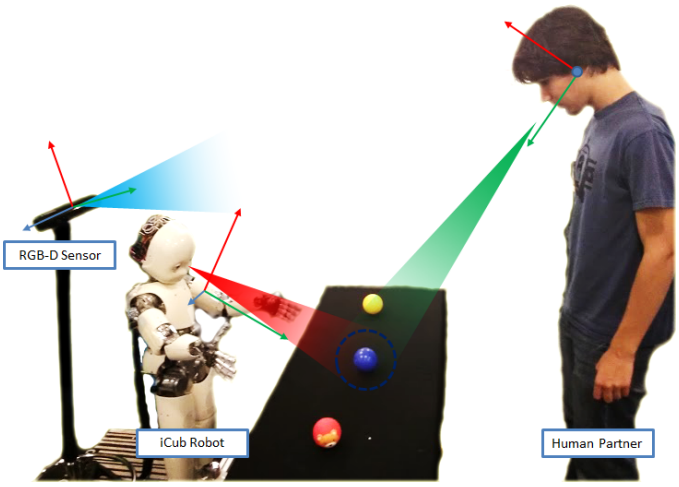


Fig. 1. The experimental setup.

This paper takes a typical developmental social robotics scenario and focuses on the interdependency between the HRI performance and the evolution of the interaction by the human caregiver. We show how a robot can learn information about simple objects in a social way, through attention sharing with a human. As shown in Fig. 1, the iCub robot, able to recognize the gaze of humans, is instructed about some objects. In a first stage, the robot learns a label for each object. In this case the object selection process is guided by the robot, using its gaze upon an object and asking a question about it. After this training step, the robot reports to its partner the learnt label: in this case, the selection process is guided by the human partner who looks towards the desired object, while the robot follows his/her gaze. Previous works such as [6] show the importance of initiating, ensuring and responding to joint attention, and those aspects are placed at the center of our experiments.

Such scenarios are not uncommon, especially when studying joint attention, but frequently many constraints are imposed on subjects, thus biasing the interaction. In short, because of the complexity of the computational aspects required to estimate the partner's proxemics variables, researchers often relies on external wearable devices or simplified object pointers, which make interaction atypical and far from natural. Constraints may include labeling items using

All the authors are with the Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie. Paris, France. Mail:{anzalone,ivaldi}@isir.upmc.fr

This work was supported by the French National Research Agency through the project MACSI (ANR 2010 BLAN 0216 01), the TecSan program (project Robadom ANR-09-TECS-012), the Investissiments d'Avenir program (project SMART ANR-11-IDEX-0004-02) within the EDHHI project, and by the European Commission, within the CoDyCo project (FP7-ICT-2011-9, No. 600716).

a paper pointer to select the target of joint attention [6], using an Eye-Tracker to achieve accurate gaze tracking results [17] or an offline gaze estimation system [18]. All those constraints negatively impact the natural aspect of the interaction, and to our knowledge, strong efforts are required to avoid them all.

In this paper, several challenges are addressed. The first is to integrate several modules in the existing iCub's software architecture [7], allowing the subject to naturally interact with a socially intelligent robot. The second is to make this interaction as natural as possible and to highlight the real perks of relying on joint attention and mutual engagement in such situations. Finally, the experimental results as well as the feedback obtained from subjects through a questionnaire show how more natural interactions make it easier and how joint attention was paramount to its success.

II. METHODS

Learning about the environment in a social way is a very high level task that requires several kind of abilities, in particular on the modeling of the human partners and of their behavior, on the understanding of the shared environment, and on the achievement of coherent behaviors which must be readable by humans.

As shown in Fig. 1, an iCub robot shares a common space with its human partner. Several objects are also placed between the robot and its partner. The iCub robot is made able to engage partners and to induce joint attention to a random object placed on the table. The robot learns about its environment by being able to capture and learn a label for each of the objects focused on. In a second stage the human is able to engage and induce joint attention on the robot to focus on objects and to ask about their label.

To achieve such complex behaviors, we enriched a previous software architecture built in the context of the MACSI project[1] [**?**] with several modules able to communicate with each other and specialized on solving a particular subproblem. Modules have been developed for both Robot Operating System (ROS) and Yet Another Robot Platform (YARP) middlewares, made able to communicate through a generic middleware bridge.[2] In Fig. 2 a sketch of the main modules is depicted.

The robotic system is able to recognize the presence of humans and their behavior using a conveniently placed RGB-D sensor, while it is able to recognize objects using the cameras placed on the iCub's eyes. Communication tasks take place using both verbal and non-verbal communication: verbal communication is achieved through speech synthesis and using a simple speech recognition system; non-verbal communication calls upon the ability of the robot to recognize gazing and to behave in a similar way. Since the focus of this study is on the attention sharing between the human and the robot, rather than using complex symbol grounding systems, names of the objects are learnt in a simple way as labels associated to the respective object features.

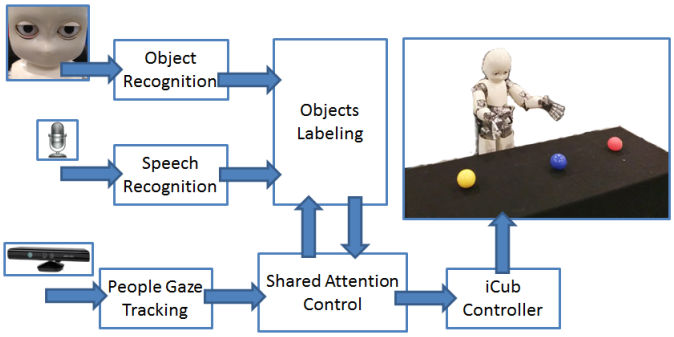


Fig. 2. A sketch of the pipeline of the system.

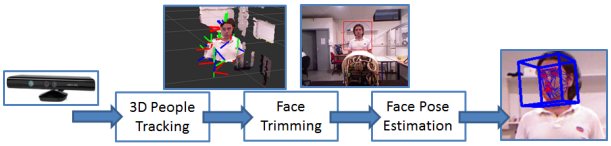


Fig. 3. Capturing the pose of the head.

A. JOINT ATTENTION SYSTEM

Humans are one of the main sources of information for social robots. Developing a model able to accurately detect their presence and their behavior is thus essential for all HRI applications. The presented robotic system is able to gather information about the human presence, about their posture, their head movements, and other facial behaviors during direct interactions. Humans are modeled through a 3D description of their body in terms of Cartesian coordinates of each joint, and in terms of the yaw, pitch and roll angles of their head.

As shown in Fig. 3, data perceived by the RGB-D sensor is elaborated by a Skeleton Tracking system able to detect the presence of humans. When a human enters the field of view of the sensor, his/her articulations are detected and tracked in the environment, in three Cartesian dimensions. The position of the head is then estimated by a Kalman Filter, and is then used to find, in the RGB image, an area in which the face should appear. A Face Tracking algorithm is then applied to the trimmed image to retrieve the head's pose. In order to prevent additional noise, this information is stabilized using a second Kalman filter. The pose is eventually buffered on a port, readable within the finite state machine controlling the different phases of the interaction.

1) 3D People Tracking: A multiple skeleton tracking system provided by OpenNi is used to detect human activity and trace in the space human bodies in terms of their joint positions [13]. In the present interaction, only the head pose is used. In more details, the algorithm performs a background subtraction applied to the depth data perceived by the sensor to distinguish the body of each person from the static

[1] http://macsi.isir.upmc.fr

[2] For more details, see `http://wiki.icub.org/wiki/UPMC_iCub_project/YARP_ROS_bridge`.

environment. Then, through a per-point approach, the depth data pixels of each body are classified according to depth invariants and 3D translation invariant features which aim at assessing which part of the body each depth pixel belongs to. A total of 31 patches distributed among the different parts of the human body have been considered and classified. The training processes performed by OpenNi employes a database of 500k frames captured in several scenarios, such as running, kicking or dancing. From each patch, it is then possible to extract the position of each joint of the body according to its density.

2) Head Pose Estimation: The estimated 3D position of the head of the human partner can be back-projected to the RGB image captured by the sensor to select the area in which the head should appear. A face tracking algorithm can be applied to the accordingly trimmed image to estimate a model of the head in terms of yaw, pitch and roll. Getting the head orientation information solely from the head pose is not extremely accurate, but child developmental research [11] shows that this is the only way of estimating the gaze until a certain stage of development, usually near the tenth month. Informing the human partner of this inability of the robot to use the gaze information given by the eyes was however made necessary by this simplification.

The Constrained Local Models (CLM) [5] algorithm has been chosen as the method to estimate the pose of the head of the human partners. Such an algorithm is a particular case of the Active Appearance Models (AAM) [4] and tries to model human faces through a statistical description based on a set of landmarks: the face shapes are deformed iteratively according to the landmark positions in order to find a best match with the actual image.

The CLM algorithm is a slight variation of the AAM algorithm, in which the appearance model is built by using local features, patches of pixels around each key-point instead of the wrapping of the whole image as used by AAM. To adapt the actual face to the current model, the Nelder-Meade simplex algorithm is also employed.

B. OBJECT LEARNING

The presented system focuses on learning the name of simple objects in a social way. The learning process is guided by a mix of verbal and non-verbal communication between a social robot and a human teacher. Without losing its general aspect, the system has been constrained to the recognition of objects strongly described by their colorimetric characterization, such as baby's toys. Moreover, to capture the names of the objects, a labeling system based on a simple speech recognition system has been developed.

In the training session the robot guides the gaze of the human partner towards a random object. Then, it asks for a label and it associates such a label with the features of the perceived object. After this learning stage, the human partner can guide the gaze of the robot to an object: if the perceived features are close to the ones of an already labeled object, the robot assumes it as a known acquaintance and communicates verbally the associated label.

1) Objects Characterization: The object recognition system used in the presented architecture relies on the information perceived by the camera mounted on the eyes of iCub. The robot's head first points to the location gazed upon by the human partner over a table placed in front of them. The object recognition system aims at detecting, locating and extracting significant features of the objects perceived, as shown in Figure 4, assuming they are characterized by their color.

Fig. 4. The objects located by the robot's camera.

A first elaboration of the image is performed through an edge detection process to locate circles, using the Hough Gradient method [8] which retrieves the part of the image in which objects should appear. After a trim of the image according to the obtained results, in order to remove any contour not located on the table, a second stage of refinement is performed through a Canny edge detector [3] and trough the Suzuki algorithm [14] for finding contours. Since the eligible object is the one pointed by the gaze of both the human partner and the robot, the closest shape found near the center of the image is considered and selected as the object to classify. Finally, an average of the color of the pixel inside the contour is calculated and used as main feature of the object. The color space considered is CieLab space, in which the distances between colors match the differences perceived by the human eye [15]. A Delta-E[3] distance is used along with a threshold to distinguish already known objects from new ones.

2) Names Labeling: A speech recognition system based on CMU Sphinx is used to capture the name of the objects presented to the robot. Through a Lavalier microphone, the voice is captured in a clear way, without limiting the movements of the human subject and without being limiting the spontaneity of the interaction with the robot.

As shown in Table I, a restrictive grammar has been created to achieve the simple task of capturing the color of the object. The grammar is designed to give the possibility to the subjects to include the color of the object in a more complex sentence. Such sentence can be composed by an optional verb, an article, and the name of the object. This grammar helps filtering unwanted step words, extracting just the most important informative content, the color of the object.

[3] $\Delta E^* = \sqrt{(L_1 - L_2)^2 + (a_1 - a_2)^2 + (b_1 - b_2)^2}$ where L, a and b are the coordinates of the colors to be compared in the CieLab color space.

```
#JSGF V1.0;

grammar explanation;

public <explanation>=
      <verb><article><label><object>;

<verb>= [ its | it is ];
<article>= [ /5/ al/1/ the ];
<label>= ( green | yellow | orange | red | blue );
<object>= [ ball ];
```

TABLE I

THE GRAMMAR USED BY THE SYSTEM.

III. EXPERIMENTAL PROTOCOL

The robot learning process is supervised by the human partner. At the beginning, the robot waits to establish mutual eye engagement with the human partner. In the first training phase, the robot learns the labels of the objects by the subjects. In the second test phase, the human gazes to one of the objects, and the robot responds with the learnt label. Gaze is important for both agents: during the training phase the robot gazes and induces joint attention with the human; in the second stage the subject gazes and induces a joint attention response from the robot.

Algorithm 1 Robot's Training Session

```
procedure TRAINING(partner,objects[])
    Waiting for partner-robot mutual engagement
    for all Objects on the table do
        Robot gazes to the object
        Robot says "What is this?"
        Robot waits for the label
        Robot greets
    end for
end procedure
```

Algorithm 2 Robot's Testing Session

```
procedure TESTING(partner,objects[])
    Waiting partner-robot mutual engagement
    loop
        Robot says to "Look at an object"
        Robot waits for partner's gazing
        Robot gazes accordingly
        Robot recognizes the object and vocalizes the label
    end loop
end procedure
```

3) Two steps interaction: In the training phase, described by Algorithm 1, the robot leads the interaction, driving the focus of attention of the person towards an object, using its gaze. Here, the readability of the robot's behavior is the key factor for a successful interaction. In the second phase, described by Algorithm 2, the human gazes first, and in response the robot gazes back at the same point. Here, the key ability of the robot is the estimation of the gaze.

4) Data Collection: A population of 7 subjects equally distributed among sex (4 men, 3 women), and age (25.5 years, $\sigma = \pm 2.5$) has been randomly chosen mostly from the ISIR laboratory but also from the UPMC campus. Each of them has been separately involved in the experiment. Contamination bias has been prevented by keeping the people who already took part away from those who had not yet. Each subject has been informed about the teaching task with the iCub robot through an instruction paper, in order to give to each of them the same amount of knowledge about the system. In order to keep the interaction natural, subjects were mostly asked to be attentive to the cues given by the robot, to be able to follow the evolution of the interaction. Three colored balls were placed on a shared table between the human and the robot, on the left, on the center and on the right, and used as objects to label. Those objects were deliberately put sparse, not because of the contour detection module which would work with those further packed, but because the yaw estimation of the subject's head would be made harder, since the thresholds used for gaze estimation would have to be reduced. This would have certainly had a negative impact on the overall results of the experiments.

Six additional subjects were presented with a variant of the experiment, in which the object is selected by the human during the learning phase, through his gaze. This made the first phase the most error-prone as it required a correct estimation of the human gaze and of the color object, both through image processing and voice recognition.

In all presented results, the main focus is the first population of 7 subjects. However, feedback statistics were obtained by comparing both populations.

IV. EXPERIMENTAL RESULTS

Experiments of the system have been performed in order to evaluate the system in terms of HRI performances and objects learning.

5) HRI Evaluation: A first evaluation was conducted on the gazing behavior of both the robot and the human. During the training stage, the robot tries to induce attention towards an object placed on a shared table: here, the frequency of gazing by the human partner toward the same object was counted. As shown in Table II (line 1), the high level of such frequency can be considered as a very good readability from the humans of the gaze of the robot. As shown in [**?**], this high readability does not mean that the information brought by the robot's gaze is at the same level as information one could get from another human partner's gaze. Indeed, subjects were in the experiment highly focused on the robot's gaze, given the fact that the interaction relied on the proper recognition of the gazed at object.

During the test stage, the human gazes upon the objects, trying to induce attention on the robot. Here, the main ability of the robot is to recognize the gazing of the human partner. The frequency of the robot's correct response on human gazing was taken in account. As shown in Table II (line 2), the high level of such frequency evaluates the robot ability on sharing attention with its partner over the same object.

To better understand this second result, a pre-evaluation of the system was conducted to evaluate the performances of the CLM-based algorithm: through the use of a motion capture system composed by 3 CodaMotion CX1-800 units, the ground truth of the head movements of three people was retrieved. Each person was equipped with three markers placed on the head, one over the left, one over the right ear and one on the front of the head. Each subject stood up in front of the robot at around 1.5m of distance from an RGB-D sensor, moving the gaze on the left, on the right, upward and downward. Table III shows the results of the comparison between the ground truth obtained with the motion capture system and the information obtained through the RGB-D sensor, using the gaze recognition algorithm presented above. The low recognition rate on the head's pitch of the human partner explains well the difficulty of the system to retrieve correctly the attention focus of the human.

Inducer Subject	JA Induced
Robot	100%
Human	65.79%
Overall	82.89%

TABLE II

JOINT ATTENTION INDUCTION EVALUATION: THE ROBOT INDUCES JA ON THE HUMAN; THE HUMAN INDUCES JA ON THE ROBOT.

Head Pose	CLM
Pitch	49%
Yaw	93%
Overall	71%

TABLE III

THE HEAD'S POSE ESTIMATION PERFORMANCES USING THE CLM BASED APPROACH.

System	Performance
Object Recognition	92.72%
Speech Recognition	76.19%
Overall	84.46%

TABLE IV

PERFORMANCES OF THE TWO MAIN OBJECT LEARNING SYSTEMS: THE OBJECT RECOGNITION AND THE SPEECH RECOGNITION MODULES.

6) Objects Learning: The performances of the labels learning process is influenced mainly by the capability of the object recognition system to correctly extract the correct feature from the object perceived, and to accurately perceive the speech of the person. An evaluation of both subsystems was performed in each case in which these modules were used.

As depicted in Table IV, while the performance on features extraction obtained by the color based object recognition system is good, lower performance was obtained by the speech recognition system. This result is very likely biased because the speech recognition system was in English, which was not all the subjects' mother tongue.

A final analysis of the whole learning system was taken in account, by simply matching the object gazed by the human in the test stage, and the label asserted by the robot. In this case a sum of the different errors of each module occurs: the errors on object recognition, on speech recognition and, mainly, on head gaze estimation, reduces the whole performance of the system, that is able to correctly assess the taught label in the 50% cases over 38 trials.

	Average Scoring (1-5)	
	Main Group	Variant
The robot is intelligent	2.86	2.33
The robot is a better partner for a cooperative task	3	2.33
The robot is involved in the naming process	4	4.5
The robot behaves like a child in development	2.29	3.17

TABLE V

AVERAGE LIKERT SCALE SCORING FOR BOTH SUBJECT GROUPS.

7) Subjects Feedback: Right after the interaction, a questionnaire containing the same questions for both variants of the experiment was submitted to all the involved subjects. Questions presented aimed at assessing how people felt about the robot as a partner and about its behaviors. Subject rated each statement from 1 to 5 following the model of a Likert scale, according on how much they agree with them. The most significant results are given in Table V.

In the main experiment during the training stage the robot guides the selection of the objects to be labeled: in this case the overall behavior of the robot is seen as more intelligent and the robot is perceived as an efficient partner for cooperative tasks. Statements evaluating technical aspects, such as how easy it was to identify the object designated by the robot, or how easy it was to figure out when the robot expected an answer, were also ranked higher.

In the experiment variant in which the objects to learn are selected by the human, despite the technical aspects that were ranked with a low score, the robot was seen as more involved in the naming process: this response is probably given due to the fact that the robot responds well to the joint attention and arouses a sense of presence in the human partner. Moreover, such learning stage tends to remind subjects of teaching activities with a child, rather than with an adult: this explains why the robot is perceived in this experiment variant as a developing child. Such results show that, despite the fact that joint attention and mutual engagement bring some technical difficulties, they are able to arouse on the human partners a sensation of 'intelligence' of the robot, as the robot is involved in the interaction.

At last, a high variance ($\sigma^2 = 2.57$), associated with the question related to how easily people felt the robot could determine their own gaze, is found. This can be explained by the fact that subjects handle the gazing in different ways: some people choose to look to the items without moving too much their head, while others exaggerate their head movements, trying to facilitate the recognition of their

behaviors by the robot. In any case, both behaviors trigger a decrease of the performances of the recognition of people's gaze.

V. CONCLUSION AND FUTURE WORK

We presented a shared attention system that improves the capability of robots to learn objects labels through social tutoring. Our system is aimed to ensure an effective joint attention mechanism avoiding the use of wearable gaze tracking devices on the human, which would have improved the quality of human gaze estimation but perturbed the natural interaction between the human and robotic partners. Despite this intrinsic limitation, we were able to ensure the quality of the gaze estimation, and realize HRI experiments in a more natural interaction setting.

The results obtained both through the success rates of the different subsystems as well as through the subjects feedback sheds light on the importance of joint attention for staging a natural HRI. Our results relate to the ones obtained by Huang & Thomaz in [6], where the interaction between the human and the robot was simplified by tags on the objects.

Future work includes the strengthening of our preliminary results, by doing more experiments with more naive subjects, and presenting more statical evaluations about the effective performance of the gaze tracking estimation with a ground truth (an external device such as an eye-tracker): this will be done only to assess the robustness of our system, because the crucial aspects of our experiments is the preservation of the naturalness of the interaction between partners, without intermediate elements or devices. We may include the possibility of a fine-grained temporal analysis, including the states sequencing of the finite state machine and some cause-effect comparison obtained focusing on the head's pitch and yaw data from both the human and the robot. Such studies would produce some accurate estimation of the adequate timings and sequencing of the non verbal and verbal communications with humans [12]. However, the learning process does not include the expression of uncertainty from the robot, causing the subject to sometimes repeat the label, causing difficulties in such an analysis. More complex machine learning based interactions such as the one presented in [**?**], have the robot give various feedback to its human teacher, providing updates on its learning status, including through uncertainty. Works such as [**?**] go even further, by allowing the robot to react to incomplete information given by the human partner, thus not only letting the partner know there is uncertainty, but also reacting properly to such situations.

Finally, the interaction task presented in this paper is limited to a simple labeling case. It could benefit from facing more complex symbol grounding situations such as the one addressed in [16] [19] or from associating people to known items, making the robot's world representation more complex as well. Although it would make the technical aspects more challenging, it would eventually trigger interesting feedback results from human partners and would enable further progress in the field of developmental robotics.

REFERENCES

[1] H. Admoni and B. Scassellati. Robot gaze is different from human gaze: Evidence that robot gaze does not cue reflexive attention.

[2] S. Anzalone and M. Chetouani. Modeling social signals during human robot interactions. In *Fifth Workshop on Gaze in HRI, 2013 Human-Robot Interaction Conference*. IEEE, 2013.

[3] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida. Cognitive developmental robotics: a survey. *IEEE Trans. Aut. Mental Dev.*, 1(1):12–34, 2009.

[4] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Analysis Machine Intell.*, (6):679–698, 1986.

[5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Trans. Patt. Anal. Mach. Intell.*, 23(6):681–685, 2001.

[6] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *Proc. British Machine Vision Conference*, volume 3, pages 929–938, 2006.

[7] C.-M. Huang and A. L. Thomaz. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *Proc. IEEE RO-MAN*, pages 65–71, 2011.

[8] S. Ivaldi, N. Lyubova, D. Gérardeaux-Viret, A. Droniou, S. M. Anzalone, M. Chetouani, D. Filliat, and O. Sigaud. Perception and human interaction for developmental learning of objects and affordances. In *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, Osaka, Japan, 2012.

[9] S. Ivaldi, S. Nguyen, N. Lyubova, A. Droniou, V. Padois, D. Filliat, P.-Y. Oudeyer, and O. Sigaud. Object learning through active exploration. *IEEE Transactions on Autonomous Mental Development*, pages 1–18, 2013. to appear.

[10] C. Kimme, D. H. Ballard, and J. Sklansky. Finding circles by an array of accumulators. *Communications of the Association for Computing Machinery*, 18:120–122, 1975.

[11] G. Knoblich, S. Butterfill, and N. Sebanz. 3 psychological research on joint action: Theory and data. *Psychology of Learning and Motivation-Advances in Research and Theory*, 54:59, 2011.

[12] A. Lockerd and C. Breazeal. Tutelage and socially guided robot learning. In *Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 4, pages 3475–3480. IEEE, 2004.

[13] A. N. Meltzoff. 'like me': a foundation for social cognition. *Developmental science*, 10(1):126–134, 2007.

[14] A. N. Meltzoff and R. Brooks. Eyes wide shut: The importance of eyes in infant gaze following and understanding other minds. *Gaze following: Its development and significance, ed. R. Flom, K. Lee & D. Muir. Erlbaum.[EVH]*, 2007.

[15] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. Recognizing engagement in human-robot interaction. In *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction (HRI)*, pages 375–382, 2010.

[16] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. Which one? grounding the referent based on efficient human-robot interaction. In *RO-MAN, 2010 IEEE*, pages 570–575. IEEE, 2010.

[17] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *IEEE CVPR*, 2011.

[18] S. Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.

[19] M. Tkalcic and J. F. Tasic. *Colour spaces: perceptual, historical and applicational background*, volume 1. IEEE, 2003.

[20] Y. Yoshikawa, T. Nakano, M. Asada, and H. Ishiguro. Multimodal joint attention through cross facilitative learning based on μx principle. In *IEEE Int. Conf. Development and Learning*, pages 226–231, 2008.

[21] C. Yu, M. Scheutz, and P. Schermerhorn. Investigating multimodal real-time patterns of joint attention in an hri word learning task. In *ACM/IEEE Int. Conf. Human-Robot Inter.*, pages 309–316, 2010.

[22] Z. Yucel, A. A. Salah, C. Merigli, and T. Meriçli. Joint visual attention modeling for naturally interacting robotic agents. In *Int. Symp. Comp. Inf. Sciences*, pages 242–247, 2009.

[23] O. Yürüten, K. F. Uyanık, Y. Çalışkan, A. K. Bozcuoğlu, E. Şahin, and S. Kalkan. Learning adjectives and nouns from affordances on the icub humanoid robot. In *From Animals to Animats*. 2012.

Computational Model of the Biologically Inspired Cognitive Architecture for Human Robot Interaction

Evren Dağlarlı, Hatice Köse, and Gökhan İnce *Member, IEEE*

***Abstract*— People having disabilities (especially hearing impaired) need better interaction with their environment and humanoid robots can help in this mission. This is a challenging issue because high level linguistic capabilities are required for communication and the humanoid robot should realize sign language processing in the case of interaction with hearing impaired people. In the human brain, some cortical and cerebral zones are responsible from cognitive functions. Therefore in this paper, it is required to consider a novel brain inspired robot control architecture consisting of artificial emotion imitated computational cognitive and limbic system anatomical structure of human brain includes some critical zones named cerebral cortical zones such as Orbitofrontal Cortex, Sensory Cortex, Thalamus, and Limbic system components (e.g., Basal-Ganglia, Amygdala). Designed architecture is embedded into humanoid robot NAO H25. An interaction game play was demonstrated by a simulation. Finally results are observed and discussed.**

I. INTRODUCTION

People having disabilities (especially hearing impaired) need better interaction with their environment and humanoid robots can help in this mission. This is a challenging issue that high level linguistic capabilities are required for communication and the humanoid robot should realize sign language processing in the case of interaction with hearing impaired people [1]. As a visual communication tool, the sign language constitutes complicated gestural tasks including all upper torso physical behaviours of the humanoid robot such as arm, hand, finger, neck (head) movements. These can be achieved by having a complex cognitive architecture in the humanoid robot.

Modern AI architectures developed for Human-Robot Interaction (HRI) have been based on biologically inspired deliberative transactions, which constitute the principles of cognitive science and neuroscience. Therefore, in the global society, there is an increasing demand to an automated solution, which solves the communication problems. In the nature, these concepts are biologically realized in the human brain. The computational approximation of human brain should exhibit features of higher cognitive abilities such as deliberative planning (task allocation), re-organizing (learning), communication, consciousness and self-awareness. When investigating from the computational intelligence, computer science and engineering viewpoint, consciousness and self-awareness are still uncovered philosophical concepts. Consciousness includes the creatures having perception, thoughts instinctual feelings, autonomous planning of behaviours, self-acting, awareness of themselves, selective attention, performing simultaneous decision making and recursively task processing (behaviour execution) in an infinite loop fashion.

Thus, the purpose of the proposed Artificial Intelligence (AI) framework is to enable that humanoid robot -as an artificial live form- provides the perfect interactional conformity with humans and uncertain dynamic environment; and when the robot establishes an interaction with a disabled person, designed computational cognitive architecture of the humanoid robot takes over the mission of carrying out a service or rehabilitation task. In this paper, we demonstrate the capabilities of our cognitive architecture on an interaction game between a humanoid robot and people having disabilities to provide a suitable testing environment. Section II provides background information about nature inspired approach and the following section gives the detailed design instructions of the computational cognitive architecture. Section IV presents the simulation results followed by the discussion and potential applications of the proposed method in the next section. The final section concludes the paper.

II. NATURE INSPIRED APPROACH

In the human brain (Figure 1), some areas of cortical and cerebral zones are responsible from cognitive functions. These cortical zones are sensory cortex, orbitofrontal/prefrontal cortex, associative cortex, thalamus, cerebellum, brain stem, limbic system (e.g., amygdala, basal ganglia, hippocampus) [2][3].

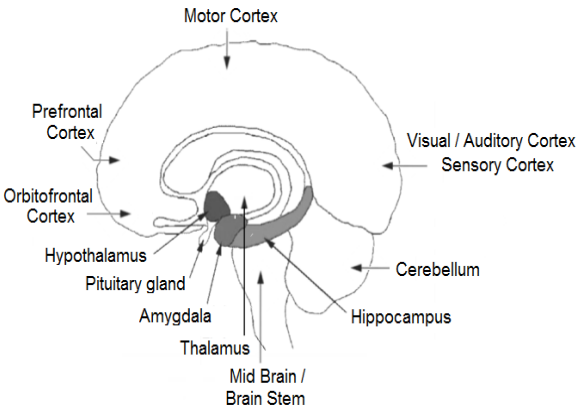


Figure 1. Human brain cortical zones [3]

This work was supported by The Scientific and Technological Research Council of Turkey under the contract TUBITAK KARIYER 111E283.

* Evren Dağlarlı, Hatice Köse and Gökhan Ince are with the Faculty of Computer and Informatics, Istanbul Technical University, Istanbul, Turkey.
Email: {evren.daglarli, hatice.kose, gokhan.ince}@itu.edu.tr

According to neuroscientific evidence, associative learning of emotional responses (states) and gestural expressions come from the amygdala [4][5]. Also behaviour selection events are realized in basal ganglia [6][7]. Hippocampus is an important area for short term/long term memory representation and provides information about spatial navigation. Associative cortex is responsible of high level cognitive activities such as meta-level reasoning and inference, high level deliberative decision making. Motor cortex contributing to planning of low level actions, representation and generating behaviours, complex task decomposition, permissions of task access, realizing actuator commands [2]. Sensory cortex is divided into some sub-modules, but it mainly works on feature representation and recognition. Thalamus is responsible of segmentation of perceptual sensorial raw data and applying pre-processing on them [2].

Next, we discuss computational viewpoint of self-awareness, consciousness and describe developing a computational model imitated by artificial emotion of the brain inspired by general cognitive architecture which constitutes a limbic system based on these concepts. In fact, we are trying to develop a model, which captures the more realistic properties of the limbic system which are mainly known as the Amygdala-Orbitofrontal Cortex system.

III. BRAIN INSPIRED COGNITIVE ARCHITECTURE

Nowadays, in order to comply with humans and achieve improved human-robot interaction, the computational model of AI architecture in the humanoid robots needs higher level cognitive abilities which support awareness and consciousness aspects. The biological nature of human brain contains these features. Therefore, in Figure 2, it is required to consider a novel brain inspired robot control architecture consisting of artificial emotion imitated computational cognitive and limbic system.

Anatomical structure of human brain includes some critical zones named cerebral cortical zones such as *Orbitofrontal Cortex*, *Sensory Cortex*, *Thalamus*, *Limbic system components (Basal-Ganglia, Amygdala)*. Functional analogies of these units and their pathways between them allow us to construct brain-inspired frameworks of computational models of these cerebral cortical zones.

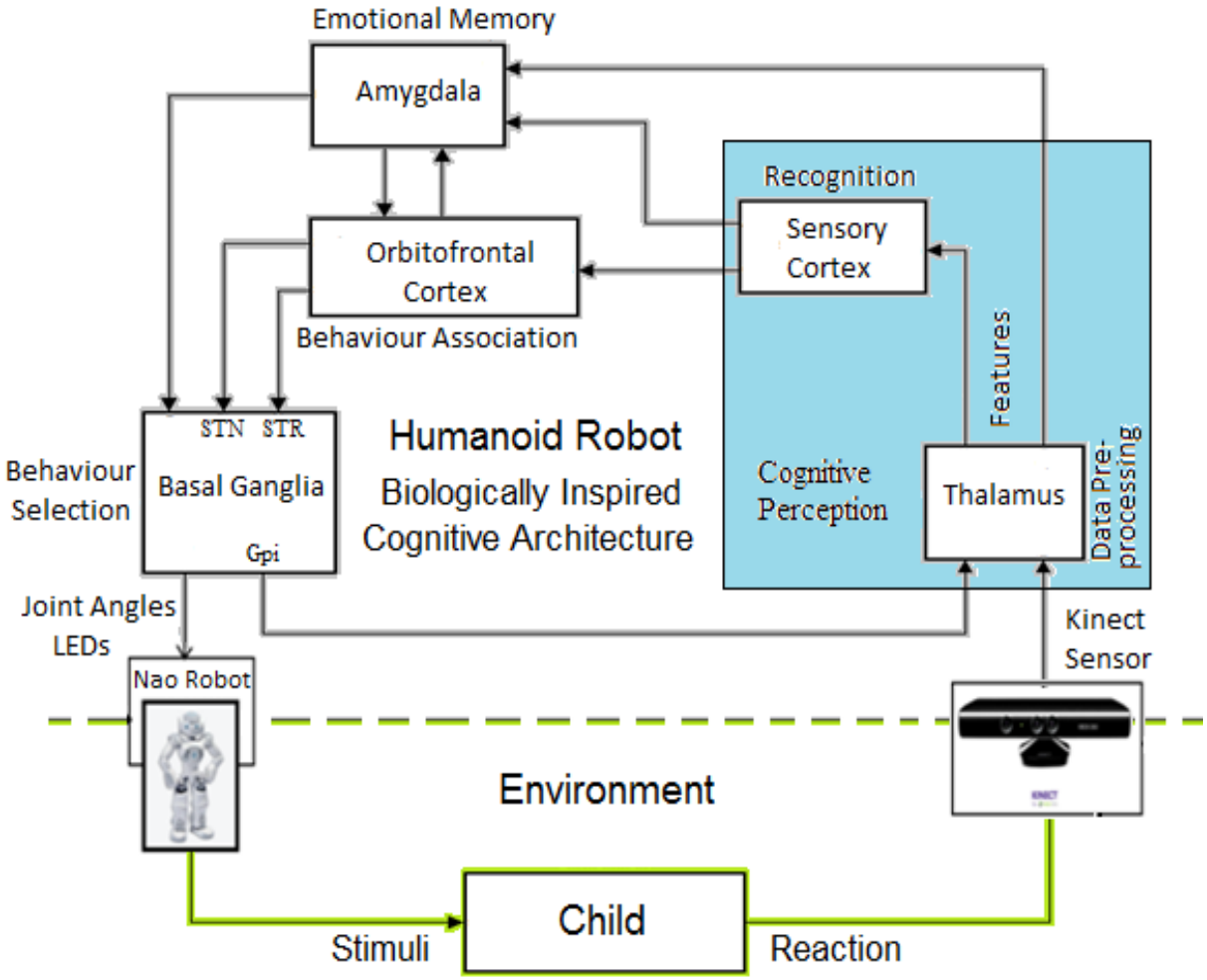


Figure 2. Computational framework of the biologically inspired cognitive architecture

As a nonlinear dynamic model, the major developmental novelty in our paper is that the Cellular Neural Network (CNN) model is employed to construct the framework of biologically inspired cognitive architecture and better representation of the input-output topology of the system. Also in this model, an adaptation procedure is applied on some cortical units (e.g., *Sensory Cortex, Orbitofrontal Cortex and Amygdala*). In this way, the internal states of these units are converged to some specific points. These points describe characteristics of cortical units in the architecture. The adaptation procedure allows the system to learn the decision boundary for behavior selection.

All components of the architecture are modeled in this neural network.

$$x_{k+1} = \Gamma . x_k + A . y_k + B . u_k + I \quad (1)$$

Where x_k represents the states of the system, y_k the output values of the each component, u_k input values of the each component [8]. In the cellular neural network structure, which describes a mathematical model of the cognitive architecture, the components x_k, y_k and u_k are the vectors with 7x1 dimension as the default case of one executable behavior. Also indices (k and k+1) are expressed as current and next values in a variable of the system. Depending on the increase or decrease of the number of the behaviors in the architecture, dimensions of the matrices and vectors are grown or contracted accordingly.

$$X_k = \begin{bmatrix} Str(k) \\ Stn(k) \\ Gpi(k) \\ Th(k) \\ SC(k) \\ OC(k) \\ Am(k) \end{bmatrix} \quad (2)$$

where parameters Str, Stn, Gpi, Th, SC, OC, Am are striatum, subthalamic nucleus, globus pallidus, thalamus, sensory cortex, orbitofrontal cortex and amygdala respectively. The output vector in the system model has a form of $y_k = [f(X_k)]$. This output vector is obtained after the state vector X_k is transferred via a neural activation function. In our system, the activation function is selected to restrict the incoming signal, which is in the form of

$$f(x) = \frac{1}{2}(\tanh(3(x - 0.5)) + 1) \quad (3)$$

$$y_k = \begin{bmatrix} f(Str(k)) \\ f(Stn(k)) \\ f(Gpi(k)) \\ f(Th(k)) \\ f(SC(k)) \\ f(OC(k)) \\ f(Am(k)) \end{bmatrix} \quad (4)$$

The matrix Γ describes the inner dynamics of the system components. The values of this matrix will be updated by using an adaptation mechanism. Matrix *A* gives us a neural topology between output connections of the system components (cerebral zones in the computational model of the brain). Matrix *B* defines a neural topology between input connections of the system components.

$$A = \begin{bmatrix} 0 & 0 & -a_{13} & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{23} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -a_{34} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{45} & 0 & a_{47} \\ 0 & 0 & 0 & 0 & 0 & a_{56} & a_{57} \\ a_{61} & a_{62} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{73} & 0 & 0 & 0 & 0 \end{bmatrix} \quad (5)$$

$$B = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & b_{16} & 0 \\ 0 & 0 & 0 & 0 & 0 & b_{26} & 0 \\ b_{31} & b_{32} & 0 & 0 & 0 & 0 & b_{37} \\ 0 & 0 & b_{43} & 0 & b_{45} & 0 & 0 \\ 0 & 0 & 0 & b_{54} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & b_{65} & 0 & 0 \\ 0 & 0 & 0 & b_{74} & b_{75} & 0 & 0 \end{bmatrix} \quad (6)$$

Finally, the vector I having 7x1 dimensions as a bias element, expresses outer effects to the system. These effects may be disturbance signals or unknown exciting signals from the environment.

Since our biologically inspired cognitive architecture is composed by seven cortical zones (modules), matrices Γ, A, B inside the model can be considered as weight matrices with the dimensions of 7x7 by default.

$$\Gamma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_{SC} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda_{OC} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{Am} \end{bmatrix} \quad (7)$$

The matrix Γ in the cellular neural network structure for cognitive architecture is the matrix with 7x7 dimensions in case of one behavior to be executed.

Because the parameters λ_{SC}, λ_{OC}, λ_{Am} in this matrix are updated by an adaptation mechanism, they are also expressed as the inner dynamics of Sensory Cortex, Orbitofrontal Cortex and Amygdala modules respectively.

When we deal with a biologically inspired cognitive architecture, the cerebral zones (*Thalamus, Sensory Cortex, Orbitofrontal Cortex, Amygdala and Basal Ganglia*) form the architectural components (modules), which correspond to the related rows in the cellular neural model.

A. *Cognitive Perception*

The general perception architecture namely "Cognitive Perception" performing cognitive abilities such as recognition, generalization, selective attention, classification/clustering, (supervised/unsupervised) learning, sensorial data fusion and interpretation are introduced here.

Before performing these tasks, some support module named *Thalamus* is considered. It is responsible for data preparation and interpretation tasks such as segmentation, skeletal modeling. For broadcasting, the input stream coming from *Thalamus* module is opened and fed into computational model of "Sensory Cortex".

B. *Thalamus*

Thalamus module can be considered as feature extractor part of Cognitive Perception structure. This module accepts raw sensory information stream [2]. Some preprocessing steps are applied on this raw data stream such as noise cancellation. Source of sensory information may come from Kinect of Microsoft Xbox. The RGB-D sensor (Kinect), which generates joint spatial coordinates (x,y,z) of skeletal structure for each joint (node) is activated and sends information about visual data (motion) frame by frame [1][9]. Then, spatial coordinates (x,y,z) of all skeletal nodes (joints) are transformed to roll, pitch, yaw angles for all frames of motion. These frames of data stream as the feature broadcast to the Sensory Cortex.

C. *Sensory Cortex*

Broadcasted data (features) coming from "Thalamus" are processed in two main tasks. One of the tasks is recognition task and the other one is selective attention task. The recognition task, which includes supervised learning activities, is usually based on Hidden Markov Models (HMM). In this structure, the recognizer module of sensory cortex generates a dynamic model for every distinct behavior (gesture). According to the data coming from input stream, it determines hidden states (node) and observable variables [9]. As a target vector, data coming from semantic and long-term memory (gestural behavior database) seed into recognizer cycle to perform supervised training algorithm (e.g. Baum-Welch) [1]. This process throws likelihoods related to the generated dynamic model of a gestural behavior (sign words) in the sense of joint -relevant information. Produced raw likelihood values provide us information about recognition process.

$$x(5)_{k+1} = \Gamma(5,j)x_k + A(5,j)y_k + B(5,j)u_k + I(5) \quad (8)$$

The dynamics of the *Sensory Cortex* is represented in fifth row of the cellular neural model. The fifth row of matrix Γ, which expresses inner characteristics of the Sensory Cortex is in the form of $\Gamma(5,j) = [0\ 0\ 0\ 0\ \lambda_{sc}\ 0\ 0]$

$$\Delta\lambda_{SC} = \gamma . u'_k . (y_k - u_k) \quad (9)$$

The equation (9) gives an adaptation rule, which updates the parameter λ_{SC} expressed as inner dynamics of the Sensory Cortex module. As an adaptation coefficient, γ is an empirically set value between [0,1].

D. *Orbitofrontal Cortex*

Orbitofrontal Cortex is one of the major cerebral cortical zones in the human brain. The orbitofrontal cortex (OFC) is a prefrontal cortex region in the frontal lobes of the brain involved in the cognitive processing of decision-making [3]. The OFC is sometimes considered also to be a part of the limbic system [4].

According to neuroscientific connections, Markovian and cellular neural modeling based machine learning techniques are used to achieve the proposed tasks in this paper.

$$x(6)_{k+1} = \Gamma(6,j)x_k + A(6,j)y_k + B(6,j)u_k + I(6) \quad (10)$$

The dynamics of the *Orbitofrontal Cortex* is represented in sixth row of the cellular neural model. The sixth row of matrix Γ, which expresses inner characteristics of the Orbitofrontal Cortex is in the form of $\Gamma(6,j) = [0\ 0\ 0\ 0\ 0\ \lambda_{OC}\ 0]$.

$$\Delta\lambda_{OC} = \alpha . u'_k . (y_k - I_k) \quad (11)$$

The equation (11) gives an adaptation rule, which updates a parameter λ_{OC} expressed as inner dynamics of the Orbitofrontal Cortex module. As an adaptation coefficient, parameter α is an empirically set value between [0,1].

E. *Basal Ganglia*

The basal ganglia (Figure 3) formed by a set of nuclei, primarily takes on action selection tasks, the decision of several possible behaviors to perform at a given time [6][7].

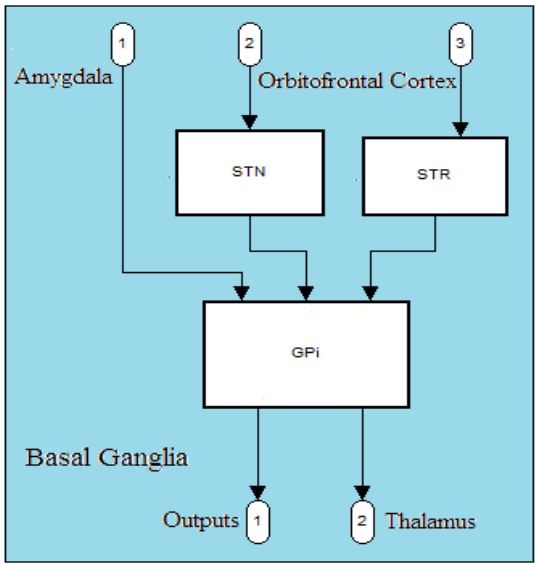


Figure 3. Basal ganglia model

Selection mechanism of the basal ganglia utilizes an inhibitory effect on various motor signals. According to this, some motor signals can be activated or inactivated by inhibition that takes place within the basal ganglia [2]. Also a group of motor signals can be activated together in some time slices [7].

Recent studies show that behavior switching of the basal ganglia is influenced by signals from many parts of the brain, including the thalamus, prefrontal cortex and limbic system components such as hippocampus and amygdala [6].

F. Amygdala

As an architectural overview, amygdala is responsible of emotional memory storage. Also in this memory, activations of the behaviors may be boosted or diminished according to robot's goals [3][4].

The amygdala sends impulses to the basal ganglia for activation of the emotional responses. In this architecture, the amygdala plays the primary roles on the formation and storage of memories associated with emotional events [10]. The stored memory is associated with weighting matrix. Also coefficients of this matrix are updated during the simulation.

$$\Delta\lambda_{Am} = \beta . u'_k . (I_k - x_k) \quad (12)$$

The equation (12) gives an adaptation rule which updates a parameter λ_{Am} expressed as inner dynamics of the *Amygdala* module. As an adaptation coefficient, parameter β is an empirically set value between [0,1].

As a simplification, the amygdala resembles to an engine that generates reward and penalty signals by reinforcement learning on the association between stimuli and motor systems. Emotional responses released by amygdala using this reinforcement influence, are composed by long-term memory stored in the weights [10].

IV. SIMULATION RESULTS AND PERFORMANCE TESTS

In the run-time, biologically inspired cognitive architecture of the humanoid robot requires a self-adaptation procedure for the environmental uncertainties such as difficulties in the detection of behavioral discrimination while the architecture performs its cognitive abilities. During the human robot interaction game play designed as a test bed for the biologically inspired cognitive architecture, some cortical zones such as *Sensory Cortex, Orbitofrontal Cortex, Amygdala* have modified gain or weighting parameters as shown in Fig. 4. Adaptation signals tune the internal states of system characteristics.

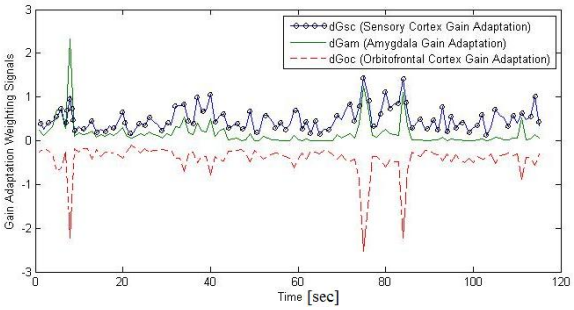


Figure 4. Gain adaptation of cortical areas

Also general system error of the computational architecture is observed by the error function, Err(t). The error term is expressed as a distance metric between ideal (accepted as true) outputs and system outputs retrieved by the architecture.

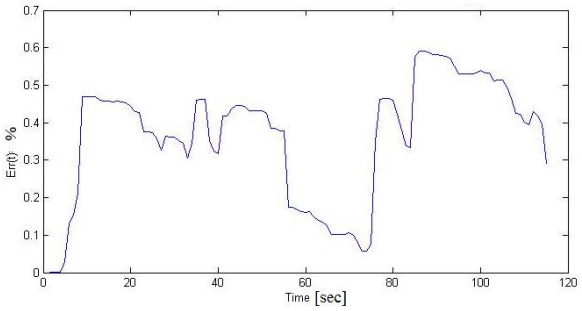


Figure 5. System error

Figure 5. shows the system error term, which is computed by Err(t)=y(x(t))-y(x(t-1)), normalized and presented in percentage. The system outputs are sent to humanoid robot as joint angles. In this project, NAO H25 humanoid robot is used. Due to the upper torso working, there are eight angles used.

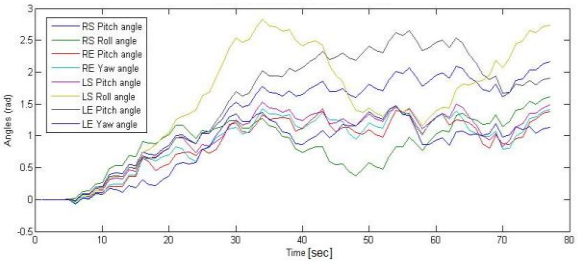


Figure 6. System outputs to robot's joints

Figure 6 explains the response of the humanoid robot to the sign language word "the car". In this figure, R, L represent right and left arm. Also S, E represent shoulder and elbow. For example, elbow of right arm is expressed as RE. Angles of joints are Roll, Pitch, Yaw. According to this, when the card with the car picture is shown to the robot, sign language action designed to represent the sign language word was executed in the simulation.

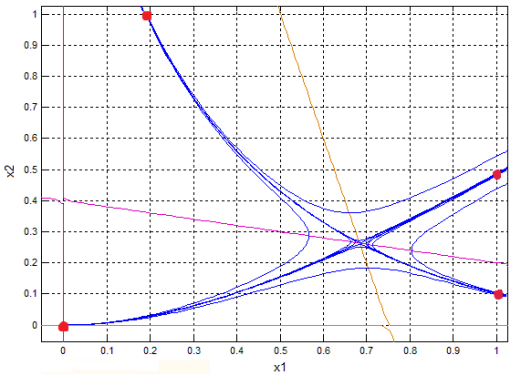


Figure 7. Decision boundaries and attraction basin points for behavior selection

Figure 7 depicts a behavior switching plane with decision boundaries. For a simplified visual representation, the figure illustrates two behaviors, which are competing with each other. The recognized behaviors release with likelihood values between [0,1]. In case of two behaviors switching, x_1 represents the behavior corresponding to the sign language word "car". The second behavior represents the sign language word "table", which is expressed by x_2. Two decision boundary curves can divide behavior switching plane into four regions. For example, if the likelihood values related to behaviors are given such as x_1=0.2, x_2=0.4, it converges to a point x_1=0.2, x_2=1. Thus, the x_2 action (sign language word "the car") is selected. If the likelihood values related to behaviors are x_1=0.8, x_2=0.1, it converges to a point x_1=1, x_2=0.1. Thus the x1 action (sign language word "the table") is selected. If the likelihood values related to behaviors is x_1=0.3, x_2=0.1, it converges to a point x_1=0, x_2=0. As a consequence, none of the two actions are selected. If the likelihood values related to behaviors is x_1=0.8, x_2=0.7, it converges to a point x_1=1, x_2=0.5 in the fourth region. In this region, all actions are selected. Selection of three or more behaviors requires the behavior switching hyper plane with three or more dimensions.

V. DISCUSSION AND FUTURE CASE STUDIES OF APPLICATION SCENARIOS

The studies introduced in this paper have been realized as part of an on-going research, which aims to develop a computational framework of brain model; and artificial emotion based cognitive architecture to be utilized in the humanoid robots for the social and rehabilitation purposes such as assisting sign language tutoring due to the incompetency of 2-D instructional tools developed for this goal and the lack of sufficient educational material [1].

In the proposed system, it is intended that a child-sized humanoid robot performs and recognizes various elementary signs (currently basic upper torso gestures and words from Sign Languages (SL)) so as to assist teaching these signs to children with communication problems [9]. This will be achieved through interaction games based on non-verbal communication, turn-taking and imitation designed specifically for robot and child to play together. In the first versions of the game, the robot was telling a short story verbally and through the story for some selected words, the robot was able to express words in the SL among a set of chosen words using hand movements, body and face gestures and having comprehended the word, the child was encouraged to give relevant feedback in SL or visually to the robot (using a colored card visualizing the word), according to the context of the game. The proposed game is based on the visual cards, the cards will be shown to the robot to select among several signs from American Sign Language (ASL) and Turkish Sign Language (TSL) and basic upper torso motion (hands side, forward, up etc.) Then, the robot performs the sign and waits for the child to imitate [9]. The imitated action is evaluated using an RGB-D camera (Kinect) and the robot will give a motivating comment, when the action is imitated with success [1].

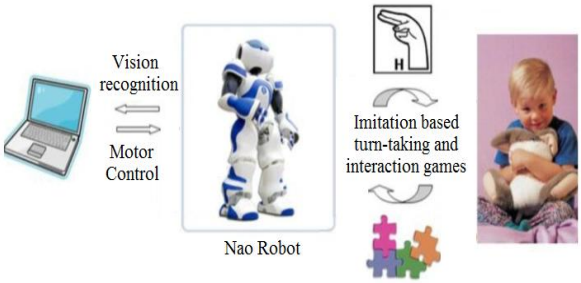


Figure 8. Interaction game play based on SL [1].

The game (Figure 8) is designed for children with special needs in teaching non-verbal communication skills, imitation and turn-taking [9]. The sign language versions of the game were conducted with adults, sign language students, children with normal development, and hearing impaired children with success [1]. We are accepted for a special school for children with special needs for a long term study and will test the game in this school with both autistic and hearing impaired children, shortly. The main aim of this interdisciplinary study is to build a bridge between the technical know-how and robotic hardware with the know-how from different disciplines to produce useful solutions for children with communication problems [1][9].

Another alternative scenario summarizes the attempt to extend this study to autistic children. This paper presents one of the projects, which is produced as an output of this collaboration, and it is planned to use the system and the game in the collaborative special schools on autism. Many such children show interest in robots and find them engaging. Robots can facilitate interaction between the child and teacher. Every child with autism has different needs. Robot behavior needs to be changed to accommodate individual children's needs and as each individual child makes progress.

VI. CONCLUSIONS

In this paper, we designed a brain inspired cognitive architecture embedded into a humanoid robot NAO

H25. As a nonlinear dynamic model, the major novelty in our paper is using Cellular Neural Network to enhance the representation of the input-output topology of the system. Also in this model, an adaptation procedure is applied on some cortical units (e.g. *Sensory Cortex, Orbitofrontal Cortex and Amygdala*). In this way, internal states of these units are converged to some specific points. The adaptation procedure allows the learning of the decision boundary for behavior selection. The interaction game play was simulated and tested in the MATLAB environment so that the system was demonstrated with robot and preschool children from Special School for Hearing Impaired Children. Finally results are observed to check the system performance. Also gain adaptation results on the *sensory cortex, orbitofrontal cortex and amygdala* are observed. In the future, American (ASL) and Turkish (TSL) sign languages will be implemented and tested.

REFERENCES

[1] Ertugrul, B.S., Kivrak H., Daglarli E., Kulaglic A., Tekelioglu A., Kavak S, Ozkul A, Yorgancı R., Kose H., "iSign: Interaction Games for Humanoid Assisted Sign Language Tutoring", International Workshop on Human-Agent Interaction (2012)

[2] Shanahan M, "Consciousness, Emotion, and Imagination A Brain-Inspired Architecture for Cognitive Robotics", In Proc. Artificial Intelligence and Simulation of Behavior, 2005 Workshop: "Next Generation Approaches to Machine Consciousness", pp. 26–35

[3] Shahmirzadi D., "Computational Model of the Brain Limbic Sysyem and its Application in Control Engineering", M.Sc. Thesis, Texas A&M University, 2005

[4] Beheshti Z. and Hashim S.Z.M., "A Review of Emotional Learning And It's Utilization in Control Engineering", Int. J. Advance. Soft Computing Applications, Vol. 2, No. 2, July 2010 ISSN 2074-8523; ICSRS Publication

[5] Yau C-Y, Burn K and Wermter S, "A Neural Wake-Sleep Learning Architecture for Associating Robotic Facial Emotions", Neural Networks, 2008. International Joint Conference on Neural Network, 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pp. 2715 – 2721

[6] Karabacak Ö., Şengör N.S., "A Dynamic Model of a Cognitive Function: Action Selection", 2005, International Federation of Automatic Control.

[7] Denizdurduran B., "Learning How to Select an Action: From Bifurcation Theory to the Brain Inspired Computational Model", 2012, MSc Thesis

[8] Leon O. C. and Roska T., "The CNN Paradigm", 1993, IEEE Transactions on Circuits and Systems-1: Fundamental Theory and Applications, vol 40. No.3, pp 147-157

[9] H. Kivrak, Ertugrul, B.S., R. Yorgancı , E. Daglarli, A. Kulaglic, H. Kose, "Humanoid Assisted Sign Language Tutoring", 5th Workshop on Human-Friendly Robotics (HFR 2012), October, 2012, Brussels, accepted

[10] Mor´en J, Balkenius C, "A Computational Model of Emotional Learning in the Amygdala", 2000, MIT Press

Authors Index

Anzalone, Salvatore 31
Awaad, Iman 1
Baddoura, Ritta 25
Chetouani, Mohamed 31
Daglarli, Evren 37
Grigore, Elena Corina 22
Hertzberg, Joachim 1
Ince, Gökhan 37
Ivaldi, Serena 31
Kawamura, Kazuhiko 15
Kose, Hatice 37
Kraetzschmar, Gerhard 1
Nagai, Yukie 9
Oztop, Erhan 9
Rousseau, Woody 31
Scassellati, Brian 22
Sigaud, Olivier 31
Tan, Huan 15
Ugur, Emre 9
Venture, Gentiane 25
Wilkes, Don 15