Learning for Human-Robot Interaction Modeling

RSS 2010 WORKSHOP
June 27, 2010, Zaragoza, Spain.

http://www.ensta.fr/~tapus/RSS2010/RSS2010-WSHRI.html
ORGANIZERS

Prof. Mohamed Chetouani
ISIR
Pierre and Marie Curie University (UPMC)
Pyramide - T55/65, 4 Place Jussieu, 75005
Paris, France
e-mail: mohamed.chetouani@upmc.fr

Prof. Adriana TAPUS
Cognitive Robotics Lab
ENSTA-ParisTech
32 Blvd Victor, 75015
Paris, France
e-mail: tapus@ensta.fr
<table>
<thead>
<tr>
<th>Time</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:30</td>
<td><strong>Welcome - Introduction</strong></td>
</tr>
<tr>
<td></td>
<td><em>Organizers: Mohamed Chetouani and Adriana Tapus</em></td>
</tr>
<tr>
<td>9:30 - 10:00</td>
<td><strong>Decisional Issues in Human-Robot Interaction</strong></td>
</tr>
<tr>
<td></td>
<td><em>Rachid Alami, LAAS-Toulouse, France</em></td>
</tr>
<tr>
<td>10:00 - 10:30</td>
<td>Learnable adaptivity in situated dialogue for human-robot interaction</td>
</tr>
<tr>
<td></td>
<td><em>Geert-Jan Kruijff, DFKI GmbH, Germany</em></td>
</tr>
<tr>
<td>10:30 - 11:00</td>
<td><strong>Considerations on Gesture in Space and Time</strong></td>
</tr>
<tr>
<td></td>
<td><em>Anna Esposito, Second University of Naples, Italy</em></td>
</tr>
<tr>
<td>11:00 - 11:30</td>
<td><strong>Coffee Break</strong></td>
</tr>
<tr>
<td>11:30 - 12:00</td>
<td><strong>Tactile Guidance for Policy Adaptation</strong></td>
</tr>
<tr>
<td></td>
<td><em>Eric Sauser, EPFL, Switzerland</em></td>
</tr>
<tr>
<td>12:00 - 13:00</td>
<td><strong>Motivation, Emotion, and Learning for Human-Robot Interaction: A</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Spectrum of Research at Georgia Tech’s Mobile Robot Lab</strong></td>
</tr>
<tr>
<td></td>
<td><em>KEYNOTE SPEAKER: Ron Arkin, Georgia Tech, USA</em></td>
</tr>
<tr>
<td>13:00 - 14:30</td>
<td><strong>Lunch Break</strong></td>
</tr>
<tr>
<td>14:30 - 15:00</td>
<td><strong>End-user Programming of a Social Robot Using Dialogues</strong></td>
</tr>
<tr>
<td></td>
<td><em>Miguel Salichs, Carlos III University of Madrid, Spain</em></td>
</tr>
<tr>
<td>15:00 - 16:00</td>
<td><strong>Poster Session</strong></td>
</tr>
<tr>
<td>16:00 - 16:30</td>
<td><strong>Coffee Break</strong></td>
</tr>
<tr>
<td>16:30 - 17:50</td>
<td><strong>Panel of Discussion - Road map of existing robotics learning technologies for social and cognitive robotics</strong></td>
</tr>
<tr>
<td>17:50 - 18:00</td>
<td><strong>Workshop Wrap-up</strong></td>
</tr>
</tbody>
</table>
1. Welcome

Even if human-robot interaction (HRI) is in its early stages of development, the next decade promises systems that will be used in hospitals, schools, homes, and daily lives to assist their users. Therefore, we believe that social cognitive robotics present great potential uses and grand multi-faceted challenges both from the point of view of the user benefit and for the potentialities offered for learning from the robot point of view.

The objective of this workshop is twofold. On the one hand the intention is to gather researchers working in different areas of HRI and social cognitive robotics in order to give a broad panorama of the recent technological advances. On the other hand it will address the learning aspect of these technologies. More precisely, the aims of this workshop include:

• To bring together researcher from interdisciplinary fields, so as to focus on the different aspects of learning all this with the main objective of creating solutions to build robotic systems with high cognitive capabilities, and therefore being novel and complimentary to previous related workshops, which had a broader scope;
• To focus on issues of behavioral models that are designed and naturally adapted according to the user’s profile;
• To provide a general overview of the critical issues and key points in building effective, acceptable and reliable human-robot interaction systems and providing indications for further directions and developments in the field, based on the current diverse expertise of the participants.

The main areas of interest include but are not limited to:

• Machine Learning
• Adaptive Cognitive Systems
• Human-Robot Interaction
• Social learning
2. Invited Speakers

• KEYNOTE SPEAKER:
  Prof. Ron Arkin, Georgia Tech, USA
  
  Title:
  Motivation, Emotion, and Learning for Human-Robot Interaction: A Spectrum of Research at Georgia Tech’s Mobile Robot Lab
  
  Abstract:
  The major portion of the presentation will address ongoing research using TAME as part of a project funded by Samsung Electronics in involving machine learning, affective behavior, and humanoid robots. The use of case-based reasoning and learning momentum (a form of reinforcement learning) are presented in the context of entraining a humanoid robot to its user over time, with the ongoing goal of creating affective human-robot interaction consistent with creating a lifelong partner that can adapt to the varying affective states of its human user over time.

• Dr. Rachid Alami, LAAS-Toulouse, France
  
  Title:
  Decisional Issues in Human-Robot Interaction

• Dr. Geert-Jan Kruijff, DFKI GmbH, Germany
  
  Title:
  Learnable adaptivity in situated dialogue for human-robot interaction
  
  Abstract:
  People don't talk the same way. Person by person, situation by situation, they typically express content in different ways. This requires adaptation on the side of the robot, if it is to deal smoothly with spoken dialogue in human-robot interaction. It needs to adapt how to understand something, what (communicative) goal to undertake when working together with a human, and how best to realize that goal in a given context. The talk discusses recent efforts in developing online learning methods to achieve this adaptivity, in a range of scenarios set in lab- and real world-environments.

• Prof. Anna Esposito, Second University of Naples, Italy
  
  Title:
  Considerations on Gesture in Space and Time
  
  Abstract:
  This talk will discuss two different theories on the significance of gesture in the body to body communication. Are they acting as a support or do they share with speech similar
semantic and pragmatic functions, rejecting the hypothesis that either gesture or speech alone might have a primary role in the communicative act? Both the above theories find some support in experimental data leaving open the question of how to integrate and evaluate these various approaches. The results we are going to present may be relevant in evaluating the relative merits of the above theories. In previous work (Esposito et al., 2001, 2002, 2003; Esposito & Marinaro, 2007) we adopted Kendon (2004) and McNeill’s focus (McNeill 2005) on speech that is synchronized with gesture (in particular hand gesture). The theoretical framework that emerges from this focus holds that, while speech clearly contributes semantic content to the message, gestures engage the visible part (visible actions in Kendon 2005) or the imagistic content (McNeill 1992, 2005). Starting with these assumptions, we tried to link gesture and gesture entities (such as holds) to speech and in particular to speech pauses and we investigated to what degree holds and speech pauses synchronize. A careful review of speech and gesture in narrative discourse data showed that in fluent speech contexts, holds appear to be distributed similarly to speech pauses and to overlap with them and that this phenomenon is observed not only in adults but also in 9 year old children with a rich linguistic vocabulary and advanced language skills (Esposito et al., 2001, 2002, 2003; Esposito & Marinaro, 2007). The findings were also independent from the language (results were similar both for Italian and American English) and were evident in discourse elicited by different tasks, suggesting that at least some aspects of speech and gesture reflect a unified planning process, that appear to be the same for all human beings. Considerations are made on how this research contributes to the embodiment theories and on its use in robotic applications and assistive technologies.

- Dr. Eric Sauser, EPFL, Switzerland

Title:
Tactile Guidance for Policy Adaptation

Abstract:
Demonstration learning is a powerful and practical technique to develop robot behaviors. Even so, development remains a challenge and possible demonstration limitations, for example correspondence issues between the robot and demonstrator, can degrade policy performance. This work presents an approach for policy improvement through a tactile interface located on the body of the robot. We introduce the Tactile Policy Correction (TPC) algorithm that employs tactile feedback for the adaptation of a policy learned from demonstration. The TPC algorithm is validated on humanoid robot performing a grasp positioning task. The performance of the demonstrated policy is found to improve with tactile corrections. Tactile guidance also is shown to enable the development of policies able to successfully execute novel, undemonstrated, tasks. We further show that different modalities, namely teleoperation and tactile control, provide information about allowable variability in the target behavior in different areas of the state space.
Prof. Miguel Salichs, Carlos III University of Madrid, Spain

Title:
End-user Programming of a Social Robot Using Dialogues

Abstract:
Most of current robots must be programmed by experts with a significant knowledge of robotics and programming techniques. Future social robots will be used by common people, which in most cases have no programming skills. It should be desirable that end-users of social robots could program them in a simple and natural way. We present a system, based on mixed-initiative dialogues, that allows end-users to teach activities to a social robot. The program can include not only simple sequences of actions and switching conditions, but also alternative conditions and concurrent sequences of actions. The dialog management system is based on VoiceXML, with an extension to permit multimodal dialogues. The system has been implemented in Maggie, a social robot developed by the UC3M Robotics Lab.
### 3. Posters

<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pet robots with social skills for satisfactory interaction with hospitalized children</td>
<td>Marta Diaz, Joan Saez, Diego Pardo and Cecilio Angulo</td>
</tr>
<tr>
<td>2</td>
<td>Postural expressions of emotions in a humanoid robot for assistive applications</td>
<td>Mustafa Suphi Erden and Adriana Tapus</td>
</tr>
<tr>
<td>3</td>
<td>Learning ability models for human-robot collaboration</td>
<td>Alexandra Kirsch and Fan Cheng</td>
</tr>
<tr>
<td>4</td>
<td>Learning cross-modal translatability: grounding speech act on visual perception</td>
<td>Mario Gianni, Geert-Jan M. Kruijff, and Fiora Pirri</td>
</tr>
<tr>
<td>5</td>
<td>Robot self-initiative and personalization by learning through repeated interactions</td>
<td>Martin Mason and Manuel Lopes</td>
</tr>
<tr>
<td>6</td>
<td>A basic cognitive system for interactive learning of simple visual concepts</td>
<td>Danijel Skocaj, Matej Kristan, Ales Leonardis, Alen Vrecko, Miroslav Janicek, Geert-Jan M. Kruijff, Pierre Lison, Michael Zillich</td>
</tr>
<tr>
<td>7</td>
<td>Learning to Perceive Human Intention and Assist in A Situated Context</td>
<td>Hee-Tae Jung, Shichao Ou and Roderic Grupen</td>
</tr>
<tr>
<td>8</td>
<td>Single Trial Recognition of Error-Related Potentials During Observations of Robot Operation</td>
<td>Inaki Iturrate, Luis Montesano, Javier Minguez</td>
</tr>
<tr>
<td>9</td>
<td>Cooperative Robot Movements for Guiding and Regrouping People using Cost Function Evaluation</td>
<td>Anais Garrell and Alberto Sanfeliu</td>
</tr>
<tr>
<td>10</td>
<td>Interactive Artificial Learning in Multi-Agent Systems</td>
<td>Yomna M. Hassan, Salman Ahmed, and Jacob W. Crandall</td>
</tr>
<tr>
<td>11</td>
<td>Unsupervised learning for interactive behavior for HRI</td>
<td>Yasser F. O. Mohammad and Toyoaki Nishida</td>
</tr>
</tbody>
</table>
Unsupervised Learning of Interactive Behavior for HRI

Yasser F. O. Mohammad
Faculty of Engineering
Assiut University
Egypt, 71516
Email: yasserm@aun.edu.eg

Toyoaki Nishida
Graduate School of Informatics
Kyoto University
Japan, 606–8501
Email: nishida@i.kyoto-u.ac.jp

Abstract—In this paper, we present our efforts toward building interactive robots that can learn how to interact naturally with human partners in different environments and contexts. The main feature of our approach is that it relies completely on unsupervised learning and time series analysis techniques that allow the robot to build its own interaction protocol representation from the bottom up. The final controller of the robot learned this way is a hierarchy of either dynamical systems or probabilistic networks with complexity that is automatically adjusted to the interaction protocol to be learned. We report two examples of applying this technique to learn an explicit interaction protocol in a master-slave settings (guided navigation) and an implicit protocol in a teammate settings (a listener robot).

I. INTRODUCTION

The body of research targeting achieving natural interactive behavior in HRI is increasing every year but most of this research utilizes learning late in the design process. Usually, the designer starts by deciding the required interaction protocol based on studies of human-human interaction or Wizard of Oz experiments and implements this protocol into the robot. Learning may be used to improve the original protocol by adjusting its parameters but the structure of the protocol is usually pre-defined by the designer. One successful example of this approach is the listener robot developed and evaluated by Kanade et al. [1]. The main disadvantage of this approach is the hard-coding of the interaction protocol which reduces the ability of the robot to learn and adapt. It is also time consuming because of the careful analysis required to decide the interaction protocol in enough details to get the robot operational.

Our approach tries to utilize machine learning earlier within the design loop. The main aim of this approach is to allow the robot to learn the interaction protocol directly from samples of the interaction it watches and then to improve its learned protocol continuously during engagement in HRI. The following sections provide a brief presentation of this approach and two real world robot controllers that used it.

II. ARCHITECTURE AND ALGORITHMS

The architecture utilized in this system (Fig. 1) consists of layers of interaction control processes (ICPs) that implement the interaction protocol at multiple levels of abstraction. Each ICP has a forward and reverse components (FICPs and RICPs). The forward component is responsible of generating some kind of behavior (e.g. mutual gaze) while the reverse component is responsible of detecting this behavior in partner’s actions. Each role in the interaction protocol (e.g. listener, instructor, operator, etc) is represented by a set of these ICPs. The architecture assumes operationally that the partner has the same layers and processes (Theory of Simulation). The effect of each ICP on the final behavior is controlled by its activation level. The processes of the first layer are called Basic Interactive Acts (BIAs) and are connected directly to sensors and actuators. Other ICPs take as input the activation level of lower layer ICPs passed through a set of delay elements and the ICPs of other roles in the same layer, and their output is one of the factors controlling the activation level of lower level ICPs. This generates a set of control loops within the architecture that can generate complex behavior from simple BIAs. ICPs can be implemented using either Baysian Networks or Dynamical Systems (e.g. Recurrent Neural Networks).

Learning is done in three stages. In the first stage the BIAs are learned using constrained motif discovery [2]. This is called the interaction bubbling stage. In the second stages the rest of the ICPs are learned using a novel algorithm called Interaction Structure Learning [3]. By the end of this stage, the robot becomes able to interact using the learned protocol and the Interactive Adaptation Algorithm (IAA) [3] is then used.
to continuously adapt the parameters of the learned hierarchy during the interaction.

The main point to emphasize here is that this three stage learning approach starts with no BIAs or ICPs defined and with no knowledge even about their numbers, and ends up with the complete hierarchy given a set of training examples. The number of sessions needed depends on the complexity of the behavior (the number of ICPs involved).

III. LEARNING TO LISTEN

The first example of applying the proposed approach is learning how to use human-like nonverbal behavior while listening to an explanation by a human instructor. The protocol in this case is implicit (no a-priori known components), and complex but with very simple BIAs (e.g., look at the partner, look at most recently looked at object, etc). Instructor's and listener's behaviors were captured using motion capture markers attached to their heads and hands. The ICPs were represented by Radial Basis Function Neural Networks. The most difficult stage in this scenario is the second stage (ISL).

The training data consisted of 22 sessions (around 5:30 minutes each) of human-human interactions in which one subject instructed the other about how to assemble a simple device. The evaluation experiment involved the robot listening to a human instructor explaining the operation of a different device in a different arrangement (e.g., a table between the two partners). Fig. 2(a) shows example frames from the evaluation experiment. The system learned 10 BIAs (including Look@Partner, Look@Salient, Align2Partner, nod). The system also learned three other stages of ICPs converging to a single ICP representing nonverbal listening. The learned controller was compared using third-party subjective evaluation to a gaze controller that used the traditional approach with its parameters adjusted using GA. The proposed controller outperformed the traditional controller in terms of naturalness (p=0.0019), human-likeness (p=0.0099), and comfort of the instructor (p=0.046). The proposed robot also outperformed the traditional controller in the overall subjective evaluation as shown in Fig. 2(b) (p=0.0243). It also performed higher than the expectation of the subjects (p=0.00183).

IV. LEARNING GUIDED NAVIGATION

The second example involved learning an explicit simple protocol that uses complex BIAs. The scenario in this case is guided navigation in which one partner (operator) is guiding the other (actor) using free hand gestures to follow a predetermined path known only to the operator. Gestures were captured using accelerometers attached to both hands of the operator, while actor's behavior was captured using motion capture markers. ICPs were represented using Baysian Networks. The most difficult stage in this scenario is the first stage (Interaction Bubbling).

The training data consisted of a single guided navigation session of 10 to 20 minutes in which a human operator guided a WOZ operated robot to follow a predefined path (Fig. 3(a)). The proposed approach was used to train the robot and then the robot interacted as an actor with the same operator to follow a different path. Performance was compared to a WOZ operated robot using first-person subjective evaluation. Fig. 3(b) shows the scores assigned to each condition by the subjects. It is clear that the proposed controller was able to perform at the same level as a WOZ operated robot in a single training session.

To test the ability of the robot to adapt its learned protocol, another controller was trained using all training session except the one from the current subject. This robot failed in the first four sessions to complete the task but by the final 8 sessions its performance was at the same level as the original controller and the WOZ operator.

V. CONCLUSION

This paper presented an unsupervised technique for learning interactive behavior in three stages. The proposed techniques requires no a-priori information about the number of control layers or processes involved. Two evaluation experiments were also briefly reported showing the effectiveness of the proposed approach in learning both complex protocols of simple behaviors and simple protocols of complex behaviors.

In the future, the proposed approach will be used in more complex situations involving verbal behavior. Long term adaptation using IAA will also be studied in more details.

REFERENCES

Interactive Artificial Learning in Multi-agent Systems

Yomna M. Hassan, Salman Ahmed, and Jacob W. Crandall

Many real-world problems, in which intelligent machines can be useful, including transportation systems, futuristic power grids, and assistive robotics, require interactions between multiple intelligent agents. These intelligent agents often have different goals and allegiances, and hence their behavior is not controlled by a single entity. To be successful, each agent in these systems must learn to adapt to the dynamic and unknown behavior of the other agents to effectively compete and collaborate. However, the development of such agents is challenging, expensive, and time consuming, as encoding these machines with successful behavior requires both domain and technology expertise. Such expertise prohibits these technologies from being customized effectively to the needs of end users.

Two methods are typically cited to overcome these challenges. First, easy-to-use programming interfaces can be developed that require less technology expertise. For example, commercially available robot platforms come equipped with software that allows non-expert users to program robot behaviors [1, 2]. While such software helps users of these systems to create customized autonomous behaviors, the complex nature of the world makes it difficult for users to know what behaviors to create.

A second method to allow novice end users to encode intelligent behavior in dynamic and unknown multi-agent systems is machine learning. Successful artificial algorithms allow an agent to automatically adapt in unknown environments, thus theoretically eliminating the need for the programmer to have domain expertise. Unfortunately, deriving effective features, distance metrics, and other representations used by these algorithms often requires significant domain expertise.

To overcome these challenges and limitations, we are evaluating the potential of a third approach, called interactive artificial learning (IAL). In IAL, a non-expert user interacts with a machine learning algorithm as the agent seeks to learn successful behavior. Ideally, by combining human-machine interactions with artificial learning, the end user and agent can jointly determine and create autonomous behaviors that satisfy the user’s desired needs and goals.

IAL can take on many forms, many of which are illustrated by the five-step machine learning process depicted in Figure 1. In the first step of the process, the algorithm’s internal representations, features, and parameters are configured. In steps two through five, the algorithm repeatedly plans its behavior, performs actions, observes consequences, and updates its internal representations based on its experiences. In IAL, the human can potentially interact with the human in any step of the process [3].

To date, most IAL algorithms focus almost exclusively on interactions in the third and fourth steps of the learning process [4, 5]. This approach has often been successful when the end user knows the desired autonomous behavior and the domain is well-understood. However, since novice end users are unlikely to know how an agent should behave in dynamic and unknown multi-agent systems, we believe that end users must be more fully involved in the learning process. As such, our current research focuses on developing IAL algorithms and interface technologies that allow novice end users to intuitively collaborate with learning agents throughout the learning process. Our in-progress research focuses on the following application areas: smart power grids, dynamic task scheduling in transportation systems, and assistive robotics for autism therapy.

Smart Power Grids

When a power grid is powered in large part by renewable energy sources, the amount of affordable clean energy at any given time will fluctuate depending on the supply generation of renewable energy sources.

---

1 Y. M. Hassan, S. Ahmed, and J. W. Crandall are in the Computing and Information Science Program at the Masdar Institute of Science and Technology, Abu Dhabi, UAE. Email: {yhassan, sahmed, jcrandall}@masdar.ac.ae
Under such situations, demand response programs, in which users are encouraged to reduce or shift their electricity consumption, can be used to match supply and demand. One way to facilitate effective demand response is by providing real-time information and incentives that encourages consumers to alter their consumption patterns in desirable’s ways.

However, such programs are only likely to be useful if autonomous devices can be employed by users to respond to different pricing signals in real time. This creates a multi-agent system in which autonomous devices must learn and adapt to each other. While current multi-agent learning algorithms are incapable of performing effectively in such large-scale systems, we believe that IAL could potentially be effective. This belief is somewhat validated through learning by demonstration (LbD) [4] in repeated stochastic games, wherein LbD algorithms are shown to learn successfully in many instances, even when human input is sub-optimal [6].

One method we are considering in this domain is interactive evolutionary learning, wherein human input is provided to a genetic algorithm. Previous work on interactive evolutionary learning in distributed, collaborative tasks has shown that human input can allow genetic algorithms to learn more effectively [7]. However, such successes require heavy user interactions [7]. Previous work in interactive evolutionary learning in single-agent systems has analyzed methods for decreasing the amount of necessary human interaction in interactive evolutionary learning [8, 9]. Our goal is to create interactive evolutionary algorithms that learn successfully in multi-agent systems with minimal human input.

**Task Scheduling in Multi-Vehicle Transportation Systems**

A second area in which we are applying IAL is dynamic task scheduling in personal rapid transit and taxi-routing systems [10]. In these systems, the routes and destinations of individual vehicles must be determined based on the behavior of the other vehicles in the system. We are considering the situation in which a user must encode successful autonomous behavior in a subset of the vehicles in the system. We are initially investigating how learning by demonstration (LbD), wherein an agent’s behavior is derived by observing user demonstrations, can be used to derive this autonomous behavior. While effective in many domains, current LbD algorithms are highly dependent on pre-defined state features and distance metrics. Our research includes investigating how such information can be derived from user input so that system designers are not required to have as much domain expertise.

**Assistive Robotics for Autism Therapy**

We are also considering IAL in the task of programming robots for autism therapy. Research suggests that autistic children show an affinity for robots [11]. Thus, if used correctly by therapists and caretakers, these robots can potentially be used therapeutically if robot behaviors can be easily programmed and adapted by individual without technology expertise [12]. Given that therapists and caretakers are typically not technology experts, and that each case of autism is distinct, effective IAL in this dynamic and unknown multi-agent system could be critical to the development of these capabilities. We are currently looking into developing IAL algorithms and techniques that will allow people to, quickly and effectively, program robots to express specific emotions.
References

Cooperative Robot Movements for Guiding and Regrouping People using Cost Function Evaluation

Anaís Garrell and Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
08028 Barcelona, Spain

Abstract—The objective of this research is to optimize robots’ cooperative work and obtain the minimum displacement of humans in a guiding people mission, where some individuals can escape from the formation and must be regrouped by robots. The guiding mission is done in an urban environment, with obstacles and building constraints, where people can move freely and using multiple mobile robots which work cooperatively. In this paper, the forces that actuate toward robots will be studied, which are the forces between robots and humans and the forces between humans. Our goal is to find out the minimum work required for robots to lead and regroup people. We have developed a cost function that minimizes the work required in order to do this cooperative task.

I. INTRODUCTION

The interaction between social robotics and cooperative robotics areas is a new field of study. Therefore, the number of publications that exist nowadays is quite short, specifically, if we refer to the study of guiding a group of people in urban areas with several robots. We can find some works using a single robot leading people in exhibitions and museums [2], or in hospitals [3]. In previous work [4], a model for guiding people in a dynamic environment using several robots working in a cooperative way was presented. This model is called “Discrete Time Motion” (DTM), which is used to represent people and robot motions.

In this research we present a method to optimize locally the tasks assignment to robots for doing their missions. Robots’ assignation are done by analyzing the minimum work required to do such task, where the function to minimize is based on one hand, by robot’s motion, and, on the other hand, by the impact of such motions on people’s displacement.

II. OPTIMAL ROBOT TASK ASSIGNMENT FOR COOPERATIVE MISSION

The cost function, described below, speaks in Work terms, and it can be divided as: (i) Robot work motion, and (ii) Human work motion.

In order to know what robots’ tasks are, we have considered the following situations: (i) The leader robot has to guide people, (ii) one robot has to look for the person (or people) that can potentially escape from the crowd formation and push him (or them) to regroup him (or them) into group, and, (iii) one robot has to go behind the people in order to push them in case that the crowd formation is broken down.

The Robot tasks that we are considering are, Leader task which computes a path planning and moves to the next point, where exists a drag force that will attract people behind the robot. Looking for a person that goes away task: The robot moves to the estimated position of the individual who goes away from the crowd formation, the estimation is computed using a Particle Filter [1]. Pushing task: The robot pushes a person that has gone away in order to reach the crowd formation. Crowd traversing task: The robot has to move through the formation to achieve the estimated position of the person that goes away from the crowd formation. In this work we are not taken into account this situation, due to safety reasons. We use the equations defined in previous works on human behavior with other individuals [6]. Working with autonomous mobile robots, the robot \( i \) work motion is expressed by:

\[
f_{m}^i = m_ia_i
\]

\[
W_{m}^i = \int f_{m}^i \Delta x_i
\]

where \( m_i \) is the mass of the \( i \)-th robot, \( a_i \) its acceleration and \( \Delta x_i \) the space traversed by the robot to achieve its goal.

The effect of robots on people as forces is: leader robot: attractive (dragging) force, it is inversely proportional to the distance, until a certain distance, and, shepherding robot: Repulsive (pushing,traversing) force, has a repulsive effect inside people’s living space.

The dragging force is necessary when the leader robot guides the group of people from one place to another. It acts as an attractive force, the force applied by robot leader \( i \) to each person \( j \) and the dragging work are defined by:
\[ f_{ij}^p(t) = -C_{ij} n_{ij}(t) = -C_{ij} \frac{x_i(t) - x_j(t)}{d_{ij}(t)} \]  
\[ d_{ij}(t) = \|x_i(t) - x_j(t)\| \]  
\[ W_i = \sum f_{ij}^p \Delta s_j, \forall \text{ person } j \]

where \( d_{ij}(t) \) is the normalized vector pointing from person \( j \) to robot \( i \) at instant \( t \). See [5] for more information about the parameter \( C_{ij} \), which reflects the attraction coefficient over the individual \( j \), and it depends on the distance between the robot leader and person \( j \). Where \( \Delta s_j \) is the distance traveled by the person \( j \).

The **Pushing force** is given by the repulsive effect developed by shepherding robot on the group of people, for regrouping a person (or the broken crowd) in the main crowd formation. The territorial effect may be described as a repulsive social force, and the work can be computed as:

\[ f_{ij}^p = A_i \exp^{(r_{ij} - d_{ij})/B_i} n_{ij} \left( \lambda_i + (1 + \lambda_i) \frac{1 + \cos(\Phi_{ij})}{2} \right) \]  
\[ W_p = \sum f_{ij}^p \Delta s_j, \forall \text{ person in } \Omega_i \]

Where \( A_i \) is the interaction strength, \( r_{ij} = r_j + r_i \) the sum of the radii of robot \( i \) and person \( j \), usually people has radii of one meter, and robots 1.5 m, \( B_i \) parameter of repulsive interaction, \( d_{ij}(t) = \|x_i(t) - x_j(t)\| \) is the distance of the mass center of robot \( i \) and person \( j \). Finally, with the choice \( \lambda < 1 \), the parameter reflects the situation in front of a pedestrian has a larger impact on his behavior than things happening behind. The angle \( \Phi_{ij} \) denotes the angle between the direction \( \vec{e}_{ij}(t) \) of motion and the direction \( -n_{ij}(t) \) of the object exerting the repulsive force. See [5]. Where \( \Omega_i \) is the set of people in which one of the helper robots have reached the living space.

And last but not least, the **Traversing force** is determined by the forces applied by the robot when is traversing the crowd. As we have said before, in this research we consider this force infinity. The cost function for robot \( i \), given a specific task, is the following one:

\[ W_i = \delta_m W_{ip}^m + \delta_d W_{ip}^d + \delta_p W_{ip}^p + \delta_t W_{ip}^t \]

Where \( k \) could be pushing, dragging, traversing or motion and \( \delta_k \) is 1 if this task \( k \) is assigned, and 0 if this task is not assigned. Finally, the task assignment for the robots will be the one which locally minimizes the minimum assigned work cost, which is the one required to do the global task. It is computed by the following way:

\[ C = \text{argmin} \{W_{\text{total}}(c)\}, \forall \text{ configuration } c \]

where the **Configurations** mean how the tasks are distributed among the robots, for each configuration \( c \) robots compute \( W_{\text{total}} \) which is the addition of all \( W_i \) for all robots \( i \) that are working cooperatively.

### III. Experiments and Results

The results we will expose correspond to a synthetic experiment. We have considered one scenarios that robots can find in the North Campus of UPC, with open areas and cross areas. In this experiment, the dynamical models of the persons, we have considered a group of 9 persons, will follow the models described by Helbing et al. [6].

In Fig. 2(a) we present the evolution of the cost function computed using different robots behaviors, it can be seen that the behavior that obtains the lower cost is the one which follows the optimization of the cost function presented previously. In Fig. 2(b) the trajectory the group has followed is presented. Hence, the cost function minimizes globally the work of the group of robots along all the mission.

### IV. Conclusions

We have presented a new cost function for optimizing cooperative robot movements for guiding and regrouping people in a guiding missions. In contrast to existing approaches, our method can tackle more realistic situations, such as dealing with large environments with obstacles, or regrouping people who left the group. For that reason, this work can be applied in some real robots applications, for instance, guiding people in emergency areas, or acting as a robot companion. In future work, we are going to study the relation between the number of robots required and the number of people who are guided.

### REFERENCES


Single Trial Recognition of Error-Related Potentials During Observation of Robot Operation

Iñaki Iturrate, Luis Montesano, Javier Minguez

I. INTRODUCTION

Event-Related Potentials (ERP) are signals that are elicited by the presence of an internal or external event [1], usually recorded by means of an electroencephalogram (EEG). In cognitive neuroscience, it is well known the usage of the ERP to study the underlying mechanisms of human error processing, sometimes referred to Error-related Potentials (ErrPs) [2]. This is because the observation/execution of an incorrect action for the user triggers an activity or potential. This potential codifies the difference between the user’s expected outcome and the actual one. Different ErrPs have been described, for instance, when a subject performs a choice reaction task under time pressure and realizes that he/she has committed an error [3] (response ErrPs); when the subject perceives an error committed by another person (observation ErrPs) [4]; when the subject delivers an order and the machine executes another one [2] (interaction ErrP); and recently when the subject perceives an error committed by a simulated robot [5].

Recent studies have shown that it is possible to use these error potentials in a Brain-Computer Interface context. In [2], the authors demonstrated the feasibility of detecting these potentials online, and proposed their use to recover from BCI errors when operating a wheelchair controlled by asynchronous EEG activity. In a similar way, in [6], the authors proposed the use of single-trial detection of error potentials to detect misinterpreted commands in a P300-based speller. Finally, in [5], the authors proposed the detection and use of these signals as a reward for a simple Reinforcement Learning task. However, the question is whether this framework is potentially usable in a real robotic context (e.g. a rehabilitation prosthesis), that is, whether the error mechanisms of the brain are also elicited by observing a real robot operation, and the feasibility of detecting these signals in real-time. The benefits of this framework would be the possibility to detect online an incorrect operation of the robotic device and correct its behavior, and opens the door to develop prostheses that could learn users’ desires and adapt to time-specific requirements.

To study this question, this paper presents an experiment developed with a real robot. Experimental results have been obtained with 4 participants observing the operation of a 5 dof robotic arm performing correct/incorrect reaching tasks, while an EEG system recorded their brain activity. The results suggest that: (a) the brain areas that play a role in detection and monitoring of errors also play a role when observing the operation of a real robot; (b) a brain discriminative response is elicited during the observation of a correct/incorrect operation of a real robot, (c) this response is consistent among different subjects, (d) it is possible to learn a classifier that provides online categorization with high accuracy (~ 80%).

II. METHODOLOGY AND NEUROPHYSIOLOGICAL RESPONSE

We designed an experiment to determine if a specific brain potential is elicited during the observation correct/incorrect operation of a real robot, if it is consistent among different subjects, and if it is possible to learn a classifier that provides online categorization with enough accuracy.

The general setting of the experiment was a user observing the operation of a robot arm (a Katana300 with 5 degrees of freedom) while the EEG was recorded (Figure 1a). The robot continuously operated by developing reaching tasks to five predefined positions (Figure 1b), where a motion towards the center (colored in red) was a correct operation and a motion towards the locations placed on the side (left or right) were small (yellow) and large (red) errors.

Four male, right-handed, 24-aged people participated in the experiments. For each participant, 100 ERP brain responses of each action were recorded.

After recording the EEG data, we characterized the brain response as a possible ERP. The time-locked averaged ERP potentials for each participant were constructed. This averaged ERP was then averaged for the four participants for error versus correct responses. The results shown on Figure 2 demonstrate that the ERPs resulting from the robot correct/incorrect operations are different, which implies that on average, there are different brain processes involved.

Additionally, a Source Localization technique to analyze the active areas was performed. The main active areas of the conform a brain activity on the frontier of the Pre-Supplementary Motor Area (Pre-SMA, BA 6 and 4), Anterior...
Cingulate Cortex (ACC, BA 24) and Posterior Cingulate Cortex (PCC, BA 31 and 5). As discussed on [7], the results suggest the existence of an activity related with error detection and processing (due to ACC and Pre-SMA). The error-related areas also agree with several results that obtained the same areas in the most prominent negativity in reaction, observation and interaction errors [7], [3], [4]. These results support the hypothesis that a discriminative (correct/incorrect) Event-Related Potential is elicited during the human monitoring of the robot operation.

III. CALIBRATION AND REAL-TIME CLASSIFICATION

The classification process is composed of two different phases: feature extraction and classification. The used features were the RAW data subsampled to 64 Hz. In order to select which features use (i.e., which channels and time window), a statistical measure that shows the areas significantly different between error and correct responses was performed. Concretely, we computed the $r^2$ analysis [8], widely used in neurophysiology for this purpose. Figure 3 shows the $r^2$ for error versus correct responses, for every channel within the time window [0 – 800] ms, averaged for all the participants. The activity on the ERP is clearly centered on the FCz and Cz electrodes, suggesting a fronto-central activity and thus agreeing with the analysis performed on the previous section. We selected (by visual inspection) for classification the following fronto-central channels: Fz, FC1, FC2, Cz, CP1, CP2 and CPz. The time window was also fixed by visual inspection according to the $r^2$ results (see footnote 1), selecting the range [200 – 800] ms.

The previous features were normalized on the range [0 – 1], and they were used to train a Support Vector Machine (SVM). We used the $\nu$-SVM classifier with a radial basis function kernel.2

We selected 25% of each type of incorrect movement so as to have balanced data, thus having 100 error examples and 100 correct examples. Finally, in order to minimize overfitting effects, we used a ten-fold cross-validation strategy to train the classifier. Furthermore, the normalization values were calculated with the 90% of the data and applied to the remaining 10%.

The classification performances for each participant are shown on table I.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>CLASSIFIER PERFORMANCE: ERROR VS CORRECT RESPONSES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P1</td>
</tr>
<tr>
<td>Error</td>
<td>84.67%</td>
</tr>
<tr>
<td>Correct</td>
<td>90.56%</td>
</tr>
</tbody>
</table>

The results show a high detection rate of the ERPs, being roughly an 80% on average. The accuracies are always higher on the correct responses. In general, these results demonstrate the feasibility of detecting these signals on single trial when elicitated by a robotic arm.

REFERENCES

Learning to Perceive Human Intention and Assist in A Situated Context

Hee-Tae Jung*, Shichao Ou† and Roderic Grupen*

*Department of Computer Science, University of Massachusetts Amherst, Amherst, Massachusetts 01003, USA
Email: {hjung, grupen}@cs.umass.edu

†Network Equipment Technologies, New Jersey 07724, USA
Email: chao.ou@net.com

Abstract—In this paper, we propose that a robot can create a series of monitors describing its own pointing gesture and reaching behavior and that these models can be used to infer human intention delivered by a pointing gesture and assist accordingly. This extends our previous work where we demonstrated that an intrinsically motivated robot can be designed to seek controllable relationships with the world and employ them to solicit assistance from a nearby human via expressive gestures. Preliminary experimental results demonstrate that our approach enables a robot to respond appropriately after learning a receptive quality of gesture.

I. INTRODUCTION

It is envisioned that personal robots will assist human daily activities in various settings [1]. In these scenarios, robots should be able to parse communicative signals and understand the intention delivered by evocative gestures so that even naive users and those with disabilities can receive aid when necessary. Drawing inspiration from the psychology literature [2], Ou and Grupen studied how a pointing gesture can emerge from a reaching behavior in under-actuated contexts and demonstrated that the pointing gesture can successfully solicit human assistance [3] as in Fig. 1a. In this work, we address how this expressive behavior knowledge can be exploited to perceive human intention via a pointing gesture and assist accordingly in the same context associated with the very behavior. This coincides with the recent hypothesis in biology literature such that there may exist a common coding for expressing and perceiving behaviors [4].

In Section II, we first visit the psychology and biology literature that inspired this work. Then, in Section III, we discuss how a robot learns a receptive behavior which perceives human intention and assists in a situated context using the control basis and action schema frameworks. Section IV concludes this paper.

II. BACKGROUND

A. Manual Behavior and Communicative Gestures

Psychologists acknowledge a tight connection between manual behavior and communicative gestures. For instance, Vygotsky noted that a pointing gesture originated from an unsuccessful attempt to grasp certain objects [2]. As infants attempt to reach for out-of-reach objects, even though they inevitably fail, in the presence of a caregiver. The action is recognized and interpreted as the “intention” to acquire the objects. Thus, the action becomes a means of conveying the intention to the caregiver. In this work, this insight is applied to robotics wherein a robot learns to perceive the pointing gesture of a human as an evocative gesture and assists by handing the object to the human.

B. Common Coding for Expressive and Receptive Behaviors

Mirror neurons were first found in the inferior frontal gyrus (region F5) and the inferior parietal lobe of the macaque monkeys [4]. It is observed that mirror neurons discharge both when a monkey performs an action and when it observes others do a similar action. Recently, a similar research result on human subjects have been reported [5]. Specifically, mirror neurons which discharge when a human grasps an object would also discharge when the human perceives another person grasp the object. Inspired in part by this finding, we designed the frameworks and a representation that can support the expressive and receptive symmetry that some models of mirror neurons propose.

III. LEARNING RECEPTIVE BEHAVIORS

A receptive behavior, in this work, consists of recognizing human pointing and assisting human grasping. Throughout expriments, a robot is situated in a naturally cluttered lab environment where a human stands across a table and points at an out-of-reach object as in Fig. 1b. Human detection was done by using [6]. Due to the lack of space, detailed explanation
Subjects were asked to solicit robot assistance by evaluating the performance of the learned behavior schema. After the robot acquired a stable schema, i.e., hands the referenced object to the human using Q-learning within about 20 episodes, in the presence of a human, the robot learned a positive experience for behavior acquisition. Within about 20 episodes, the robot learned how to observe intention and how to respond directly by interacting with humans rather than simply executing a preprogrammed response. The success rate was 78% for 4 objects and 96% for 2 objects respectively. It is important to note that the robot learned how to observe intention and how to respond directly by interacting with humans rather than simply executing a preprogrammed response.

In this paper, we address the problem where a robot learns the intention of a human pointing gesture and assists accordingly. The preliminary experimental results demonstrate that a robot can learn assisting behavior through intrinsically motivated learning. In the future, we plan to conduct further experiments where a robot uses a suite of monitors to recognize diverse human gestures more robustly.

### ACKNOWLEDGMENT

This research was funded in part by a gift from Microsoft Corporation and award N00014-07-1-0749 from Massachusetts Institute of Technology/ONR Muri. The first author gratefully acknowledges Robin Popplestone Fellowship and Cambridge Culture Foundation Scholarship.

### REFERENCES

A basic cognitive system for interactive learning of simple visual concepts

Danijel Skočaj, Matej Kristan, Miroslav Janiček, Geert-Jan M. Kruijff, Danijel Skočaj, Matej Kristan, Miroslav Janiček, Geert-Jan M. Kruijff, University of Ljubljana, Slovenia
Aleš Leonardis, Alen Vrečko, Pierre Lison, DFKI, Saarbrücken, Germany
Michael Zillich, Vienna University of Technology, Austria

Abstract—In this work we present a system and underlying representations and mechanisms for continuous learning of visual concepts in dialogue with a human tutor.

I. INTRODUCTION

Two common and important characteristics of cognitive systems are the ability to learn and the ability to communicate. By combining both competencies such a system could also be capable of interactive learning, i.e., learning in dialogue with a human, which should significantly facilitate the incremental learning processes. In this work we briefly describe the representations and mechanisms that enable such interactive learning and present a system that was designed to acquire visual concepts through interaction with a human [1].

Fig. 1 depicts our robot George engaged in a dialogue with a human tutor1. In this scenario, the main goal is to teach the robot about object properties (colours and two basic shapes) in an interactive way. The tutor can teach the robot about object properties (e.g., 'H: This is a red thing.'), or the robot can try to learn autonomously or ask the tutor for help when necessary (e.g., 'G: Is the elongated thing red?'). Our aim is that the learning process is efficient in terms of learning progress, is not overly taxing with respect to tutor supervision and is performed in a natural, user friendly way.

To enable interactive learning, the system has to be able to, on one hand, perceive the scene and (partially) interpret the visual information and build the corresponding representations of visual objects, and on the other hand, to communicate with the tutor and interpret the tutor's utterances, forming the corresponding representations of the linguistic meaning. The system should then relate these two types of modal representations and on top of them create new, a-modal, representations that enable further communication and allow for incremental updating of visual models, therefore facilitate incremental learning. In the following section we will describe the robot system we have developed, and the underlying representations and mechanisms that implement the premises mentioned above.

II. THE SYSTEM

The implementation of the robot is based on a specific architecture schema, which we call CAS (CoSy Architecture Schema) [2]. The schema is essentially a distributed working memory model, where representations are linked within and across the working memories, and are updated asynchronously and in parallel. The system is therefore composed of several subarchitectures (SAs) implementing different functionalities and communicating through their working memories. The George system is composed of three such subarchitectures: the Binder SA, the Communications SA and the Visual SA, as depicted in Fig. 2.

1The robot can be seen in action in the video accessible at http://cogx.eu/results/george.

Fig. 1. Continuous interactive learning of visual properties.

(a) Scenario setup. (b) Observed scene.

Fig. 2. Architecture of the George system.
The Visual SA processes the visual scene as a whole using stereo pairs of images and identifies regions in the scene that might be interesting for further visual processing (3D spaces of interest or SOIs that stick out of the plane). These regions are further analysed; the potential objects (proto-objects) are segmented using 3D and colour information and are then subjected to feature extraction. The extracted features are then used for learning and recognition of qualitative visual attributes, like colour and shape. These visual concepts are represented as generative models that take the form of probability density functions over the feature space. They are based on online Kernel Density Estimator (oKDE) [3], that we have developed, and are constructed in an online fashion from new observations. The oKDE estimates the probability density functions by a mixture of Gaussians, and is able to adapt from the positive examples (learning) as well as negative examples (unlearning) [4]. Our approach also does not assume the close world assumption; at every step the system also takes into account the probability that it has encountered a concept that has not been observed before. Therefore, during online operation, a multivariate generative model is continually maintained for each of the visual concepts and for mutually exclusive sets of concepts (e.g., all colours, or all shapes) the optimal feature subspace is continually being determined. This feature subspace is then used to construct a Bayesian classifier for a set of mutually exclusive concepts, which is used for recognition of individual object properties.

The recognized visual properties are then forwarded to the Binder SA, which serves as a central hub for gathering information from different modalities about entities currently perceived in the environment. Based on the available information, the binder seeks to fuse the perceptual inputs arising from the various subarchitectures, by checking whether their respective features correlate with each other. The probability of these correlations are encoded in a Bayesian network. We call the resulting (amodal) information structure a belief. The task of the binder is to decide which perceptual inputs belong to the same real-world entity, and should therefore be unified into a belief. The decision algorithm uses a technique from probabilistic data fusion, called the Independent Likelihood Pool (ILP) [5]. A belief encodes also additional information related to the specific situation and perspective in which the belief was formed, such as spatio-temporal frame, epistemic status and a saliency value. The beliefs, being high-level symbolic representations available for the whole cognitive architecture, provide a unified model of the environment which can be efficiently used when interacting with the human user.

Beliefs can also be altered by Communication SA. It analyses an incoming audio signal and parses the created word lattice. From the space of possible linguistic meaning representations for the utterance, the contextually most appropriate one is chosen [6]. We represent this meaning as a logical form, an ontologically richly sorted relational structure. Given this structure, the Communication SA establishes which meaningful parts might be referring to objects in the visual context. The actual reference resolution then takes place when we perform dialogue interpretation. In this process, we use weighted abductive inference to establish the intention behind the utterance.

If the intention was to provide to the system a novel information that can be learned (Tutor driven learning), a belief attributed to the human is constructed from the meaning of the utterance. This event triggers a learning opportunity in Visual SA, where the corresponding visual concepts are updated. The learning process can also be initiated by the system itself (Tutor assisted learning). In the case of missing or ambiguous modal information (interpretation of the current scene), Visual SA can send a clarification request to Communication SA, which formulates a dialogue goal given the information the system needs to know and how that can be related to the current dialogue and belief-context. Dialogue planning turns this goal into a meaning representation that expresses the request in context. This is then subsequently synthesised, typically as a polar or open question about a certain object property, and the tutor’s answer is then used to update the models.

III. Conclusion

In this work we briefly presented the system and underlying representations and mechanisms for continuous learning of visual concepts in dialogue with a human tutor. We have made several contributions at the level of individual components (modelling beliefs, dialogue processing, incremental learning), as well as at the system level; all the components presented in this work have been integrated into a coherent multimodal distributed asynchronous system. Building on this system, our final goal is to produce an autonomous robot that will be able to efficiently learn and adapt to an everchanging world by capturing and processing cross-modal information in an interaction with the environment and other cognitive agents.

ACKNOWLEDGMENT

The work was supported by the EC FP7 IST project CogX-215181, and partially by the Research program Computer Vision P2-0214 (RS).

REFERENCES

Robot Self-Initiative and Personalization by Learning through Repeated Interactions

Martin Mason1,2, 1Mt. San Antonio College, USA, profmason@gmail.com
Manuel Lopes2, 2University of Plymouth, UK, manuel.lopes@plymouth.ac.uk

I. INTRODUCTION

The aim of this project is to create an efficient personal, or service, robot that is able to accept commands from a user, and anticipate the user’s needs. We envision a robot that comes out-of-the-box with a large set of skills to accept commands from the user, and the capability to learn and create profiles to anticipate user needs and pro-actively fulfill them.

In this work we make several contributions: a) provide a system that adapts to the user through repeated interactions; b) a novel system to describe users’ preferences using a feature based description of the world; c) a classifier to identify good and bad world states based on the features d) a system that can pro-actively plan, for a given user’s profile, how to change the world from a bad state to a good state through a sequence of steps e) an integrated system using verbal instructions and a real platform that achieves all these behaviors and is evaluated in terms of user acceptability and efficiency of interaction.

An efficient human-robot interaction requires several components [1], [2]. A natural language interface allows users to issue commands easily and observe the results. The user can direct the robot to perform various actions and as the robot’s capabilities increase, the user’s role can transition from direction to clarification. This concept of semi-autonomy has been already analyzed in terms of the tasks that require different levels of autonomy [3] and it has been proposed that higher levels of autonomy can be achieved through user adaptation. These mechanisms can be used in robots that can implement the user’s desires in a very autonomous way [4], [5], and with collaborative task execution [6].

Allowing the user to train the robot for their own ends [7], will not only provide a more useful tool but also a tool that is easier to understand. As each user will have different definitions of successful goal completion, having the robot achieve such goals according to user preferences should lead to an increased acceptance of the system. Expected end-users will not have the technical skills to program robots for their needs and so the robots must have strategies in-place to allow such adaptation. Our approach relies on having the robot learn interaction histories [2], [8] while being commanded by the user to perform useful tasks, this contrasts with Learning by Demonstration where an explicit training phase exists.

II. SYSTEM SPECIFICATION

Our system is composed of three main components: a) dialog system; b) user profiling and adaptation and c) task planning and execution by self-initiative.

A dialog system is used to receive the users’ commands and clarification responses and to provide cooperative supervision of the task execution [9]. The system is not limited to a fixed sentence structure and allows users to speak naturally to the system. The grammar is defined from the concepts that can be grounded in the robot’s own perceptual and motor skills. The system implements an optional confirmation and clarification request system. Although such confirmations must be ensured in critical systems, for personal robots it is a cause of fatigue and makes the interaction very unnatural. Upon hearing a command the robot executes the most probable inference from the information available. This can be seen as a primitive form of self-initiative where the robot executes the most likely command. After some time such decisions will be aligned with the particular user, in the beginning they just correspond to an average profile we created for the task at hand.

User Profiling and Adaptation is developed by learning a probabilistic model that for each world state gives the probability of being a desired state for each user. This is done using a kernel method based on beta-binomial probability models [10]. We consider a world representation that can generalize between different environments and the number/types of objects. For this we created a large bank of features from which we selected relevant ones taking into account an online survey about the task at hand. We do not restrict ourselves to copy the final state of the environment by memorizing the desired locations of objects, but learn what makes a desired state and what does not.

Self-initiative is only achieved after a certain number of interactions. The model is acquired incrementally by assuming that each user request moves the world from a less- to a more-desired state. With such reasoning, the final state of each interaction episode is a good state and all the intermediate ones are bad states. With such background knowledge, the robot will then self initiate actions based on the user profile. We assume that the robot is able to plan how to reach an object while avoiding obstacles, but for the sake of brevity will not discuss this topic. What is new is the way our system uses the learned user profile to achieve the desired goal and allowing for generalization. First from an initial state we assume that we do not know the desired final state and, second assume that not all desired states are reachable. To allow for generalization, we take an optimization view and use a method that, from an initial condition, finds a realizable path to a state whose features correspond to a desirable state. This contrasts with other approaches that separate the process of selecting the desired state from the problem of planning how to reach there. This approach is more complex than doing these separately but avoids having to re-plan if an unachievable state is generated.

In our experiments a breath-first search method (see [11]) was able to plan tasks with 4 objects in an environment where each object could be at 45 different locations, corresponding to more than 3 million states. Typical plans consisted of a sequence of nine actions.

III. SETUP AND EXPERIMENTS

We consider a robot with a holonomic base and a humanoid torso that is able to grasp, transport and release objects, and navigate and avoid obstacles, see Figure 1. We rely on an overhead camera and on-board sensors.
To validate what image features are relevant in developing user profiles we built an online survey. We consider the task of tidying a room that is cluttered with different objects. The robot can receive commands from the user to clean the room. After some interactions the robot should be able to clean the room according to the particular user preferences.

A ten question online survey asked participants to rate images based on the perceived tidiness of the objects displayed. The results from the twenty six participants were used to calculate a weighted tidiness score for each image. The images were analyzed for a set of thirty six geometric and relational features such as the intra class dispersion between objects and the distance between the center of each object class and each obstacle in the image. The resulting features and weighted scores were used to train the general user profile. From the results we saw that users prefer objects close to corners or borders, and that clustering objects is also important. There were lesser preferences for clustering objects by color and aligning them.

A. Experiment I - Interaction Validation

This experiment will study how well can a user instruct the robot. Table I summarizes the results. Two users were asked to interact with the robot to get it to transition a scene from the same initial state of scattered objects to a tidy state defined as having all the objects in the corner of the scene. A third user was asked to interact with the robot to transition a simplified scene from an untidy to a tidy state. The efficiency of the interaction was measured by the ratio of the successful interactions to the total number of interactions required to complete the task. The robot was deemed to have failed the interaction if it either failed to complete the task, or if the user asked the robot to do something that was outside of its domain. Language failures are defined as failures in the natural language system to parse the user input into usable commands. The exit survey asked users to rate their satisfaction with the behavior of the robot.

<table>
<thead>
<tr>
<th>User Gender</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>Non-Native</td>
<td>Native</td>
</tr>
<tr>
<td>Number of Interactions</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Robot Failures</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Language Failures</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.53</td>
<td>0.82</td>
</tr>
<tr>
<td>Interaction Time (sec)</td>
<td>301</td>
<td>276</td>
</tr>
</tbody>
</table>

To achieve the desired state, the robot needs to move all the four objects.

**TABLE I**

**VALIDATION OF THE HUMAN MACHINE INTERACTION SYSTEM**

The robot is able to pro-actively act in different environments and produce substantially different results that correspond to the users’ preferences, see Figure 2. To achieve the desired state, the robot needs to move all the four objects.

**IV. CONCLUSIONS**

In this work we presented a system that interacts naturally with a user to execute room tidying tasks. The robot is able to learn user profiles (defined as a database of good and bad states) by acquiring knowledge from the different interactions episodes it has with the user. Using such knowledge the robot is able to pro-actively act in different environments and world configurations taking into account the user’s (expected) preferences. We show results with a real platform performing room tidying tasks.

**REFERENCES**


---

1 Videos can be seen at: http://profmason.com/?p=1284
Learning cross-modal translatability: grounding speach act on visual perception

Mario Gianni\textsuperscript{1}, Geert-Jan M. Kruifff\textsuperscript{2}, and Fiora Pirri\textsuperscript{1}

\textsuperscript{1}Dipartimento di Informatica e Sistemistica, Sapienza, Universita’ di Roma
\textsuperscript{2}Language Technology Lab, German Res. Center for Artificial Intell.(DFKI GmbH)

May 28, 2010

The problem of grounding language on visual perception has been nowadays investigated under different approaches, we refer the reader in particular to the works of \cite{7, 11, 3, 13, 2, 10, 12, 6, 5, 1}. It is less investigated the inverse problem, that is, the problem of building the semantics/interpretation of visual perception, via speech-act.

In this abstract we face the two problems simultaneously, via learning both the language and its semantics by human-robot interaction. We describe the progress of a current research facing the problem of simultaneously grounding parts of speech and learning the signature of a language for describing both actions and states space, while actions are executed and shown in a video. Indeed, having both a language and a suitable semantics/interpretation of objects, actions and states properties, we will be able to build descriptions and representations of real world activities under several interaction modalities.

Given two inputs, a video and a narrative, the task is to associate a signature and an interpretation to each significant action and the afforded objects, in the sequence, and to infer the preconditions and effects of the actions so as to interpret the chronicle, explaining the beliefs of the agent about the observed task.

We start, thus, with two sets of observations the set \(\{Y_i\}_{i=1}^N\) of speech-acts and the set \(\{D_k\}_{k=1}^C\) of descriptors of the action and objects space, both suitably extracted from the audio and video sequence (there are several methods to do that, for the visual sequence here we mention \cite{9}). There are two sets of hidden data, namely the speech-act labels \(\{X_i\}_{i=1}^N\), and the properties \(\{P_j\}_{j=1}^H\) induced by actions, specifying how actions dynamically change both what is visible and what can be reported. The hidden variables \(P\) are indexed by time, and the hidden speech-act labels \(X\) are indexed by time and contextual links. We call these indices the states, thus, for all visual states \(j \in S\) there exists a cluster of contextual links \(\{j_1, \ldots, j_k\}\) formed by \(S_k\) specifying a neighbour system for the speech act labels.

The (simplified) dependency relation among the random variables is as follows. Speech-acts are independent of any other visual state, given the state at which the commented action is uttered, induced by the visual stimuli. The action descriptors are independent of both speech-act and the other visual states, given the state at which the action is expected. The interpretation of each phrase (the labels) is independent of any visual state given the time at which the action is seen and it depends on the interpretation of other speech-acts only via the neighbouring system. The specific dependencies of these variables is represented in Figure 1, right.

The variables interplay is accounted for by the interaction of these two different processes, during an experiment. Here an experiment is specified by a task, like pouring some water from a jug to a glass, visually represented by a video, which is described by a narrative. The narrative refers to both the simple actions and objects space and to the beliefs about what is going on in the scene. The narrative, however, goes beyond the direct denotation as it describes the beliefs concerning preconditions and effects of actions on the afforded objects, in terms of temporal and spatial relations, and eventually other deictic expressions. In this initial formalisation, learning the cross-modal translatability is achieved via the two mentioned processes. As gathered above, each process is defined by an observable and an hidden part. The first process serves the linguistic part and it is defined by a hidden
Markov random field (HMRF), to capture the qualitative-spatial structure of the multi-modal contexts of parts of speech. As said above speech-acts are the observed random variables, while labels provide an interpretation, or word contexts, and thus are the hidden part (see also [4]). The second process is an hidden Markov model with mixture of Gaussian observations (GHMM). Here the observations are the visual features dynamics; these are represented by descriptors of the images sequence which, in turn, are obtained by attention based interpretation of light and motion features. From these processes we obtain a speech-act space, a configuration space, an action space and a state space. The structure and connections of the two processes are schematically illustrated in Figure 1, left.

The HMRF defines a joint probability distribution of sequences of observations, i.e. the speech-parts, and sequences of labels, i.e. the language and interpretation of speech-parts. Assume a collection of states $S^K$ has been learned, so that we have a double indexed graph. Observations are formed by a finite set of phrases $\{y_1, \ldots, y_n\}$ (we do not consider here the speech analysis which is assumed to yield the correct association with a predefined and recorded language corpus) having a $d$-dimensional term space $Y$ indexed by $S^K$, and the random field $p(y)$ is defined by a $n$-dimensional space $Y = \prod_{i=1}^{n} Y_{j_n}$, where $Y_{j_n} = \{y_{j_n} | y_{j_n} \in \Sigma^n\}$, with $\Sigma$ the signature. The hidden field is determined by labels defined by a finite space of terms $X_{j_m} = \{x_{j_m} | x_{j_m} \in \Sigma\}$, with $\Sigma$ the language, i.e. $L = (\Sigma, D, I)$, that is, the language is defined by a signature, a domain and an interpretation. For example the predicate $Q(t_1, t_2)$ is a 2-dimensional term, its interpretation is defined according to the HMRF by any suitable set of pairs of objects whose denotation is specified by a term $t \in \Sigma^n$. Note that we are simply referring to a language and an interpretation in terms of elementary structures, not models (in the logical sense). Models of the language can be induced (see [9]) by the probabilistic relational structure, but are not treated here. The product space $X$ is the space of configurations of the labels. A probability distribution $p(x)$ on the space of configurations of the hidden labels is another random field; on this random field we define a neighbouring system $\delta$ that specifies how labels form subgraphs affine to the time of the visual stimuli. These subgraphs are, then, incrementally extended by the learning algorithm that we cannot describe here for lack of space. Labels specify via the neighbours a set of available interpretations. The random field equipped with the neighbour system $\delta$ is a Markov random field iff $p(x_j | x_{\tau(j)}) = p(x_j | \delta(x_j))$ and the joint probability of the two processes is $p(x, y) = p(y|x)p(x)$, this implies that any function $f : X \mapsto \mathbb{R}$ is supported by precisely the cliques of the graph, and $p(x_j | \delta(x_j)) = \frac{1}{Z} \exp(\sum_{C} V_C(x_j))$, with $V_C = \sum_{j \in C} \lambda_j^C f_j^C(x) = \lambda_j^C f_j^C(x)$, with $\lambda_j^C \in \mathbb{R}$ the parameters of the model and $f_j^C(x) \in \{0, 1\}$ the features of the field. However, as gathered above the joint probability involves also the HMM. In fact, speech-acts (observations of the HMRF) are given in sequences and thus these are synchronised together with the action descriptors, therefore they turn out to be also observations of the HMM (see the Figure 1 on the left), if their
lengths satisfy specific conditions (that is, they are terms not phrases).

For example, suppose that the task is to pour water into a glass. Then the video sequence is interpreted to generate descriptors for the action space. The parameters of the HMM can be learned as usual as far as we assume that observations are multivariate in \( \mathbb{R} \) and states are in \( \mathbb{N} \). Suppose, instead, that by a suitable action space construction, from the video analysis, it is possible to build an action space and that states can be given an interpretation. Thus two more variables are involved, the denotation of variables and the inference of the state properties, as extracted from the speech-acts. Thus, let \( \{ \alpha_i \}_{i=1}^M \) be the generated action space and let \( S \) be the state space of the visually interpreted actions. At time \( t \) a phrase and sparse denotations will be uttered, in the context of the observed scene. Thus the realization of the variables is 
\[
p(y, D|s, \alpha_t, x)P(x|\delta(x)).
\]
However the graph topology is locally induced by the visual stimulus and the utterance. Learning the dependency between the HMM states \( S \) and the HMRF states \( S^K \) is achieved by an incremental learning algorithm that follows closely [8]. The difference is mostly on the initial steps. Here, instead of a normal distribution, the random field is built as a set of cliques induced by the simultaneous association of descriptors and phrases.

In conclusion, labels are ground by the narrative which, on the other hand, describes both pointwise actions and state changes, by speech act explaining the action course and specific modalities concerning time and space features of the action effects and preconditions. The connection of the two learning processes ensures both grounding and signature learning. For example, after the action is executed a specific change of spatial relations is the action effect, a speech act shall serve to designate it. For this task the following objects require a relation to be established: the hand, the jug, the glass, the table and the water. The actions are: approaching the jug, grasping the jug handle, rising the jug, inclining the jug so that the water can spill out, putting down the jug. On the other hand there is an infinite set of possible world states associated with these actions and objects. However we are interested only in a finite state space, in which states are just those that can be specified in a finite time lag. That is, those states that can be uttered by the narrator. For example “now the hand is grasping the glass”, or “the glass now is on the table while before it was on the hand”, or “the glass is on the table and it is full of water but before you filled it it was empty”. Similarly: “pouring the water into the glass has been successful, because now the glass is full of water” and “you want to pour water in the glass because someone wants to drink it”.

References


Learning Ability Models for Human-Robot Collaboration
Alexandra Kirsch and Fan Cheng

Abstract—Our vision is a pro-active robot that assists elderly or disabled people in everyday activities. Such a robot needs knowledge in the form of prediction models about a person’s abilities, preferences and expectations in order to decide on the best way to assist. We are interested in learning such models from observation. We report on a first approach to learn ability models for manipulation tasks and identify some general challenges for the acquisition of human models.

I. PLAN-BASED CONTROL FOR HUMAN-ROBOT COLLABORATION

We are interested in planning and plan execution mechanisms that allow a robot to actively participate in joint tasks with a human partner, especially in assistive scenarios. We assume that in such a task there will only be limited explicit communication similar to the way when humans participate in a shared task.

One crucial factor in the development of such joint planning abilities is knowledge about the capabilities and preferences of the partners involved. When helping a person, a robot should take over those tasks that are difficult to perform for the person, but feasible for the robot. For example, an elderly person with a walking impairment can be supported by bringing items she needs to prepare a meal. But the robot should not attempt to cut the ingredients: first, current robots are not capable of doing such sophisticated manipulation tasks reliably, so the probability of failing would be very high and second, the robot should leave work to do for the person to keep her active and healthy and not to intrude into her private life more than necessary.

There is a wide variety of models that are interesting for a robot to interact with a human. Models describing the workspace of robots and humans [4] can be used to coordinate actions in joint workspaces or to choose high-level actions that avoid spatial conflicts. Another approach is to model criteria of social comfort into planning algorithms [1]. In our work, we try to describe capabilities on an action level, for example if a specific manipulation task would succeed in a given situation. We are interested in criteria such as success probability, effort for a person, efficiency and social acceptability (for example it might not be appropriate for a robot to touch food).

Because capabilities and preferences vary strongly among individuals, we would like a robot to learn models about itself and its human partner from experience. In the following we present our observations from a first attempt to learn the capabilities of agents to pick up and put down objects. We report on the challenges that we found in this work and summarize them in a general way to identify problems for similar learning problems.

II. LEARNING ABILITY MODELS

For our research on joint human-robot activities we use a physical simulation of two agents (that are both displayed as robots): one is acting as an autonomous robot, the other one is controlled by a human via the keyboard. A person can move such an agent freely in the world and give commands for gripping and putting down objects. The gripping and put down actions are executed autonomously based on heuristics. These manipulation actions are implemented in different ways for the autonomous robot and the human-controlled one, so that the capabilities are not identical.

In a user study [2] we acquired data from nine subjects who had the task to set and clear the table in two simulated kitchen environments. In total, we observed about 60 gripping and put-down tasks respectively from the execution of complete plans for each participant. For learning capability models, we assumed that all participants were equally skilled in the manipulation tasks, because those were executed autonomously. This is not completely true, because the success of the task also depends on the position where the agent is standing while gripping. But we doubted that any learning algorithm would succeed with the small number of samples we had for each participant and with this simplification we had around 700 examples in total (This includes gripping and put-down tasks that were performed in incomplete runs that were aborted for some reason. This data was not used for evaluation in the user study, but can be used for our purposes here). Beside the data from the user study, we also collected analogous data for an autonomous robot.

As a first approach, we tried to learn prediction models of when such tasks succeed or fail by using decision trees (using the Weka J48 algorithm), for example this function for putting down an object: \textbf{object-goal-position} × \textbf{object-type} → \textit{success/failure}. The object positions were given relative to the piece of furniture they were standing on or had to be put on, which was general enough given the samples from predefined scenarios from our user study. The result of these learning attempts was not surprising, but still disappointing: the decision tree judged that both the robot and the human-controlled agent will succeed in all cases. Looking at the learning experience, the reason for this result was very obvious: only about 4% of the gripping tasks from the user study were identified as failures and 7% of put-down task. The rates for the autonomous robot are similar.

Beside the low failure rate overall, there was no obvious structure when manipulation tasks succeed and fail. One hypothesis was that the wall behind the worktop would be
a crucial factor. Since we used only one kind of kitchen furniture, the way we specified positions should have accounted for the closeness to the wall. But it seems there were not enough samples in the training set to identify this in the learning process.

We got slightly better results when including the distance of the agent to the robot as an input variable. This doesn’t give a model about the capability of gripping the object, because the agent might move, but it could be interesting as a comfort model to predict if a person would have to move and if so this might be an indication of low comfort and a good opportunity for assistance.

Another drawback of our approach is the way we define success and failure of gripping. We used a local decision after each grip and put down task to decide if the object was in the agent’s gripper or the object was put near enough the goal position. However, when observing longer tasks, it becomes obvious that this is not the only source of failures that can occur. Figure 1 shows the end position of a scenario in which two plates, cups and knives had to be brought from the table to the worktop. In both runs all the put down actions were rated as successful. In the left picture, this is true, but in the right picture, the agent involuntarily changed the positions of other objects and the result was anything but satisfactory.

This leads to the question of how to model the state space of such activities. Which other objects apart from the one to be manipulated might be affected by the task and which parameters of these objects are important for the action execution? We are not aware of learning algorithms dealing with a varying number of input parameters. Therefore, the existence of a varying number of objects would have to be coded into a fixed number of parameters.

We continue our work on how to learn prediction models in several ways:

- We are currently exploring situations where two objects are affected by a manipulation task (one is an obstacle, the other one has to be moved) and try to work towards general representations of the state space that is suitable for learning and could be generalized to multi-object cases. One problem here is that the state space will be bigger, which means that more experience is needed for learning. As long as we can use a simulation to generate examples, this is fine, but acquiring enough information in real-world environments where the observations can be expected to be more noisy, this will be a serious problem.
- The Boolean outcome of success or failure seems to be inappropriate for describing the outcome of actions. A continuous value might be more appropriate both from the viewpoint of how to learn it and from the decision-making process of the robot. The robot can then decide at runtime, which level of quality is needed for a task in a certain situation and can trade off different possibilities for distributing the tasks of joint plans.

III. CHALLENGES AND FURTHER RESEARCH

From our first approach to learning predictions models, we identify some general problems of representing and learning models about human and robot activities. By tackling these issues we hope to develop general mechanisms for action modeling that can equally be applied to other predictions like social acceptability, the effort needed for a task, or personal comfort and for all kinds of actions on different levels of the action hierarchy.

These are the main challenges that we have identified:

- The structure of our models is currently a simple function. However, we think that the temporal component of actions (e.g. modeled with hybrid automata) and their effects in the world must also be represented. In our example we realized that the action as such is not the crucial question, but the situation in the world plays an important role. And it is not sufficient to find a clever representation for one specific action: there are too many to be modeled individually. We rather need general representation methods and possibly ways to acquire them automatically by the robot using and observing those actions.
- We will have to consider other forms of adaptation than currently used learning algorithms. Standard machine learning methods have two drawbacks for learning the models we are interested in: 1) they need a large number of samples, which might be difficult to obtain from observing humans, 2) they work on functional representations, which is not directly compatible with the temporal representations we would like to use.
- In our attempt to learn models of actions, we relied only on learning. However, there are explicit reasoning methods, e.g for spatial reasoning [3], that could help to make predictions. When reasoning explicitly about the geometry in the world, the individual differences in capabilities and personal preferences of humans are hard to include. Besides, there are the same questions of how to model the relevant part of the world to make geometric reasoning feasible. We think that a combination of explicit reasoning and learned predictions would be a promising way to model human and robot actions.

REFERENCES

Abstract—In this paper, we present the development of postural expressions of emotions for the humanoid robot Nao in an assistive context. The approach is based on the adaptation of human body postures to the case of the Nao robot. In our paper, the association between the joints of human body and the joints of Nao robot are described. The postural expressions are studied for three emotions - anger, sadness, and happiness. In our experimental design, we generated 32 postural expressions for each emotion and based on the questionnaire-based interview of ten external observers, we selected the best five postural expressions for each emotion. The results of this work will be integrated in a study designed for children with autism.

Keywords-human-robot interaction, social behavior, emotions.

I. INTRODUCTION

In the past few years, many studies have tried to equip robots with emotional expressiveness [1, 3, 4]. Having robots with naturally social behaviors in interaction with humans is a challenge in the human-robot interaction field. Postural expressions of emotions constitute an important part in the social interaction between the humanoid robot and the human. This study aims at developing emotional postures to be used with Nao robot in an assistive context, designed especially for children with autism.

To the best of our knowledge, a lot of the existing studies that are equipping robots with emotions, are focusing on emotional facial expressions. The Ekman six emotions: happiness, fear, surprise, sadness, anger, and disgust have become to be known as the six basic emotions [2]. Breazeal in [3], describes a social behavior architecture that allow the Kismet robot to express various emotions through the different parts of the facial construct. In [4], the authors, are describing a humanoid robot capable of facially expressing Ekman’s six emotions. Compared to facial expressions, body postures are less studied in humanoid robot applications. The authors in [4] used body postures to accompany the facial expressions. However, in their work, a single body posture for each emotion was generated. As far as we know, no extensive research work in robotics concentrates on generating sets of body postures for conveying specific emotions.

Most of the existing studies try to generate human postures in virtual environments by using simplified simulation of human body [5,6,7]. In these studies various postures are generated and shown to observers on computer screen. The feedback of the observers is used to verify the emotional content of body postures. The work of Coulson [8] is remarkable since it provides not only the visual content but also the quantitative joint angle values for various postures for the six emotions. In this work, we take inspiration from Coulson’s work and try to generate postural expressions of emotions with our humanoid robot, Nao.

II. POSTURAL EXPRESSIONS OF EMOTIONS WITH NAO ROBOT

The research question that we would like to address in this paper is whether it is possible to generate postural expressions of emotions with a humanoid-robot. Moreover, our research question can be reduced to whether the postural expressions of emotions of human body convey the same kind of emotions when expressed through a robot body.

In order to answer this question, the human postural expressions of emotions [8] are adapted to the Nao robot. This adaptation requires slight modifications of the postures considering the constraints of the robot’s joints which do not exist in human body. The range of joint angles of the Nao robot is not as large as the human body. The number of degrees of freedom of the joints is also less. Another constraint is due to the difference in mass distribution throughout the robot body. In some postures, the human body can stand stably, which is not the case of Nao robot because the ratio of the mass of the head to body is larger in Nao.

Among the six postural expression of Ekman’s emotions studied by Coulson [8], we chose to focus on the most successful ones - anger, sadness, and happiness. These three emotions were attributed to a large number of postures with some of them reaching a consensus level of 90% among the participants. The most successful postures, all of which reached a consensus level of more than 90%, are given in Figure 1 for the three emotions (i.e., anger, happiness, and sadness). In our study, we use the 32 postures for each of these three emotions and determine five most successful postures for each emotion for the Nao robot. The data of Coulson [8] is transformed to fit the joint configuration of the Nao robot.
III. EXPERIMENTAL RESULTS WITH NAO ROBOT

Our aim is to come up with five successful postures for each of the chosen emotions.

The process of elimination was based on the feedback of ten external observers. A video for each emotion with 32 different postural expressions was generated. The participants watched the three videos of postural expressions of the chosen emotions (i.e., anger, sadness, and happiness) and gave feedback about the success of postures to convey the emotions. Each video showed the postures of the associated emotion sequentially. These videos are available on our website [9]. The participants did not know about the related emotion before watching the videos. After seeing a video they were asked to relate it with one of the Ekman’s six emotions - anger, disgust, fear, happiness, sadness, and surprise. The first results show that although the associated emotions were mentioned by some participants they were not the most selected ones in the three cases: the postural expressions of anger were associated 32% of the times with sadness, 27% with surprise, 22% with anger, 14% with fear, and 5% with happiness; the postural expressions of happiness were associated 50% of the times with happiness, 30% with surprise, 15% with fear, and 5% with anger; and finally the postural expression of sadness were associated 35% of the times with fear, 20% with disgust, 20% with surprise, 15% with sadness, and 10% with anger. This is an indicator that the postures related to an emotion are not consistently conceived as representing that specific emotion as a whole. It is also an indication that only the best within the groups should be chosen to represent the associated emotion with the Nao robot.

As a second phase of our experimental design, the participants were informed about the emotion associated with each video. They were asked to stop the video when they feel that the postures shown at that time is “strongly conveying” the associated emotion. In this phase, all the participants were convinced that the postures were related to the associated emotion and they could identify a number of postures that strongly represented the correct emotion. Five postures were determined for each of the three emotions. This determination was based on the number of times the postures were chosen and very close postures were eliminated. For example, for the case of happiness very close postures were consistently chosen by the participants: the arms were raised in a similar way and only the inclination of chest or head was different. In such cases, only one of those similar postures is chosen as the best representative; the other best representatives are determined among other postures chosen by the participants.

For the postures of anger the statistics were as follows. The participants stopped the video for 51 times. These 51 instances correspond to 23 different postures of anger among the 32. The first two of the chosen postures, anger_32 and anger_18 were the ones that were indicated the most, 5 and 4 times respectively (see Figure 2). Seven postures were indicated three times. Among those, three of them were chosen, anger_3, anger_5, and anger_24, with consideration of not being similar to the previous. The indexing of the postures follows the sequence shown on the videos on our website [9].

For the postural expressions of sadness the participants stopped the video for 65 times, corresponding to 24 different postures of happiness. The first two of the chosen postures, happiness_5 and happiness_29 were the ones that were indicated the most; 8 and 7 times respectively (see Figure 3). One posture was indicated for six times, and two postures were indicated for five times; however, these postures were very close to the first two selected by differing only with slight head or chest bending. Therefore, they were not selected. Four postures were indicated three times, among those two, happiness_3 and happiness_4, were chosen with the consideration of not being similar to the previously chosen ones. The last selected posture was indicated only once, happiness_18, but it was quite different from all previously chosen postures.

For the postural expressions of sadness the participants stopped the video for 28 times, corresponding to 13 different postures. Three postures were indicated four times; two of them were chosen, sadness_4 and sadness_16 (see Figure 4). The third one was very similar to the previous ones. One posture was indicated three times; it was chosen as the third appropriate posture, sadness_7. Four postures were indicated two times; two of them were chosen sadness_15 and sadness_31.

In Figure 2, 3, and 4 the five selected postures are given for the emotions of anger, happiness, and sadness, respectively. The most successful posture of anger by Coulson [8] in Figure 1, corresponds to the posture angry_18 in Figure 2. The most successful posture for happiness by Coulson [8] in Figure 1, was indicated by the participants two times, but not chosen as one of the best five in our case. However, it is remarkable that this posture is very close to the mostly indicated posture of happiness_5. The most successful posture for sadness by Coulson in Figure 1, is not indicated by the participants. This is because the backward leaning of this posture is often perceived as conveying the disgust emotion, rather than sadness. However, considering the upper body, the posture sadness_16 is very close to that described by Coulson in [8]. In summary, the best human body postures described by Coulson in [8] and the best ones determined for the Nao robot for the three chosen emotions (i.e., anger, happiness, and sadness) are either the same or very close to each other.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Front</th>
<th>Side</th>
<th>Rear</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td><img src="image" alt="Anger_1" /></td>
<td><img src="image" alt="Anger_2" /></td>
<td><img src="image" alt="Anger_3" /></td>
</tr>
<tr>
<td>Happiness</td>
<td><img src="image" alt="Happiness_1" /></td>
<td><img src="image" alt="Happiness_2" /></td>
<td><img src="image" alt="Happiness_3" /></td>
</tr>
<tr>
<td>Sadness</td>
<td><img src="image" alt="Sadness_1" /></td>
<td><img src="image" alt="Sadness_2" /></td>
<td><img src="image" alt="Sadness_3" /></td>
</tr>
</tbody>
</table>

Figure 1. Most successful postures for anger, happiness, and sadness by Coulson, viewed from three different sides. [Adapted from [8].]
IV. CONCLUSION

The participants could not consistently name the associated emotion for the videos by only watching. However, they were all convinced about the correct emotional content of the postures after they were told - they could confidently identify the ones that best conveyed the emotion afterwards. This observation points out that context is of paramount importance in order to associate a posture with the emotion. In our case, mentioning the emotion related to a video corresponded to generate a context in which the postures should be viewed. When the observers expect an emotional content, they easily identify the postures to be related to that specific emotion.

This study demonstrates that a humanoid-robot can convey appropriate emotions through its posture. Furthermore, it also demonstrates that the postural expressions of emotions of the Nao robot can be generated based on the postural expressions of emotions as we know it from humans. The selected postures in this study should be further tested to convey the associated emotions in a contextual framework, without verbally mentioning the name of the emotion. We aim at performing such a test in an assistive context, with children with autism. The context is constructed in the form of a game. The game motivates the children to pay attention to the robot. They are explicitly and/or implicitly asked for the emotions conveyed by the robot.

REFERENCES

Pet robots with social skills for satisfactory interaction with hospitalized children

Marta Díaz, Joan Sàez, Diego Pardo and Cecilio Angulo

Abstract—In this work we identify essential requirements to generate a satisfactory social bond between hospitalized children and pet-type robots. A high quality human-robot interaction is required to build up a companion relationship between child and robot. The owner-pet relationship is more complex than the typical interaction between a child and a toy-robot during a play sequence. Companionship implies living-like autonomy and affective involvement (attachment and concern). This kind of social link might provide therapy relevant effects like entertainment, relief, support and enjoyment to hospitalized children in the way real pets do.

INTRODUCTION

Long term hospitalization is a serious event that dramatically affects children and their family's lives. Hospitalized children are confronted with stressful conditions including physical pain and fear. Social support becomes almost limited to hospital staff and relatives, which are often affected themselves by feelings of sorrow and concern. In a hospital environment, animal-assisted activities that have been proven to be effective for pediatric purposes [1] are not possible. Robots have already been proposed to be used for the study of child development [5], [6], rehabilitation [7] and autism therapy [8], [9]. Recently, companion robots have been introduced to reproduce the social-emotional benefits associated with the interaction between children and companion animals [2].

The relationship between master and pet is based on hierarchy and attachment. We assume that if a pet robot with social skills according to these two dimensions is provided, a sort of master-pet bond would result. Hierarchy means that children have an obvious higher status (he/she is cleverer, more skilful, bigger, stronger, more social effective, and more resourceful in general as a human being). This status would be enhanced and more evident if the robot-pet has a baby appearance. This relevant status has consequences not only in the pet-master interaction (dyadic interaction), but in the children social network (triadic interaction). In the dyadic relation several roles are defined: taking care, protect, educate and control and demands respect, obedience and deference.

At social level, this special and unique link with the pet must to be recognized by others agents: parents, friends, acquaintances, hospital personnel. The child will feel that this link provides something valuable and desirable that nobody else has. Furthermore this status gives the child the position to intermediate and control the interaction between the pet and other people.

SOCIAL SKILLS

We assume that in the social framework defined by the master (child) / pet (baby) interaction, desirable situations would naturally emerge:

- Engaging activities: teaching new skills, learning to understand it, care giving, playing together, humor, unexpected situations, and surprise.
- Experience the warmness of pet-robot unconditional and reliable attachment, preference and affection.
- Obtaining others appreciation, admiration, and amazement.
- Satisfaction for achievement and fulfillment: acquisition of new skills, capability to be in charge, to give care etc.

These experiences could be positive in order to enhance the child's auto esteem, auto efficacy, perception of control, expression of emotions, distraction, enjoyment, activity, social facilitation, lowing the stress -even the pain perception-, and relief feelings of sorrow, loneliness and boredom.

In this context the robot -besides considerations of appearance, mainly life-like and baby-like features- must be able to deploy (or acquire) the following social skills:

1) Social skills for effective communication.
   - Activation. Show that communication is possible: I’m awake.
   - Orientation. Show that the presence is detected and maybe a request for interaction: I’m aware.
   - Attention. I know your interacting with me: I’m focused.
   - Response (latency). I’m taking an active part in interaction: I’m involved.

2) Social skills related to hierarchy.
   - Recognition. Perform at least two clearly different behavior patterns: interacting with my master / interacting with anyone else. The difference would be in orientation, responsiveness -to call, demands-, obedience, activation. Anyway it would be desirable to introduce a third pattern to enrich the triadic dynamics: with my master/with my master’s friends/with anyone else. These three classes categorization enhance the child’s status giving him/her the power to organize the social network and to grant social privileges.
   - Submission. Respond to admonition (e.g. NO!), punishment and praise.

3) Social skills related to attachment. Clear distinctive approach behavior (looking forward physical contact,
closer interaction distance, body orientation), and positive emotion (happiness, attention, activation) when interacting with its master. For developing these skills, the robot-pet at least must be capable of:

- Perform two kind of distinct behavior: Approaching (asking for interaction) / Distancing (withdrawing from interaction).
- Express two kind of social-related emotion: Positive (happiness)/negative (fear, sorrow).
- My master / anyone else, My master / my master’s friends / anyone else.
- Respond to social stimuli:
  - Recognition of at least one negative reinforcement (e.g., NO!)
  - Recognition of at least one positive reinforcement (e.g., cuddle, tap on the head)
  - Recognition of its name

As a general rule, the interaction must provide a balance between unpredictability that accounts for autonomy and subjectivity essential for life-likeness metaphor- and consistence. Unpredicted behavior (e.g. the pet does not engage in a thrown and pick up game as usual) and may excite curiosity and reinforce exploration activity provided there is something to understand and a rule to discover (e.g., the pet is tired). The understanding of pet behavior rules would result as well in a more effective and successful interaction (e.g., pick up the pet and cuddle it instead of kicking it running around).

In this scenario, the pet robot is able to interact with its social environment and distinguish between people and other agents (e.g. based on sensor information), therefore it falls within a socially situated robot [4]. A socially situated robot does not need to possess any model of ‘social intelligence’, ‘social interactions’ emerge from the robot being situated in and responding to its environment.

**MACHINE LEARNING**

The proposed unpredictability and nature of the robot reactions impose constraints to the robot control scheme. Robot behavior can not be obtained using a deterministic approach with all the input-output relations being previously handwritten. Instead, an adaptive connectionist framework is suggested as initial approach, so robot conduct is an emergent phenomena arisen from the interplay of simple predefined units. These units contains basic behaviors, encoding instinct actions of the pet, e.g., generating adequate sounds to communicate emotions (happiness, sadness, fear, curiosity...), other basic behaviors may include primitive motions like reaching, exploring, etc. Inputs for the robot are provided by its environment, mainly from the user interaction. They are processed, combined and mapped in order to trigger a convenient behavior unit. This mapping must be adaptable, e.g. using a probabilistic approach, modifying the conduct of the robot, and building the desired owner-pet relationship presented in the previous sections.

Implementing this approach, users may have the opportunity to reinforce those connections between adequate inputs and desired/undesired behaviors units. The owner will shape the pet reactions and behaviors, generating an artificial personality, with the positive implications in the rehabilitation process.

**CONCLUSIONS**

This communication supports the use of robots to bring those therapy-related effects achieved with real pets to the hospitalization scenario. Pet robots, with adequate social skills, will help hospitalized children in their social emotional development, generating entertainment, relief, support and enjoyment, which are therapy relevant effects. In order to achieve this, here we have presented a description of minimum requirements that roboticists should find interesting in the development of a new type of companion robot.

The desired children-robot relationship is obtained enabling the robot with the described interaction capabilities. The master-pet relationship is discovered by the child, giving him/herself the opportunity to develop an active role, valuable for the child’s recovery process.

**REFERENCES**