IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2011): Workshop on

Cognitive Neuroscience Robotics

Date & Venue

September 25, 2011

San Francisco, California, US

Organizers

Kenichi Narioka (Osaka University)

Yukie Nagai (Osaka University)

Minoru Asada (Osaka University)

Hiroshi Ishiguro (Osaka University)

Global COE Program
Center of Human-friendly Robotics Based on Cognitive Neuroscience

http://www.gcoe-cnr.osaka-u.ac.jp/english/

Program

8:30 - 8:40	AM1-0	Welcome
8:40 - 9:20	AM1-1	Keynote: **GCOE for Cognitive Neuroscience Robotics** - Hiroshi Ishiguro
9:20 - 10:00	AM1-2	**Cognitive Neuroscience Meets Robotics: Perception of Human and Robot Movements** - Ayse P. Saygin (Invited)
10:00 - 10:20		(Coffee break)
10:20 - 11:00	AM2-1	**In Relation with Social Robots** - Peter H. Kahn, Jr. (Invited)
11:00 - 11:25	AM2-2	**Social Stories for Autistic Children Told by the Huggable Robot Probo** - Bram Vanderborght, Ramona Simut, Jelle Saldien, Cristina Pop, Alina S. Rusu, Sebastian Pintea, Dirk Lefeber and Daniel O. David
11:25 - 11:50	AM2-3	**Computational Audiovisual Scene Analysis for Dialog Scenarios** - Rujiao Yan, Tobias Rodemann and Britta Wrede
11:50 - 12:50		(Lunch)
12:50 - 13:50		(Poster Session)
	Pos-1	**Restricted Boltzmann Machine with Transformation Units for Laser Image Processing in a Mirror Neuron System Architecture** - Junpei Zhong, Cornelius Weber and Stefan Wermter
	Pos-2	**Towards Robust Speech Recognition for Human-Robot Interaction** - Stefan Heinrich and Stefan Wermter
	Pos-3	**On-line learning of the sensorimotor transformation on a humanoid robot** - Marco Antonelli, Eris Chinellato and Angel P. del Pobil
	Pos-4	**Our robot is more human than yours: Effects of group membership on anthropomorphic judgments of social robots** - Friederike Eyssel and Dieta Kuchenbrandt

13:50 - 14:30	PM1-1	**Mobility Training of Infants and Toddlers Using Novel Mobile Robots and Interfaces** - Sunil K. Agrawal (Invited)
14:30 - 14:55	PM1-2	**Collecting A Developmental Dataset of Reaching Behaviors: First Steps.** - Tingfan Wu, Juan Artigas, Whitney Mattson, Paul Ruvolo, Javier Movellan and Daniel Messinger
14:55 - 15:20	PM1-3	**Elevated activation of dopaminergic brain areas facilitates behavioral state transition** - Beata J. Grzyb, Joschka Boedecker, Minoru Asada and Angel P. del Pobil
15:20 - 15:40		(Coffee break)
15:40 - 16:05	PM2-1	**Mutual adaptive interaction between robots and human based on the dynamical systems approach** - Tetsuya Ogata
16:05 - 16:30	PM2-2	**The Cognitive Nature of Action - A Bi-Modal Approach towards the Natural Grasping of Known and Unknown Objects** - Kai Essig, Jonathan Maycock, Helge Ritter and Thomas Schack
16:30 - 17:10	PM2-3	Keynote: **The Future of Cognitive Neuroscience Robotics** - Minoru Asada

AM1-1

GCOE for Cognitive Neuroscience Robotics

Hiroshi Ishiguro

Dept. of Systems Innovation, Graduate School of Engineering Science, Osaka University, Japan

This COE aims to develop new IRT (Information and Robot Technology) systems that can provide information and services based on understanding the meta-level brain functions of humans. We call this new interdisciplinary research area "Cognitive Neuroscience Robotics," which integrates our studies in robotic, cognitive science, and brain science, being conducted at Osaka University and ATR.

Our robotics research is distinguished by human-oriented studies. We aim to understand human intelligence and cognitive development by developing humanoids and androids. Our studies on human-robot interaction and on modeling of cognitive development address basic problems among human and machines, i.e., what is human intelligence. Studies on cognitive science have advanced with technologies. Systems for measuring human gaze and gestures as well as biological information have contributed it. Robotics and sensory network can be next promising approach. Therefore, this COE integrates these with cognitive science in order to develop human-friendly robots as well as to reveal human cognition. Our studies on brain science will further enhance the interdisciplinary research. It will enable us to uncover human cognitive function and then to propose a new design principle for IRT systems.

Citation: Saygin, A.P. "Cognitive Neuroscience Meets Robotics: Perception of Human and Robot Movements." Workshop: Cognitive Neuroscience and Robotics. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2011) Sept 25, 2011.

Cognitive Neuroscience Meets Robotics:
Perception of Human and Robot Movements

Ayse Pinar Saygin
Department of Cognitive Science
University of California, San Diego
La Jolla, CA, USA
saygin@cogsci.ucsd.edu

Abstract— **This paper describes an ongoing interdisciplinary collaboration between robotics and human cognitive neuroscience research, focusing on body movement perception. We argue such research is beneficial to both disciplines and can be viewed a win-win. Not only it is important to understand the human factors in human-robot interaction, collaborations with robotics can also help neuroscientists answer questions about human brain and behavior, and take steps toward understanding how the human brain enables some of our most important skills such as action understanding, social cognition, empathy, and communication.**

Social robotics, neuroimaging, fMRI, uncanny valley

I. Introduction

From creation myths to modern horror stories (e.g., Frankenstein), humans have long been preoccupied with creating other entities in their likeness. Advances in technology now allow us to create increasingly realistic and interactive humanoid agents. Lifelike humanoid robots are becoming commonplace; and assistive technologies are changing the face of education and healthcare [1,2]. Research on how humans perceive, respond to and interact with these agents is increasingly important. However little is understood about human social cognition in this new context. An interdisciplinary perspective on human-robot interaction (HRI) is important, since this field will impact issues of public concern in the near future, for example in domains such as education and healthcare [3-5]

Our research brings together robotics with cognitive neuroscience to explore how humans perceive, respond to, and interact with others, including artificial agents, and specifically, robots. The research program is interdisciplinary, spanning the biomedical and social sciences and engineering. Neuroscience and psychology research exploring HRI can make valuable contributions to robotics. Conversely, experiments in collaboration with roboticists can help advance neuroscience by allowing us to ask new questions, or control parameters we cannot manipulate in biological systems.

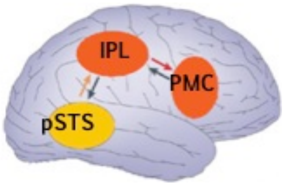


Figure 1: Schematic of the Action Perception System (APS), consisting of posterior Superior Temporal Sulcus (pSTS), Inferior Parietal Lobule (IPL), and Premotor Cortex (PMC). Figure adapted from [10].

The goal of this research program is to both improve our understanding of how the human brain enables social cognition, and to help engineers and designers in developing interactive agents that are well-suited to their application domains, as well as to the brains of their creators.

II. Action Perception and Robotics

In primates, the perception of body movements is supported by network of lateral superior temporal, inferior parietal and inferior frontal brain areas. Here, we refer to this network as the action perception system, or APS (Fig. 1). Two of the areas within the APS, (PMC and IPL) contain *mirror neurons* in the macaque brain [6]. Mirror neurons respond not only when a monkey executes a particular action, but also when it observes another individual perform the action. For instance a mirror neuron that fires as the monkey cracks a peanut, can also fire as the monkey observes someone else crack a peanut. It is thought that a similar system underlies action perception in the human brain [7-10]. Some researchers have argued that in addition to subserving action processing, the APS helps in linking "self" and "other" in the brain, and thus may constitute a basis for social cognition [6].

The finding that the visual perception of another entity automatically engages the observers' own motor system has important implications for the field of human-machine interaction. It would be fair to say that in the nervous system, simply seeing another agent automatically engages interaction.

This research was supported by the Kavli Institute for Brain and Mind (KIBM), California Institute for Telecommunications and Information Technology (Calit2), Japan Society for the Promotion of Science (JSPS), the Wellcome Trust, and the European Commission.

The APS has received intense interest from neuroscientists in the last decade and a half, and we can now use the accumulated knowledge in this field to study how the human brain supports HRI. Conversely robotics can help research on the human brain by allowing us to test functional properties of the APS and other brain areas. For example, we may ask, during interactions with robots, does the brain rely on the same or distinct processes as with interactions with humans?

Due to the presence of mirror neurons, the neural activity in PMC and IPL regions during action perception is often interpreted within the framework of "simulation": A visually perceived body movement is mapped onto the perceiving agent's sensorimotor neural representations and "an action is understood when its observation causes the motor system of the observer to 'resonate"[11]. But what are the boundary conditions for 'resonance'? What kinds of agents or actions lead to the simulation process? Is human-like appearance important? Is human-like motion?

On the one hand, it seems reasonable that the closer the match between the observed action and the observers' own sensorimotor representations, the more efficient the simulation will be. In support for this, the APS is modulated by whether the observer can in fact perform the seen movement [12]. The appearance of the observed agent may be additionally important [13, 14]. On the other hand, human resemblance is not necessarily always a positive feature in robots. The "uncanny valley" hypothesis has proposed that as a robot is made more human-like, the reaction to it becomes more and more positive, until a point is reached at which the robot becomes oddly repulsive [15]. While this phenomenon is well known to roboticists and computer animators, there is little scientific evidence in favor of or against it [16-21].

Robots can have nonbiological appearance and movement patterns – but at the same time, they can be perceived as carrying out recognizable actions. Is biological appearance or biological movement necessary for engaging the human Action Perception System (APS)? Robots can allow us to ask such questions and to test whether particular brain areas are selective or sensitive to the presence of a human, or an agent with a humanlike form, or respond regardless of the agent performing the action.

There is a small neuroscience literature on the perception of artificial agents, including robots [22-25]. Unfortunately, the results are highly inconsistent. Many studies had used toy robots or very rudimentary industrial robot arms, so the results were not informative regarding state-of-the-art robotics. Furthermore, the roles of humanlike appearance or motion were not explored in previous work. We used neuroimaging (functional Magnetic Resonance Imaging (fMRI)) along with a method called Repetition Suppression (RS) and well-controlled stimuli developed by an interdisciplinary team, allowing us to overcome limitations of previous work [26, 27].

III. Neuroimaging Study: Perception of Robot Actions

We performed fMRI as participants viewed video clips of human and robotic agents carrying out recognizable actions. fMRI is a powerful method that allows imaging the activity of the live human brain non-invasively and has revolutionized neuroscience. We worked with Repliee Q2, an android developed at Osaka University in collaboration with Kokoro Ltd [28, 29]. Repliee Q2 has a very human-like appearance (Fig. 2, Android (A)). In order achieve this, the robot's face was modeled after an adult Japanese female (Fig. 2, Human (H)). Repliee Q2 can make facial expressions, as well as eye, head, upper limb, and torso movements. It has 42 degrees of freedom (d.o.f.) in its movements, with 16 d.o.f. in the head. With very brief exposure times, Repliee Q2 is often mistaken for a human being, but more prolonged exposure and interaction can lead to an uncanny valley experience [29].

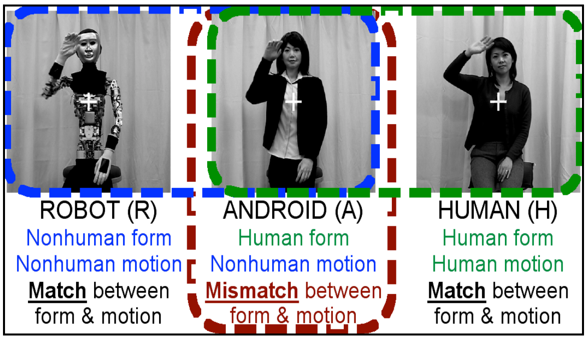


Figure 2: Stills from the videos depicting the three agents (R, A, H) and the experimental conditions (form and motion) they represent.

Importantly, Repliee Q2 was videotaped both in its original human-like appearance (A) and in a modified, more mechanical appearance (Fig. 2, Robot (R)). In this latter condition, we removed as many of the surface elements as possible in order to reveal the electronics and mechanics underneath. The silicone covering the face and hands could not be removed, so we used a custom mask and gloves to cover these areas. The end result was that the robot's appearance became obviously mechanical and nonhuman. However, since the A and R are in fact the same robot, the motion kinematics are the same for these two conditions.

There were thus three agents: human (H), robot with human form (A), and robot with nonhuman form (R). H and A are very close to each other in form, both with humanlike form, whereas R has nonhuman form. In terms of the movement, H represents truly biological motion and A and R are identical, both with mechanical kinematics. Using fMRI and RS, we explored whether the human brain would display specialization for human form (similar responses for A and H, and different for R) or motion (similar responses for R and A, and differential responses for H). Another possibility was for RS responses not to reflect biological form or motion per se, but instead pattern like the uncanny valley. In this scenario, the RS responses to the H and R would be similar to each other, even though these two agents are divergent from each other in both form and movement (Fig. 2).

The articulators of Repliee Q2 were programmed over several weeks at Osaka University. The same movements were videotaped in both appearance conditions (R and A). The human, the same female adult to whom Repliee Q2 was designed to resemble, was asked to perform the same actions as she naturally would. All agents were videotaped in the same room and with the same background. A total of 8 actions per actor were used in the experiment (e.g., drinking water from a cup, waving hand). 20 adults participated in the fMRI experiment. Participants had no experience working with robots. Each was given exactly the same introduction to the study and the same exposure to the videos prior to scanning since prior knowledge can affect attitudes to artificial agents differentially [30]. Before the experiment, subjects were told that they would see short video clips of actions by a person, or by two robots with different appearances and were shown all the movies in the experiment. By the time scanning started, participants were thus not uncertain about the robotic identity of the android.

Scanning was conducted at the Wellcome Trust Centre for Neuroimaging, in London, UK using a 3T Siemens Allegra scanner and a standard T2* weighted gradient echo pulse sequence. During fMRI, subjects viewed the stimuli projected on a screen in the back of the scanner bore through a mirror placed inside the head coil. There were blocks of 12 videos, each preceded by the same video (Repeat) or a different video (Non-repeat), which allowed us to compute the RS contrast (Non-repeat > Repeat). Every 30-seconds, they were presented with a statement about which they would have to make a True/False judgment (e.g., "I did not see her wiping the table"). Since the statements could refer to any video, subjects had to be attentive throughout the block. Data were analyzed with SPM software (http://www.fil.ion.ucl.ac.uk/spm).

RS differed considerably between the agents (Fig. 3). All agents showed RS in temporal cortex near the pSTS. For A, extensive RS was found in additional regions of temporal, parietal and frontal cortex (Fig. 3b).

In the left hemisphere, lateral temporal cortex responded to H and A, but not to R. The specific location of this activation corresponds to extrastriate body area (EBA), a region that responds strongly during the visual perception of the body and body parts [31]. Our data showed that robotic appearance can weaken the RS response in the EBA.

Aside from the left EBA, we did not find evidence for APS coding for human-like form or motion per se. Instead, for the android, whose form is humanlike, but its motion is mechanical, increased responses were found in a network of cortical areas. This was most pronounced (and statistically significant) in the IPL [27], one of the nodes of the APS (Fig. 3b, circled areas).

But why would there be an area of the brain highly selective for androids? This response pattern brings to mind the uncanny valley – except, rather than valleys, we measured "hills" in the neural responses, in the form of increased RS. A framework within which to interpret these data is the predictive coding account of cortical computation [32-34]. Predictive coding is based on minimizing prediction error among the levels of a cortical hierarchy (e.g. the APS). More specifically, during the perception of H and R, there is no conflict between form and motion of the agent. H appears human and moves like a human. R appears mechanical and moves mechanically. For A on the other hand the agent's form is humanlike, which may result in a conflict when the brain attempts to process and integrate the movement of the agent with its form. This conflict leads to the generation of a prediction error, which is propagated in the network until the predictions of each node are minimized. During this process, we can measure the prediction error in the fMRI responses. It is not possible from the current data to know the exact neural sources, the directionality, and the time course of error propagation, but it is clear that the cortical network is engaged more strongly during the perception of A compared with the agents that lead to less prediction error (R and H). Furthermore, the effect is largest in parietal cortex, which is the node of the network that links the posterior, visual components of the APS and the frontal, motor components [27, 35].

In summary, in this interdisciplinary study, we found that a robot with highly humanlike form is processed differentially compared with a robot with a mechanical form, or with an actual human. These differences are found in a network of brain areas, most prominently in parietal cortex [27]. We propose these "hills" in the brain activity reflect the prediction error that is propagated in the system. The uncanny valley may thus arise from processing conflicts in the APS, and the resultant error signals, which can in turn be measured using fMRI [34].

IV. Discussion and Future Directions

Humanoid robots are increasingly part of our daily lives [1-5]. With application in domains such as healthcare, education, communications, entertainment, and the arts, exploring human factors in the design and development of artificial agents is ever more important. This will require an interdisciplinary approach, to which we have contributed new data from cognitive neuroscience.

We have described a neuroimaging study in order to explain how such work can inform both neuroscience and robotics. This study is only a beginning. We are currently using magnetoencephalography (MEG) and electroencephalography (EEG), both of which allow for imaging brain activity with high temporal resolution so that we can study the temporal dynamics of action processing. EEG also allows neuroimaging in more dynamic and interactive settings compared with fMRI.

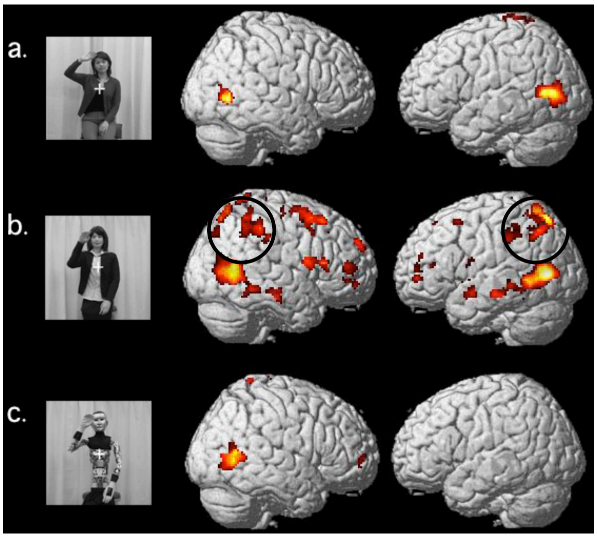


Figure 3. Repetition suppression (RS) results for the Human (a), Android (b), and Robot (c). (Non-repeat > Repeat at t>=8.86, p<0.05 with False Discovery Rate (FDR) correction for multiple comparisons, with a cluster size of at least 30 voxels). Figure adapted from [27].

Using cognitive neuroscience, we have been able to suggest an interpretation for the classic anecdotal reports of the uncanny valley hypothesis. The uncanny valley has many potential dimensions [15, 16]. While our experiments were not designed in an optimum fashion to “explain” the uncanny valley, the results suggest an intriguing link between the phenomenon, and brain responses in the APS. In a predictive coding, the android is not predictable: an agent with that form (human) would typically not move mechanically as Repliee Q2 does. When the nervous system is presented with this unexpected combination, a propagation of prediction error may occur in the APS. We suggest this framework may contribute to an explanation for the uncanny valley and future experiments will test this hypothesis.

Using robotics, we were able to answer questions regarding the neural basis of action and body movement perception, an active research area [6-14]. We found that the brain was not by and large, tuned only to our conspecifics (humans). We were able to test functional properties of human action perception system (APS). These findings help shed light on how our brains enable social cognition, an important skill, and part of what it means to be human.

Interdisciplinary collaboration between cognitive neuroscience and robotics can be a win-win for both sides. Joining forces, we can answer questions in both disciplines, and contribute to the science of tomorrow.

V. References

[1] S. Coradeschi, H. Ishiguro, M. Asada, S. C. Shapiro, M. Thielscher, C. Breazeal, M. J. Mataric and H. Ishida, "Human-inspired robots," IEEE Intelligent Sys., vol. 21, pp. 74-85, 2006.

[2] T. Kanda, H. Ishiguro, M. Imai and T. Ono, "Development and evaluation of interactive humanoid robots," Proc. IEEE, vol. 92, pp. 1839-1850, 2004.

[3] A. Billard, B. Robins, J. Nadel and K. Dautenhahn, "Building Robota, a mini-humanoid robot for the rehabilitation of children with autism," Assist. Technol., vol. 19, pp. 37-49, 2007.

[4] M. Mataric, A. Tapus, C. Winstein and J. Eriksson, "Socially assistive robotics for stroke and mild TBI rehabilitation," Stud. Health. Technol. Inform., vol. 145, pp 249-262, 2009.

[5] K. Dautenhahn, "Socially intelligent robots: dimensions of human-robot interaction," Philos. Trans. R. Soc. Lond. B Biol. Sci., vol. 362, pp. 679-704, 2007.

[6] G. Rizzolatti and L. Craighero, "The mirror-neuron system," Annu. Rev. Neurosci., vol. 27, pp. 169-192, 2004.

[7] A. P. Saygin, S. M. Wilson, D. J. Hagler, Jr., E. Bates and M. I. Sereno, "Point-light biological motion perception activates human premotor cortex," J. Neurosci., vol. 24, pp. 6181-6188, 2004.

[8] A. P. Saygin, "Biological motion perception and the brain: Neuropsychological and neuroimaging studies,". In Visual Perception of the human body in motion: Findings, theory, and practice. K. Johnson and M. Shiffrar, Eds., Oxford University Press, in press, 2011.

[9] S. T. Grafton, "Embodied cognition and the simulation of action to understand others," Ann. NY Acad. Sci., vol. 1156, pp. 97-117, 2009.

[10] M. Iacoboni and M. Dapretto, "The mirror neuron system and the consequences of its dysfunction," Nat. Rev. Neurosci. vol. 7, pp. 942-951, 2006.

[11] G. Rizzolatti, L. Fogassi and V. Gallese, "Neurophysiological mechanisms underlying the understanding and imitation of action," Nat. Rev. Neurosci., vol. 2, pp. 661-670, 2001.

[12] B. Calvo-Merino, J. Grezes, D. E. Glaser, R. E. Passingham and P. Haggard, "Seeing or doing? Influence of visual and motor familiarity in action observation," Curr. Biol., vol. 16, pp. 1905-1910, 2006.

[13] G. Buccino, F. Lui, N. Canessa, I. Patteri, G. Lagravinese, F. Benuzzi, C. A. Porro and G. Rizzolatti, "Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study," J. Cogn. Neurosci., vol. 16, pp. 114-126, 2004.

[14] T. Chaminade, J. Hodgins and M. Kawato, "Anthropomorphism influences perception of computer-animated characters' actions," Soc. Cogn. Affect. Neurosci. vol. 2, pp. 206-216, 2007.

[15] M. Mori, "The uncanny valley," Energy, vol. 7, pp. 33-35, 1970.

[16] K. F. MacDorman and H. Ishiguro, "The uncanny advantage of using androids in cognitive and social science research," Interaction Studies, vol. 7, pp. 297-337, 2006.

[17] K. F. MacDorman, R. D. Green, C. C. Ho and C. T. Koch, "Too real for comfort? Uncanny responses to computer generated faces," Comput. Hum. Beh., vol. 25, pp. 695-710, 2009.

[18] J. Seyama and R. Nagayama, "The uncanny valley: Effect of realism on the impression of artificial human faces," Presence Teleop. Virtual Env., vol. 16, pp. 337-351, 2007.

[19] S. A. Steckenfinger and A. A. Ghazanfar, "Monkey visual behavior falls into the uncanny valley," Proc. Nat. Acad. Sci. USA, vol. 106, pp. 18362-18366, 2009.

[20] A. P. Saygin, T. Chaminade and H. Ishiguro, "The perception of humans and robots: Uncanny hills in parietal cortex," Proc. 32nd Annual Cognitive Science Society, S. Ohlsson & R. Catrambone, Eds., pp. 2716-2720, (Portland, OR, USA), 2010.

[21] A.P. Saygin, T. Chaminade, B.A. Urgen, and H. Ishiguro, "Cognitive neuroscience and robotics: A mutually beneficial joining of forces." Human-robot interaction: Perspectives and contributions to robotics from the human sciences. Robotics: Science and Systems (RSS). Los Angeles, CA, July, 2011.

[22] V. Gazzola, G. Rizzolatti, B. Wicker and C. Keysers, "The anthropomorphic brain: the mirror neuron system responds to human and robotic actions," Neuroimage, vol. 35, pp. 1674-1684, 2007.

[23] C. Press, H. Gillmeister and C. Heyes, "Sensorimotor experience enhances automatic imitation of robotic action," Proc. Biol. Sci. vol., 274, pp. 2509-2514, 2007.

[24] L. M. Oberman, J. P. McCleery, V. S. Ramachandran and J. A. Pineda, "EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots," Neurocomput., vol. 70, pp. 2194-2203, 2007.

[25] Y. F. Tai, C. Scherfler, D. J. Brooks, N. Sawamoto and U. Castiello, "The human premotor cortex is 'mirror' only for biological actions," Curr. Biol., vol. 14, pp. 117-120, 2004.

[26] K. Grill-Spector, R. Henson and A. Martin, "Repetition and the brain: neural models of stimulus-specific effects," Trends. Cogn. Sci., vol. 10, pp. 14-23, 2006.

[27] A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver and C. F. Frith, "The thing that should not be: Predictive coding and the uncanny valley in perceiving human and humanoid robot actions," Soc. Cogn. Affect. Neurosci, vol. 6., In press, 2011.

[28] H. Ishiguro and S. Nishio, "Building artificial humans to understand humans," J. Artif. Organs., Vol. 10, pp. 133-142, 2007.

[29] H. Ishiguro, "Android science: conscious and subconscious recognition" Conn. Sci., vol. 18, pp. 319-332, 2006.

[30] A. P. Saygin and I. Cicekli, "Pragmatics in human-computer conversations," J. Pragmatics, vol. 34, pp. 227-258, 2002.

[31] M. V. Peelen and P. E. Donwing, "The neural basis of visual body perception," Nat Rev. Neurosci. 8(8), 636-648, 2007.

[32] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," Nat. Neurosci., vol. 2, pp. 79-87, 1999.

[33] J. M. Kilner, K. J. Friston and C. D. Frith, "Predictive coding: an account of the mirror neuron system," Cogn. Process., vol. 8, pp. 159-166, 2007.

[34] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," Philos. Trans. R. Soc. Lond. B, Biol. Sci., vol. 364, pp. 1211-1221, 2009.

In Relation with Social Robots

Paper for the IROS2011 Workshop:
Cognitive Neuroscience Robotics

Peter H. Kahn, Jr.
Department of Psychology
University of Washington
Seattle, Washington, USA
pkahn@uw.edu

Many of you at this workshop are at the cutting edge of technical work in robotics. You're doing the best in the world. There's also interest here on the side of cognitive neuroscience. In turn, I would like to show how we are creating a technological world where we will have not just social relationships with robots, but in some ways moral relationships, as well. I would also like to bring forward what I refer to as the new ontological category hypothesis.

To develop these ideas, I draw on three collaborative research studies.

The first study focused on preschool children's social interactions with Sony's robot dog AIBO (Kahn, Friedman, Perez-Granados, & Freier). In this study, we had 80 children, evenly divided into two age groups: 3-4 year olds; and 4-1/2 to 5/1/2 year olds. They came into our lab individually, and had a 20 minute session with AIBO. As a comparison condition, they also had a 20 minute session with a stuffed dog. During the session, we allowed each child to play with each of the artifacts (AIBO and the stuff dog), and in that context we asked them structured questions focused around 5 main categories (that we had developed from a previous study): animacy (e.g., Is AIBO alive?), biology (Does AIBO have a stomach?), mental states (Can AIBO feel happy?), social rapport (Can AIBO be your friend?), and moral standing (Is it ok to hit AIBO?). We also video-taped all the sessions. In terms of the behavioral results, 6 overarching behaviors emerged from the data: (1) exploring the artifact, (2) being apprehensive with the artifact, (3) being affectionate with the artifact, (4) mistreating the artifact, (5) physically animating the artifact, and (6), attempting to engage in reciprocal interaction with the artifact. Quantitatively, we coded 2,360 behavioral interactions between children and AIBO or the stuffed dog. For the purposes of this workshop, the finding I want to highlight is on reciprocal interactions. Children more often attempted to engage AIBO in reciprocal interactions than with the stuffed dog (683 reciprocal behaviors with AIBO compared to just 180 with the stuffed dog). In the moral developmental literature, going back to Jean Piaget's (1932/1969) landmark work on the moral judgment of the child, reciprocity is foundational to moral development: in terms of reciprocal exchange, perspective taking, moving away from heteronomous relationships toward mutual, freely initiated interaction. Thus our data suggest the potential for a moral relationship with animal robots.

The second study investigated whether robotic animals might aid in the social development of children with autism (Stanton, Kahn, Severson, Ruckert, & Gill, 2008). Eleven children diagnosed with autism (ages 5-8) interacted with AIBO and, during a different period within the same experimental session, a simple mechanical toy dog (Kasha), which had no ability to detect or respond to its physical or social environment. Results showed that, in comparison to Kasha, the children spoke more words to AIBO, and more often engaged in three types of behavior with AIBO typical of children without autism: verbal engagement, reciprocal interaction, and authentic interaction. In addition, we found suggestive evidence (with *p* values ranging from .07 to .09) that the children interacted more with AIBO, and, while in the AIBO session, engaged in fewer autistic behaviors. Based on this study, and other work by Scassellati (2005a,b), Dautenhahn (2003), and others, there is good reason to hold out promise for the benefits of social robots and children with autism.

The third study investigated children's and adolescents' social and moral relationships with a humanoid robot (Kahn et al, 2011). Ninety children (9, 12, and 15-year-olds) initially interacted with ATR's humanoid robot, Robovie, in 15-minute-sessions. Each session ended when an experimenter interrupted Robovie's turn at a game and, against Robovie's stated objections, put Robovie into a closet. Each child was then engaged in a 50-minute structural-developmental interview. Results showed that during the interaction sessions all of the children engaged in physical and verbal social behaviors with Robovie. Based on the interview data, the majority of children believed that Robovie had mental states (e.g., was intelligent and had feelings) and was a social being (e.g., could be a friend, offer comfort, and be trusted with secrets). In terms of Robovie's moral standing, children believed that Robovie deserved fair treatment and should not be harmed psychologically, but did not believe that Robovie was entitled to its own liberty (Robovie could be bought and sold) or civil rights (in terms of voting rights and deserving compensation for work performed). In a hypothetical scenario where humans were removed from the situation, and "aliens" sought to treat Robovie as a non-moral entity, the majority of children affirmed Robovie's moral standing across all criteria. Developmentally, while over half the 15-year-olds conceptualized Robovie as a mental, social, and partly moral other, they did so to a lesser degree than the 9 and 12-year-olds.

CONCLUSION

The data from my collaborative studies support the proposition that children's social and moral relationships with social robots of the future will be substantial and meaningful. If this proposition is correct, it will create many puzzling design decisions. For example, if we create robot nannies to help look after our children, do we design the robot to do everything the child asks of it? If so, does that reify a master-servant relationship? If we do not, if we design the robot to push back and assert its own moral selfhood, that seems strange, too; because these robots are our technological creations, and arguably have no selfhood to assert.

People will continue to develop social and moral relationships with robots, but it is unclear how exactly these relationships will take shape. One of the reasons builds on the last point I would like to make that involves what I refer to as the new ontological category hypothesis.

In philosophy, ontology refers to basic categories of being, and ways of distinguishing them. My lab's hypothesis is that a new ontological category (a) may be emerging through the creation of social robots and (b) may continue to emerge as other embodied social computational systems (e.g., "smart" cars and homes of the future) become increasingly pervasive.

Let me unpack the idea in this way. In the Robovie study, we had asked children whether they thought Robovie was a living being. Results showed that 38% of the children were unwilling to commit to either category, and talked in various ways of Robovie being "in between" living and not living or simply not fitting either category. As one child said: "He's like, he's half living, half not." It is as if I showed you an orange-colored object, and asked you, Is this object red or yellow? You might say neither and both. You might say that while you understand the question, and that aspects of the question certainly make sense, that when we combined red and yellow together we created something uniquely its own. That may be our trajectory with robots, as we create embodied entities that are "technologically alive": autonomous, self-organizing, capable of modifying its behavior in response to contingent stimuli, capable of learning new behaviors, communicative in physical gesture and language, and increasingly social. In other words, I think we are creating something that has never existed before. These are exciting times.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0905289. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Dautenhahn, K. (2003). Roles and functions of robots in human society: Implications from research in autism therapy. *Robotica, 21*(4), 443-452.

Kahn, P. H., Jr., Friedman, B., Perez-Granados, D. R., & Freier, N. G. (2006). Robotic pets in the lives of preschool children. *Interaction Studies: Social Behavior and Communication in Biological and Artificial Systems*, *7*, 405-436.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Freier, N. G., Severson, R. L., Gill, B. T., Ruckert, J. H., & Shen, S. (2011). "Robovie, You'll Have to Go Into the Closet Now": Children's Social and Moral Relationships With a Humanoid Robot. Manuscript submitted for publication.

Piaget, J. (1969). *The moral judgment of the child*. Glencoe, IL: Free Press. (Original work published 1932)

Scassellati, B. (2005a, October 12-15). *How social robots will help us to diagnose, treat, and understand autism.* Paper presented at the 12th International Symposium of Robotics Research, San Francisco, CA.

Scassellati, B. (2005b, August 13-15). *Quantitative metrics of social response for autism diagnosis.* Paper presented at the 14th IEEE International Workshop on Robot and Human Interactive Communication, Nashville, TN.

Stanton, C. M., Kahn, P. H., Jr., Severson, R. L., Ruckert, J. H & Gill, B. T. (2008). Robotic animals might aid in the social development of children with autism. *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction 2008* (pp. 97-104). New York, NY: Association for Computing Machinery.

Social Stories for Autistic Children Told by the Huggable Robot Probo

Bram Vanderborght[1], Ramona Simut[2], Jelle Saldien[1,3], Cristina Pop[2],
Alina S. Rusu[2], Sebastian Pintea[2], Dirk Lefeber[1], Daniel O. David[2,4]

[1]Vrije Universiteit Brussel, Robotics & Multibody Mechanics Research Group, Brussels, Belgium
[2] Babes-Bolyai University; Department of Clinical Psychology and Psychotherapy, Cluj-Napoca, Romania
[3]Howest University College, Industrial Design Center, Kortrijk, Belgium
[4]Mount Sinai School of Medicine, USA
bram.vanderborght@vub.ac.be
http://probo.vub.ac.be

***Abstract*— Children with autism have difficulties with social interaction and prefer interaction with objects, such as computers and robots rather than with humans. Probo is a social robot platform with an expressive head to study human-robot interaction. The social robot is in this study used for Robot Assisted Therapy (RAT) where the robot tells Social Stories to autistic children. Social Story Telling (SST) is a known therapy for autism in book form and is now for the first time tested by a scoial robot able to express different emotions by its 20 motors in the head. This paper discusses the technological innovations required to make a useful therapeutic tool for social story telling. Preliminary tests with Probo have shown that SST with Probo is more effective than with a human reader.**

I. INTRODUCTION

Children with Autism spectrum disorders (ASD) experience difficulties to engage in social interaction, and therefore lack learning opportunities in their classrooms and daily lives. Different therapies exist to attempt to lessen the deficits and family distress. For now there is no standard treatment, because the treatment has to be adapted to the specific needs of each child. To improve the efficacy of the Cognitive Behavioral Therapy (CBT) different tools are investigated to assist the therapist like music or Animal Assisted Therapy (AAT). Many children with ASD show an affinity for computers [1] [2]. Due to the recent advances in personal robots, the technology is becoming in the reach to be used as Robot Assisted Therapy (RAT) with different advantages over AAT. Emerging research shows, that autistic children proactively approach robots [3], that robots can act as a mediator between the child and the therapist [3], that robots can be used for play therapy [4] and to elicit joint attention episodes between a child and an adult [5]. Different forms of robots have been used for autism therapies from more hobby-toy style robots like GIPY-1 [3], SuperG, Cari, Robus-T, DiskCat [6], cartoon like robots like Keepon [7], mobile wheeled robots like Labo-1 [8], Pekee [9], IROMEC robot system [10]; robotic animals like Aibo [11], robotic dolls like Robota [12], and humanoid robots like Kaspar [13], HOAP-3 [14]. The use of robots also allow to provide new understanding of how human's higher cognitive functions develop by means of a synthetic approach, the so called Cognitive Developmental Robotics (CDR) [15].

The ability to express emotions is essential in communication and is a focus in different therapies for autistic children, also in social stories as developed by Carol Gray in 1991 [16]. Social Stories are short stories written or tailored to autistic individuals to help them understand and behave appropriately in social situations. Comic strip-like conversations using stick-figure and cartoon like emotions illustrate an ongoing communication and help them improve their understanding and comprehension of conversation [17]. The innovation is that for this research we use the social robot Probo (see Fig. 1 [18]) as Story Telling agent. Probo is a social robot that is developed as research platform to study cognitive human-robot interaction (cHRI) with a special focus on children. The robot Probo is designed to act as a social interface, providing a natural interaction while employing human-like social cues and communication modalities. The robot has a fully actuated head, with 20 degrees of freedom, capable of showing facial expressions and emotions. A remarkable feature is the moving trunk and the soft and huggable jacket. A user friendly Robotic User Interface (RUI) enables the operator to control the robot.

This paper discusses the technological innovations required to make a useful therapeutic tool for social story telling both on hardware (section II) as software (section III) and provides the first experimental results (section IV).

II. PROBO: HARDWARE

A. Appearance

Probo has a caricatured-zoomorphic morphology, so it does not look like a human, but has the abilities to communicate the same social cues as a human. Because it has no resemblances with existing animals or humans, it avoids the Mori's theory of the "uncanny valley" [19] and position the robot on the left hand side of the diagram. The robot has a fully actuated head with 20 DOF capable of showing facial expressions and making eye-contact [18]. In contrast with other robotic heads, a special body part, namely the trunk, is added to possibly intensify certain emotional expressions and to increase interactivity. The trunk makes the robot also easy to recognize for the children. The colour of the robot is green, this colour evokes mainly positive emotions such as relaxation and comfort [20].

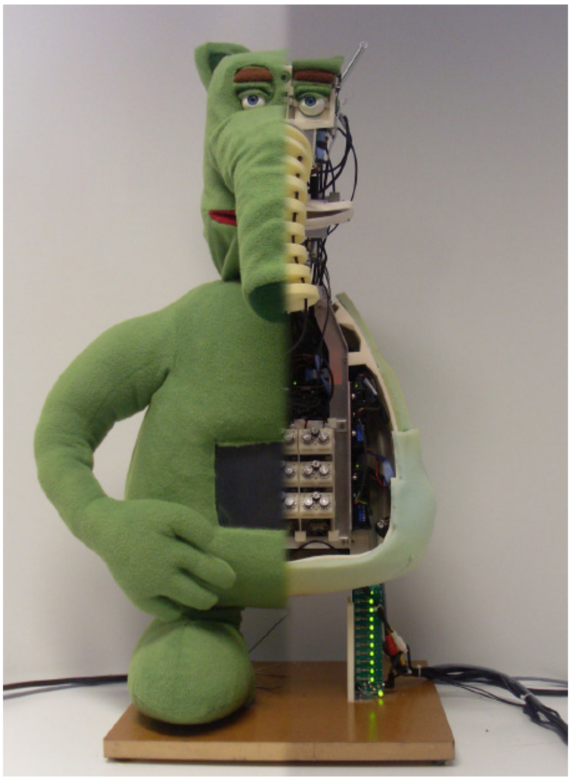


Fig. 1. Outer and inner appearance of the Huggable Robot Probo.

In evaluation studies shown by Saldien et al. [18], the facial expression of Probo are compared with other robots such as EDDIE [21], Kismet [22], Aryan [23] and Feelix [24] which did not cover their hardware, resulting in a mechanical look and the absence of a skin. The presence of the skin makes it easier to extract the facial features and explains the better scores. An overall recognition rate of 84% was achieved.

B. *Safety and huggable aspect*

Most robots have a mechanical look or are covered with plastic or metallic shells. Their actuators are stiff which gives them not only an unnatural look, but also an unnatural touch. The goal of the huggable robot Probo is to serve as robotic research platform for Robot Assisted Therapy with autistic children. Since not only cognitive interaction, but also physical interaction is targeted a new mechatronic design must be developed. During the experiments we noticed and encouraged the children to physically touch and hug the robot, therefore safety is of primary importance in the design. Possible approaches to obtain safety are: well-thought designs of light and flexible structures, the use of compliant actuators, appropriate material choices, and fail safe designs. This approach has been followed during the development of Probo. At the same time aspects as soft and flexible materials in combination with compliant actuators, contribute to the huggable and soft behavior of Probo. Traditional actuators as electrical drives with gearboxes are unsuitable to use in Probo because they are stiff, giving an unsafe behavior and an unnatural hard touch. Two different compliant actuators are developed to cope with this problem, on the one side by use of novel passive compliant actuators, Compliant Bowden Cable Driven Actuators (CBCDA), and on the other side by combining custom made servo motors, Non Back Drivable Servo (NBDS), with flexible components and materials such as springs, silicon and foam [25]. In both actuators the flexible element plays an essential role since it decouples the inertia of the colliding link with the rest of the robot, reducing the potential damage during impact [26].

In order to obtain Probo's final shape and appearance, the internal robotic hardware is covered. The covering exists of different layers. Hard ABS covers shield and protect Probo's internal hardware, shielding the internals. These covers are fixated to the head- and body-frames at the different points. The covers roughly define the shape of Probo. These are manufactured by use of rapid prototyping techniques, which means that the parts are build layer by layer. These covers are encapsulated in a PUR foam layer, that is covered with a removable fur-jacket. The use of the soft actuation principle together with well-thought designs concerning the robot's filling and huggable fur, are both essential to create Probo's soft touch feeling and ensure safe interaction. An overview of the mechanical design of Probo can be found in [27].

III. Probo: software

Since the goal is to perform interaction tests with children it was not our goal to make a fully autonomous system. Due to the current limited state of the art, we have chosen for a shared control with a human operator to allow for a higher level of social interaction. For the story telling the robot was used in a Wizard of Oz setup [28], [29]. A human operator simulates the system's intelligence and interacts with the user through Probo, without the actual need to implement all this higher level intelligence in the robot.

The software architecture of Probo is defined as a modular structure grouping all the systems in a clear control center (see Fig. 3). Black parts are already developed, red parts are under development and still need to be fully implemented in the robot. This system consists of four important blocks: the Perceptual-System, the Cognitive Control-System, the Expressional-system and the Motor-System. All those systems represented in Fig. 2 are made visible in an understandable way to the operator over the Robot Control Center (RCC). This control center has a virtual 3D model of Probo, that simulates all the movements of the robot and provides a visual feedback of the robot during control. The systems are made intuitive that make it possible for non-engineers to work with the robot within their domain of expertise, without the need of full understanding the underlying layers.

A. *Perceptual-System*

In order to interact in a social way with humans, Probo senses its environment in a human way. Since Probo's focuss lies on non-verbal face-to-face communication and since it wants to enable physical interaction by its huggable appearance, Probo is equipped with visual, auditory and tactile perception systems. Following the modular design strategy, the sensors, which are required to sense Probo's environment and physical interactions, are stand alone systems. Audio and Vision are quite commonly found in robots, while touch is rather new sensory input in robotics.

B. *Cognitive Control-System*

This part consists of different subparts:

- In the **Emotion-System** the emotions are parametrized. In our model two dimensions are used: valence and arousal to construct an emotion space. The emotion space is based on the circomplex model of Russel [30], see Fig. 3. Each emotion can be represented as a vector with the origin of the coordinate system as initial point and the corresponding arousal-valence values as the terminal point. The direction of each vector defines the specific emotion whereas the magnitude defines the intensity of the emotion.
- The **Homeostatic-System** is a system that makes Probo react emotionally with a model based on actions and needs, but is not used during this therapy.
- The **Behavior-System** is currently under development to improve the global social interaction level of the robot and also not needed for this therapy.
- The **Attention-System** provides the robot with a locus of attention using eg face detection. Also manual gazing can be used (eg when is decided to look only to the child when the story is told).

C. *Expressional-system*

Besides those autonomous systems of the cognitive controller, different modules are foreseen to steer certain groups of motors for specific tasks. The Animation module is provided that allows the user to assemble and manage sequences of motions. Animations are for example that the robot sleeps or nods with his head yes and no. In this module the different stories are preprogrammed with the appropriate animations as requested by the story. The module can also import lipsync animations which are generated from a given audio file. This creates a more lifelike talking creature. The Gaze module makes that the eyes and neck looks to a certain point of attention as defined by the Attention System. Each basic emotion correspondents with a certain position of the motors to express the facial expressions as studied in the user tests [18] and implemented in the Facial Expressions Unit. Prerecorded audio files are replayed for the social stories.

D. *Motor-System*

To make that all the motions create an illusion of life, a combination engine (Motion Mixer) and filter system is developed. Those systems take care that all the motions coming from different systems are properly mixed and that the output is one smooth and natural movement such as the motions seen in humans and animals and not like the very abrupt motions that we are expecting from robots.

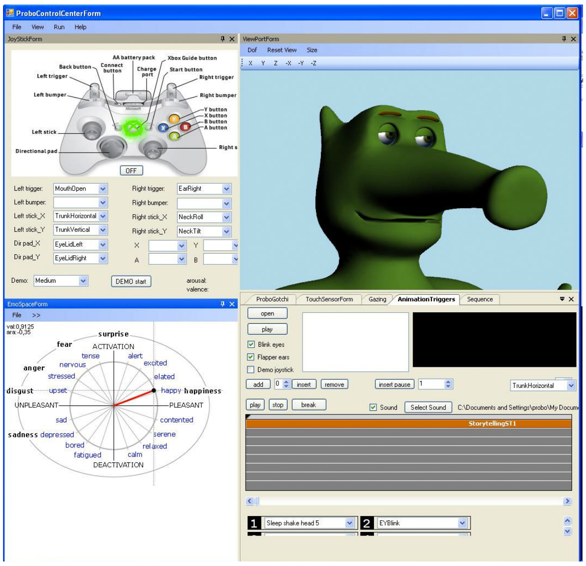


Fig. 2. Graphical User Interface of Robot Control Center (RCC)

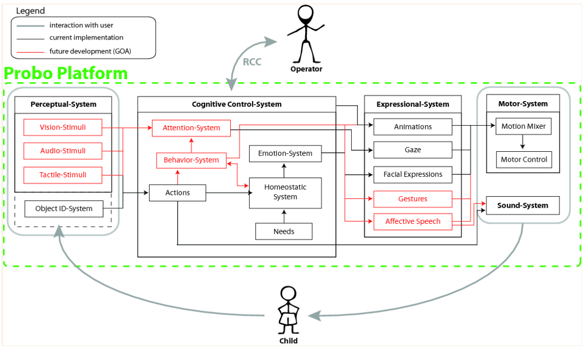


Fig. 3. Software Architecture of Probo

IV. Probo as Story Telling Agent

A. *Procedure*

Two boys (Nicu and Mihnea) and two girls (Antonia and Georgia) participated in the study, aged 4 to 9. The children were diagnosed according to the Diagnostic and Statistical Manual of Disorders IV criteria of autism. All parents were formally informed and agreed to the participation of their children in this study. The study took place at the therapy centre for children with ASD. The experiments were conducted in the same therapy room (about $20m^2$), see Fig. 4) in which the children usually participated in all the therapeutic programs. The robot operators were sitting in another room (see Fig. 5) to control the robot. By the camera of the robot the operator could see what was happening in the therapy room and he could also hear the communication to make the appropriate actions of the robot. Each child had only one intervention session a day. The language during interaction by the therapist and robot was Romanian (the mother's language of the children).

For each of the four participants, the main experimenter and the child's therapist created an individualized Social

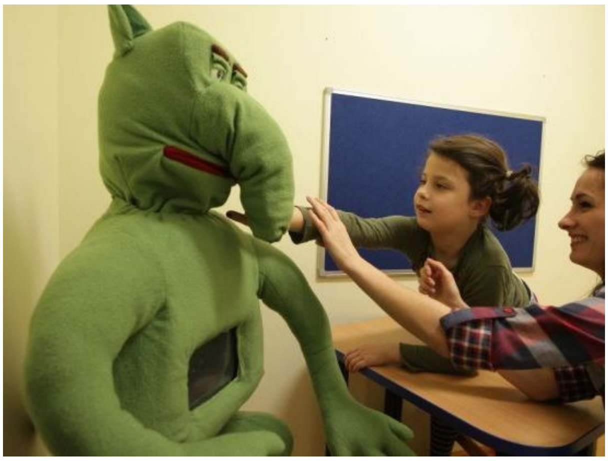


Fig. 4. Safe and huggable design of Probo allows for both cognitive as physical interaction.

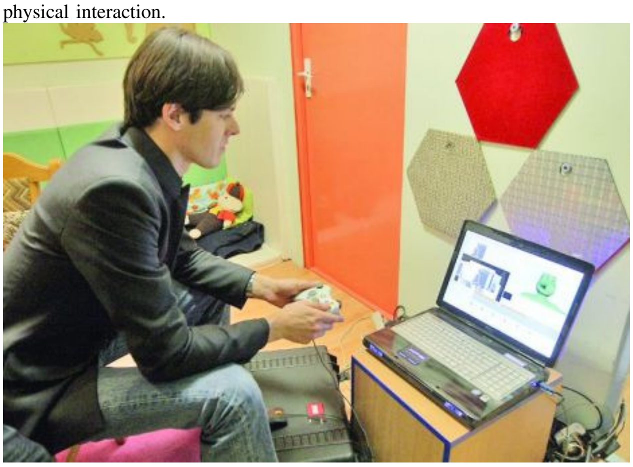
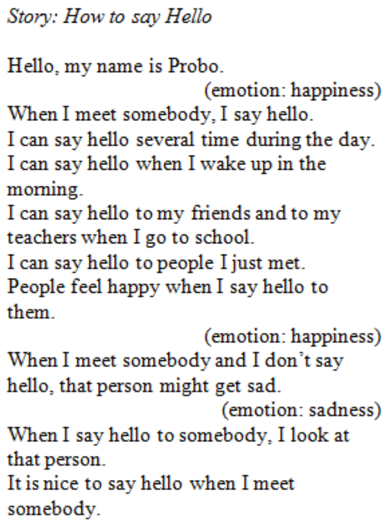


Fig. 5. The robot is used in a Wizard of Oz method, with the operator sitting in another room and using a gamepad to control the robot.



Story: How to say Hello

Hello, my name is Probo.
(emotion: happiness)
When I meet somebody, I say hello.
I can say hello several time during the day.
I can say hello when I wake up in the morning.
I can say hello to my friends and to my teachers when I go to school.
I can say hello to people I just met.
People feel happy when I say hello to them.
(emotion: happiness)
When I meet somebody and I don’t say hello, that person might get sad.
(emotion: sadness)
When I say hello to somebody, I look at that person.
It is nice to say hello when I meet somebody.

Fig. 6. Social Story designed for Mihnea to say hello when he enters in a room where someone is present based on Gray’s Social Story construction guidelines [16].

Story. The stories were developed by using Gray’s Social Story construction guidelines [16]. For each participant, a specific social skill deficit was identified. Also, the contextual factors that contribute and/or maintain that deficit were assessed, and the reinforcers for the maintenance and the generalisation of the specific social skill were selected. The Social Stories for Antonia and Nicu were designed to teach them how to share the toys when they are playing together with other children and for Georgia to teach her how to say thank you when someone gives his help to her or shares something with her, and finally for Mihnea to say hello when he enters in a room where someone is present (see Fig. 6). The stories were read by the experimenter in the Social stories phase or told by Probo in the Robotherapy and Social Stories phase of the study.

To test whether the two strategies used in increasing social abilities are effective we used an ABAC/ACAB counterbalanced design. When a stable value of the level of prompt was reached, the next phase was started. For every child six robo-sessions were foreseen. In the baseline phase (A), each child was observed during social interactions that required using the specific social skills aimed to be improved by the social story intervention (B) or by the Social Story and Robotherapy intervention (C). For each participant, the social story was offered immediately prior the observation period. After several sessions of Phase B/ Phase C, the intervention was withdrawn to the baseline conditions and the participants were observed without receiving any intervention. Before the robotherapy (phase C) an habituation phase with the robot was conducted to let the child interact as much as he/she needs it with the robot. For example the robot was sleeping and the child had to wake him up or the robot was sad and the child had to hug him to make the robot happy. The children could also ask questions to the robot which the robot answered by nodding yes or no. The session of the intervention itself lasted approximately 15 minutes. The Social Story reading was followed by a comprehensive check with the purpose of assessing participants’ understanding of the story and had to be answered with 100% accuracy. After the moment each story the child had to exercise the social ability that was targeted in the story. For example if the ability targeted was to say hello, the child was brought to a room where someone is present. All experiments were video-recorded and three experienced persons, trained by the experimenter performed the analyses of video sequences from all the phases of the study. The dependent measurements for each participant were: (1) level of prompt needed to provide the expected social response and (2) appropriate social interaction on the target social skill. The level of prompt needed to provide the expected social response was assessed using a 6-point scale rating the amount of prompting necessary for a successful social interaction. 6 meant did not respond to any prompt (gestural, physical or verbal) and a value of 0 was given if the participant independently and spontaneously engaged in an appropriate social interaction, without the need of any type of prompt.

B. Data analysis

The dependent variable (i.e. performance) was the level of prompt needed for the children to perform the requested social action at the end of each intervention session. The level of performance for the different intervention phases are presented as box plots in Fig. 7. Pay attention that we used an ABAC/ACAB counterbalanced design, but data are represented in order ABAC. The following conclusions can be drawn. The introduction of the two approaches of Social Story (with and without Probo) was associated with a decrease in the level of prompt. However, the effect of Probo was more effective than without using the robot since the average prompt in Phase C is lower than in Phase B. In the case with the robot in 40% of the interventions no need of prompt was necessary and a spontaneous engagement in the appropriate social interaction was obtained while this was only 13% in Phase B. A more detailed analysis can be found in [31].

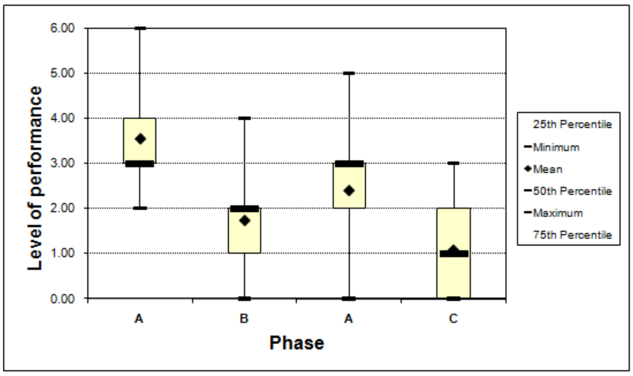


Fig. 7. Box plot of the data collected for the four children during the baseline and intervention phases. The y axis indicates the values of the level of performance from 6 to 0, where a value of 6 was given when the participant needed all types of prompt (gestural, physical and verbal) to perform the target behavior, and a value of 0 was given when the participant did not need any prompt (gestural, physical or verbal) to perform the target behavior. The x axis indicates the intervention phases (A = baseline phase; B = Social Story intervention phase; C = Social Story plus Probo intervention phase).

V. Conclusion

Probo has been designed as a research platform to perform human-robot interaction studies and develop robot assisted therapies with as special target group children. The appearance of the robot has been optimized for this task. Since besides cognitive interaction also physical interaction is targeted, a lot of attention in the design process is put to make the robot safe and give it a soft and huggable feeling. The use of compliant actuators and a soft skin is essential to achieve this goal. The robot was used in a Wizard of Oz setup and the software was optimized for this task. The Robot Control Center makes abstraction of the control of the robot so it is possible to control the robot without the need of full understanding the underlying layers. The GUI has a virtual 3D model of the robot so the robot can be controlled in an intuitive way using a gamepad and mouse. The first experiments indicate that a Social Story told with Probo leads to a decrease in the level of prompt necessary to obtain an appropriate social interaction by the autistic child, than when the story is told by a human.

Future work includes performing more therapy sessions with the robot. We also aim to construct a novel version of Probo with actuated arms to extend the range of stories with eg learning to give hugs, pointing to objects,... At the moment the Social Stories are replayed on the robot since the complete story is prerecorded; goal is also to make interactive stories besides an interactive habituation phase where the robot is able to respond to the actions of a child.

VI. Acknowledgments

The authors thank the Asociatia Autism Transilvania for the hospitality to perform the experiments.

References

[1] M. Hart, “Autism/excel study,” *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility*, pp. 136–141, 2005.

[2] A. Tartaro and J. Cassell, “Authorable virtual peers for autism spectrum disorders,” *Proceedings of the Combined workshop on Language-Enabled Educational Technology and Development and Evaluation for Robust Spoken Dialogue Systems at the 17th European Conference on Artificial Intellegence*, 2006.

[3] K. Dautenhahn, I. Werry, J. Rae, P. Dickerson, P. Stribling, and B. Ogden, “Robotic Playmates: Analysing Interactive Competencies of Children with Autism Playing with a Mobile Robot.,” *Multiagent Systems, Artificial Societies, and Simulated Organizations*, vol. 3, 2002.

[4] D. Francois, S. Powell, and K. Dautenhahn, “A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children’s proactivity and initiative-taking,” *Interaction Studies*, vol. 10, no. 3, pp. 324–373, 2009.

[5] B. Robins, P. Dickerson, P. Stribling, and K. Dautenhahn, “Robot-mediated joint attention in children with autism: A case study in robot-human interaction,” *Interaction Studies*, vol. 5, no. 2, pp. 161–198, 2004.

[6] F. Michaud, P. Lepage, J. Leroux, M. Clarke, F. Belanger, Y. Brosseau, and D. Neveu, “Mobile robotic toys for autistic children,” *Proc. International Symposium on Robotics, Montral*, 2000.

[7] H. Kozima, C. Nakagawa, and Y. Yasuda, “Interactive robots for communication-care: A case-study in autism therapy,” in *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*, pp. 341–346, IEEE, 2005.

[8] K. Dautenhahn and I. Werry, “Towards interactive robots in autism therapy: Background, motivation and challenges,” *Pragmatics and Cognition*, vol. 12, no. 1, pp. 1–35, 2004.

[9] K. Dautenhahn, I. Werry, T. Salter, and R. Boekhorst, “Towards adaptive autonomous robots in autism therapy: Varieties of interactions,” in *Computational Intelligence in Robotics and Automation, 2003. Proceedings. 2003 IEEE International Symposium on*, vol. 2, pp. 577–582, IEEE, 2003.

[10] E. Ferrari, B. Robins, and K. Dautenhahn, “Therapeutic and educational objectives in robot assisted play for children with autism,” in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, pp. 108–114, IEEE, 2009.

[11] C. Stanton, P. Kahn Jr, R. Severson, J. Ruckert, and B. Gill, “Robotic animals might aid in the social development of children with autism,” in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, pp. 271–278, ACM, 2008.

[12] A. Billard, “Robota: Clever toy and educational tool,” *Robotics and Autonomous Systems*, vol. 42, no. 3-4, pp. 259–269, 2003.

[13] B. Robins, F. Amirabdollahian, Z. Ji, and K. Dautenhahn, “Tactile interaction with a humanoid robot for children with autism: A case study analysis involving user requirements and results of an initial implementation,” in *RO-MAN, 2010 IEEE*, pp. 704–711, IEEE, 2010.

[14] P. Ravindra, S. De Silva, K. Tadano, A. Saito, S. Lambacher, and M. Higashi, "Therapeutic-assisted robot for children with autism," in *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pp. 3561–3567, IEEE, 2009.

[15] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 1, pp. 12–34, 2009.

[16] C. Gray, *The new social story book*. Future Horizons Inc, 2000.

[17] C. Gray and M. Jenison High School (Jenison, *Comic Strip Conversations: Colorful, illustrated interactions with students with autism and related disorders*. Jenison Public Schools, 1994.

[18] J. Saldien, K. Goris, B. Vanderborght, J. Vanderfaeilli, and D. Lefeber, "Expressing emotions with the huggable robot probo," *International Journal of Social Robotics, Special Issue on Social Acceptance in HRI*, vol. 2, no. 4, pp. 377–389, 2010.

[19] M. Mori, *The Buddha in the Robot*. Tuttle Publishing, 1981.

[20] N. Kaya and H. Epps, "Relationship between Color and Emotion: A Study of College Students.," *College Student Journal*, vol. 38, no. 3, pp. 396–406, 2004.

[21] S. Sosnowski, A. Bittermann, K. Kuhnlenz, and M. Buss, "Design and Evaluation of Emotion-Display EDDIE," *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pp. 3113–3118, 2006.

[22] C. Breazeal, *Designing Sociable Robots*. Mit Pr, 2002.

[23] H. Mobahi, *Building an interactive robot from scratch*. Bachelor Thesis, 2003.

[24] L. Canamero and J. Fredslund, "How does it feel? emotional interaction with a humanoid lego robot," *Socially Intelligent Agents: The Human in the Loop. Papers from the AAAI 2000 Fall Symposium*, pp. 23–28, 2000.

[25] K. Goris, J. Saldien, B. Vanderborght, and D. Lefeber, "How to achieve the huggable behavior of the social robot probo? a reflection on the actuators," *Mechatronics*, vol. 21, pp. 490–500, April 2011.

[26] M. Van Damme, B. Vanderborght, B. Verrelst, R. Van Ham, F. Daerden, and D. Lefeber, "Proxy-based sliding mode control of a planar pneumatic manipulator," *International Journal of Robotics Research*, vol. 28, no. 2, pp. 266–284, 2009.

[27] K. Goris, J. Saldien, B. Vanderborght, and D. Lefeber, "Mechanical design of the huggable robot probo," *International Journal of Humanoid Robotics*, p. accepted for publication, 2011.

[28] T. Landauer, "Psychology as a mother of invention," *ACM SIGCHI Bulletin*, vol. 17, no. SI, pp. 333–335, 1986.

[29] J. Wilson and D. Rosenberg, "Rapid prototyping for user interface design," 1988.

[30] J. Posner, J. Russell, and B. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 03, pp. 715–734, 2005.

[31] B. Vanderborght, R. Simut, J. Saldien, C. Pop, A. S. Rusu, S. Pintea, D. Lefeber, and D. O. David, "Using the social robot probo as social story telling agent for children with asd," *Interaction Studies*, under review.

Computational Audiovisual Scene Analysis for Dialog Scenarios

Rujiao Yan[1,2], Tobias Rodemann[2] and Britta Wrede[1]

Abstract— **We introduce a system for Computational Audio-Visual Scene Analysis (CAVSA) with a focus on human-robot dialogs in multi-person environments. The general target of CAVSA is to learn who is speaking now, where the speaker is, and whether the speaker is talking to the robot or to other persons. In the application specified in this paper, we aim at estimating the number and position of speakers using several auditory and visual cues. Our test application for CAVSA is the online adaptation of audio-motor maps, where vision is used to provide position information about the speaker. The system can perform this adaptation during the normal operation of the robot, like when the robot is engaging in conversation with a group of humans. Comparing our online adaptation of audio-motor maps using CAVSA with prior online adaptation methods, our approach is more robust in situations with more than one speaker and when speakers dynamically enter and leave the scene.**

I. Introduction

In most robotics scenarios, a robot is usually interacting with multiple people. Thus it should be able to learn who is speaking now, where the speaker is, and whether the speaker is talking to the robot or to other persons. Computational Audiovisual Scene Analysis (CAVSA) is aimed at fulfilling these tasks. CAVSA plays an important role in human-robot interaction, for instance it enables the robot to better understand dialog situations, improves speech recognition by assigning words to speakers, and relates visual and auditory features of a speaker. To evaluate the performance of CAVSA we employ it for the online adaptation of audio-motor maps. This task depends strongly on a correct scene representation and the performance can be measured by comparison to audio-motor maps calibrated in standard offline procedures.

In robotics, many sound localization systems use audio-motor maps, which describe the relationship between binaural cues and sound position in motor coordinates (azimuth and elevation). The main binaural cues are interaural time difference (ITD) and interaural intensity difference (IID) [1]. Using audio-motor maps one can compute the sound source position from measured audio cues. We concentrate on audio-motor maps for azimuth to ease the description of our algorithm, but the approach can be expanded to elevation. Audio-motor maps can be calibrated offline by measuring audio cues for several known positions [2]. However, audio-motor maps can change and need to be relearned whenever any relevant part of the robot or the environment was modified, for example, microphone type, microphone position, robot head and room. Additionally, it is difficult to estimate the quality of the current maps. Hence a continuous online adaptation has been considered during the normal operation of the robot [3]. It is known that humans continuously adapt the audio-motor maps to their current auditory periphery, while the dimensions of the head and external ears are growing from birth to adulthood. Even adults have sound localization plasticity, for instance when molds are placed into the external ears to alter audio-motor maps [4]. Rodemann et al. [3] suggested a purely auditory online adaptation approach, where audio provides position information of limited precision.

Vision plays an important role in calibration of audio-motor maps in humans and animals [5]–[7]. Thus vision has been used as the feedback signal for higher precision in adaptation of audio-motor maps in robotics, as per [8], [9]. However, the approach in [8] fails when more than one person or the wrong speaker appears in the camera image. Nakashima et al. [9] proposed another approach using visual feedback in a simplified environment, where a red marker was attached to the sound source and no other red object exists. These methods employ heuristics for linking visual and auditory information and can only work in limited environments. Besides, both methods need extra head motions to search for the visual marker. In comparison to state of the art, our CAVSA method is used to find the correct visual correspondence of the current sound source, and aims at enabling online adaptation to run in more complex environments. If CAVSA selects an unrelated visual signal for the adaptation, the quality of audio-motor maps may deteriorate. Given precise measurements of visual position and audio cues, the quality of maps depends on the performance of CAVSA. This is the reason why we test CAVSA in online adaptation of audio-motor maps. Our system does not require specific motions of the robot, so that audio-motor maps can be continuously online adapted during the normal operation of the robot.

For CAVSA the scene is represented with auditory and visual features using the concept of proto-objects. Proto-objects can combine an arbitrary number of features in a compressed form. For more information about proto-objects see [10], [11]. The visual and audio proto-objects for the same speaker are then integrated based on position information.

A. Comparison to related work

Currently there is a broad range of applications using audiovisual integration in multi-person environments. The first application is speaker recognition (see e.g. [12]), where

[1] Research Institute for Cognition and Robotics (CoR-Lab), Bielefeld University, 33594 Bielefeld, Germany {ryan,bwrede}@cor-lab.uni-bielefeld.de

[2] Honda Research Institute Europe GmbH, Carl-Legien-Str. 30, 63073 Offenbach, Germany tobias.rodemann@honda-ri.de

an audiovisual database consisting of all speakers is required for training. The second application is audiovisual multi-person tracking. Most methods use only sound position as an auditory feature and thus fail when a speaker leaves the scenario for a while and reappears or the speaker moves while not talking [13], [14]. In this case CAVSA could flexibly add more auditory and visual features to identify the speaker. Multi-person tracking can also be implemented in a smart-room environment [15], where many auditory and visual sensors are installed. The third group is audiovisual speaker diarization systems [16], which can index who spoke when in a video file. The training of diarization is normally an offline process, where the data can only be processed after complete recording. Another application searches for the visual part of the current speaker in a video using synchrony between lip motion and speech [17]. The approach requires that the lip of the current speaker is always in the field of view. In comparison to these methods, our CAVSA approach can combine many low or mid level auditory and visual features to achieve a high performance of audiovisual integration. It runs in an unsupervised, real-time, online and incremental manner. Besides, we use only a humanoid robot head with a pair of cameras and a pair of microphones. In this work just the left camera is employed to capture the visual signal. The head is mounted on a pan-tilt unit.

Additionally, while current approaches in Computational Auditory Scene Analysis (CASA) mostly deal with parallel sources using microphone arrays (see e.g. [18]), we focus on purely sequential sounds. Our system could be used after sound source separation in case of concurrently active sources.

II. CAVSA

In this section we introduce the concept of Computational AudioVisual Scene Analysis (CAVSA). In CAVSA the scene is represented with audio and visual proto-objects. Proto-objects for the same speaker are grouped together in auditory and visual Short-Term Memories (STM) respectively. Visual and audio proto-objects are then matched based on position. Fig. 1 schematically illustrates the system architecture of CAVSA. Proto-objects, STM and audiovisual association are described below.

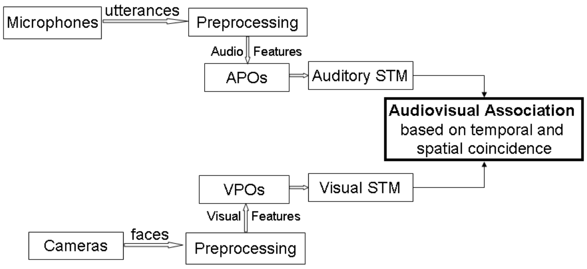


Fig. 1. System architecture of CAVSA. APO: audio proto-object, VPO: visual proto-object.

A. *Proto-objects*

Proto-objects are a psychophysical concept and are considered here as a compressed form of various features. Proto-objects can be tracked, pointed or referred to without identification and enable a flexible interface to behavior-control in robotics. For more information about proto-objects see [10], [11].

1) Visual proto-objects: In the camera field of view the visual objects can be segmented based on e.g. the similarity of color or shape to a given reference model. A frontal face detection algorithm based on [19] is used to extract visual proto-objects in multi-person scenarios. We assume that the person talking to robot is most of the time looking at the robot face. For each of these proto-objects, the center of the segment in the camera image is computed. The distance between robot head and speakers is about 1m. Using the distance information and saccade maps (see [20]), we can calculate the face position in 3D world coordinates and motor coordinates, and store them in the visual proto-object. Actually, proto-objects could contain an arbitrary number of features. Depending on the tasks, other features such as object size, orientation, color histogram and texture could also be used.

2) Audio proto-objects: A Gammatone Filterbank (GFB) as a model of the human cochlea is employed in the auditory preprocessing [11]. The GFB has 100 frequency channels that span the range of 100 -11000 Hz. To form audio proto-objects, we first segment audio streams based on energy. An audio segment begins when the signal energy exceeds a threshold and ends when the energy falls below this threshold. We then derive start time, length and energy for an audio proto-object. A filtering of audio proto-objects based on segment length and energy is then performed, since short or low power auditory signals are very probably noise. In addition, an audio proto-object contains also population-coded position cues (IID and ITD) and the estimated position encoded as a population vector, which will be explained in section III.

B. *Short-term memory*

In short-term memory (STM), proto-objects for the same speaker are grouped together. When a new proto-object appears, the procedure of entering it into STM can be described as follows:

- If the STM is empty, the new proto-object is added to the STM.
- If the STM already contains one or more proto-objects, the distance or similarity of selected grouping features are computed between the new proto-object and all proto-objects in the STM. If the distance between the new proto-object and the closest proto-object in the STM is smaller than a threshold, these two proto-objects are merged (averaged). Otherwise the new proto-object is inserted into the STM.
- Proto-objects, which are not updated for more than a certain period (200 s in our experiment) are removed from the STM.

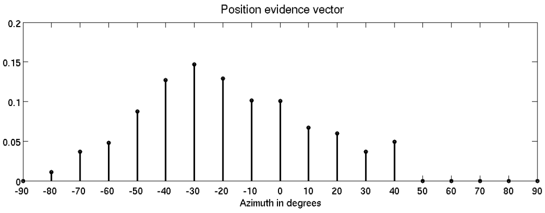


Fig. 2. An example of a position evidence vector, corresponding to output firing rates from a bank of neurons with different receptive fields. Here the estimated azimuth is $-30°$.

Using such a STM, it is not necessary to buffer all the incoming proto-objects for processing. Moreover, we can match an audio proto-object to a visual proto-object, even if it is out of sight for a while e.g. due to the movements of robot head.

In auditory STM, grouping features could be position and/or spectral energies. In visual STM one or more features among position, color and size could be employed to group visual proto-objects. In this work only position is used as a grouping feature for the auditory and visual STM. In an audio proto-object the position is represented by population code vector. For more information about this population coding see section III. The similarity of position vectors between the new audio proto-object and an audio proto-object in auditory STM is based on the scalar product of normalized position vectors (mean 0, norm 1). We set the threshold of the similarity to be 0.6 empirically. In visual STM the Euclidean distance of positions in 3D world coordinates is calculated and the threshold is set to 30 cm, so that slight movements of speakers such as head shaking are tolerated.

C. Audiovisual association

Position is also used to match a visual proto object and an audio proto-object from their STMs. Auditory position evidence vectors and visual positions in world coordinates must be converted to the same metric, for which motor coordinates (azimuth and elevation) are preferred. Speakers are about 1m away from the robot. In this paper we concentrate only on azimuth as mentioned. The azimuth of an audio proto-object is taken as the peak position in its position evidence vector, while the azimuth of a visual proto-object is estimated using saccade maps (see [20]). Fig. 2 shows an example of a position evidence vector.

The azimuth distance between audio proto-object A_i $(i \in [1,M])$ and visual proto-object V_j $(j \in [1,N])$ is denoted as $\Delta X(A_i,V_j)$, where M and N stand for the number of audio and visual proto-objects in auditory and visual STM respectively. The relative probability that audio proto-object A_i belongs to visual proto-object V_j can then be approximated as:

$$P_{common}(A_i,V_j) \approx \exp\left(\frac{-\Delta X(A_i,V_j)^2}{2\cdot\delta_{AV}^2}\right), \tag{1}$$

where the standard deviation δ_{AV} represents the average difference in estimated azimuth between an audio and a visual proto-object which are caused by the same speaker.

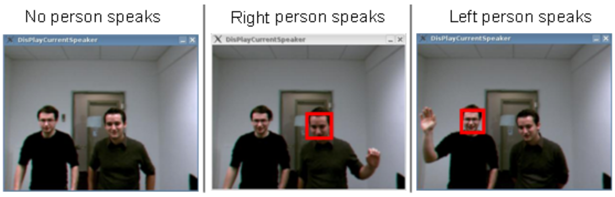


Fig. 3. Detection of the current speaker. Face detection of the current speaker is visualized. To easily evaluate the results, the current speaker was required to raise one hand.

Next, we check how certain the association between visual proto-object V_{jMax} with the maximal probability and A_i is. If one visual proto-object shows a very high probability and all other visual proto-objects have a low probability, this indicates a reliable association. Conversely, when all visual proto-objects have quasi equal probability, the association is unreliable. The uncertainty is computed using the entropy of the normalized probability. A similar usage of entropy can be found in speech recognition, as per [21]. All probabilities $P_{common}(A_i,V_j)$ for audio proto-object A_i are normalized such that they sum up to 1. The normalized probability is denoted as:

$$\widehat{P}_{common}(A_i,V_j) = \frac{P_{common}(A_i,V_j)}{\sum\limits_{j=1}^{N} P_{common}(A_i,V_j)}. \tag{2}$$

The entropy for A_i is given as:

$$H_i = \begin{cases} 0 & \text{if } N=1, \\ \dfrac{-\sum\limits_{j=1}^{N}\left(\widehat{P}_{common}(A_i,V_j)\cdot\log_2\widehat{P}_{common}(A_i,V_j)\right)}{\log_2 N} & \text{if } N>1. \end{cases} \tag{3}$$

Here, the division by $\log_2 N$ ensures that the maximal H_i is 1 to easily set a threshold Θ_H. If entropy H_i is larger than Θ_H, A_i and V_{jMax} are not associated. Θ_H was set to 0.8 empirically. The uncertainty of the whole audiovisual association can be captured by averaging over all H_i :

$$H = \frac{\sum\limits_{i=1}^{M} H_i}{M}. \tag{4}$$

Given learned audio-motor maps, Fig. 3 illustrates an example, where CAVSA is used in an online scenario to visualize the current active speaker among two speakers who stand in front of the robot.

III. Online adaptation

In this section we will describe how to find the matched visual position to the current sound and how to adapt audio-motor maps. In our system audio-motor maps represent the relation between population-coded cues and position evidence vectors. An audio-motor map M contains for each azimuth angle p (-90, -80, ..., 0, ..., 80, 90), each cue l ($l=1$ for IID, $l=2$ for ITD) and each frequency channel f $(1-100)$ a population code vector $M(p,l,f,n)$. Nodes (neurons) n have receptive field centers at $(-0.9,-0.8,...,0,...,0.8,0.9)$. We measure binaural cues in each frequency channel when

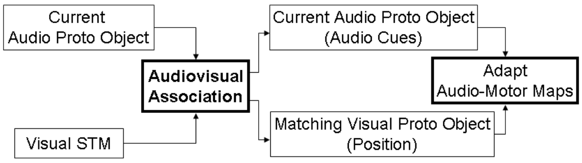


Fig. 4. System architecture of online adaptation using CAVSA

an onset appears, encode then the measured cues and store them in the audio proto-object. For encoding, the same set of neurons n is used and every measured cue IID or ITD leads to an activation in the nearest neurons, so that a population code vector is generated. In each frequency channel f, population code vectors for all measurements of cue l are then summed up in an audio proto-object. Finally, each summed population code vector is normalized to mean 0 and norm 1. Let us denote the encoded cue l in frequency channel f, at node n as $C(l,f,n)$. To acquire position evidence vector $E(p)$, population response $C(l,f,n)$ is compared with stored population responses $M(p,l,f,n)$ for all positions p by computing scalar products. The peak in position vector $E(p)$ is taken as the estimated sound source position. For more details see [11].

We use only the current audio proto-object instead of the auditory STM ($M = 1$), since only information about the current sound is required. The number of proto-objects in visual STM (N) is considered as the number of speakers in the dialog scenario. The system architecture of online adaptation using CAVSA is illustrated in Fig. 4.

The matched visual proto-object is searched for using equation (1), (2) and (3). Note that if $N = 1$, then $\widehat{P}_{common}(A,V) = 1$ and entropy $H = 0$. That means, if only one visual proto-object exists in the visual STM, the audio and visual proto-object are assumed to have a common cause. During the learning of the audio-motor maps, the standard deviation δ_{AV} in equation (1) is dynamically updated depending on the quality of the current audio-motor map. We approximate δ_{AV} by calculating the average difference between estimated position in audio and visual proto-objects over time using the following update rule:

$$\delta^t_{AV} = \begin{cases} \Delta X(A,V)\cdot w + \delta^{t-1}_{AV}\cdot(1-w) & \text{if } N = 1, \\ \delta^{t-1}_{AV} & \text{otherwise.} \end{cases} \tag{5}$$

Here, t and w represent update step and weight respectively. We set $w = 0.1\cdot\beta$ dependent on the fixed adaptation rate β, which controls the degree of adaptation for a single step. $\Delta X(A,V)$ describes the position distance between the audio and visual proto-object in the current adaptation step. δ^t_{AV} is updated only if just one visual proto-object exists. The initial value δ^0_{AV} is set to 40° empirically.

In the experiments it was found that a visual proto-object, which is not related to the current sound source but near the correct visual proto object, can also enhance the quality of audio-motor maps, particularly when the quality of maps is poor as during initialization. Thus if entropy H exceeds the threshold Θ_H, but the position distance between the visual proto objects with maximum and second maximum probability ($\widehat{P}_{common}$) is small, audio-motor maps can be updated nonetheless. The uncertainty of an adaptation step can be described by the following equation:

$$H' = H\cdot\Delta X(V_{jMax}, V_{jSecMax}), \tag{6}$$

where V_{jMax} and $V_{jSecMax}$ stand for the visual proto-objects with maximal and second maximal probability respectively. If uncertainty H' is below threshold $\Theta_{H'}$ or $H < \Theta_H$, a confidence factor c is set to 1 and the map is adapted. Otherwise $c = 0$ and the map is not updated in the current step. The threshold $\Theta_{H'}$ depends on the standard deviation δ_{AV} ($\Theta_{H'} = 2\cdot\delta_{AV}$), since the system has a high tolerance for the visual position difference when the quality of audio-motor maps is poor. The matched visual position p_v is then converted to a position evidence vector, which can be defined by a delta function δ_{p,p_v}.

The audio-motor map is updated by:

$$\begin{aligned} M(p,l,f,n,t) \;\; &= \;\; M(p,l,f,n,t-1) - F(p)\cdot \\ &\quad\;\; (M(p,l,f,n,t-1) - C(l,f,n)), \end{aligned} \tag{7}$$

where p, l, f and n stand for position, cue index, frequency channel and node, respectively. Learning parameter $F(p)$ is given by:

$$F(p) = c\cdot\beta\cdot\delta_{p,p_v}, \tag{8}$$

where c and β represent the confidence of the matching process and the fixed adaptation rate respectively. In our experiment $\beta = 0.2$.

IV. Results

Our approach was tested in real world scenarios. Offline-calibrated maps were used as reference. Our approach was compared with a heuristic method in scenarios where additional persons dynamically entered and vacated the room.

A. *Offline-calibrated audio-motor maps as reference*

In the experiment we firstly calibrated audio-motor maps offline und used them as reference for performance estimation. A loudspeaker was placed in front of the robot (0°), at a distance of 1m away and at the same height as the robot head. The head changed its orientation p_h every 10° from $-90°$ to 90°, so that the azimuth $(-p_h)$ changed correspondingly in robot-centered coordinates. At each position, 47 sound files were played and mean population responses of IID and ITD were measured. The whole offline calibration required more than 2 hours.

The performance of online-adapted audio-motor maps can be then estimated by comparison with offline-calibrated maps using normalized Euclidean distance:

$$d(M,M') = \sqrt{\frac{\sum\limits_{p}\sum\limits_{l}\sum\limits_{f}\sum\limits_{n}(M(p,l,f,n) - M'(p,l,f,n))^2}{K}}, \tag{9}$$

where M and M' represent online-adapted and offline-calibrated maps respectively. K is the total number of elements in an audio-motor map and satisfies $K = k_p\cdot k_l\cdot k_f\cdot k_n$, where $k_p = 19$, $k_l = 2$, $k_f = 100$ and $k_n = 19$ is the number of positions, cues, frequency channels and nodes, respectively.

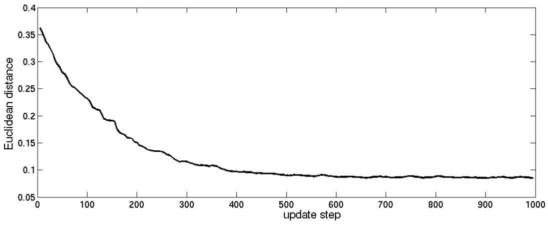


Fig. 5. Euclidean distance between offline-calibrated and online-adapted maps over time

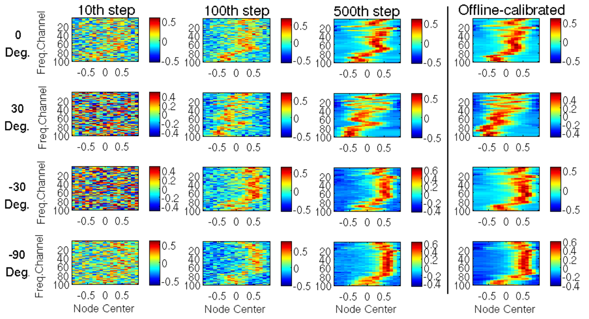


Fig. 6. Online-adapted IID maps for several azimuth angles and in different adaptation steps. Offline-calibrated maps are used as reference.

B. Basic online scenario

In the online scenario we simulated a speaker with a loudspeaker on which a picture of a face was attached. The loudspeaker was placed on the same position as in offline calibration. During online adaptation, the robot head oriented itself to a random horizontal angle in the range $[-90, 90]$ after an update step was finished. The acquisition of auditory and visual signals was interrupted during head movement, so that audio-motor maps were only adapted in still status. At the beginning maps are initialized with random numbers in the range $[-0.5, 0.5]$ using a uniform distribution. The normalized Euclidean distance over time between online-adapted and offline-calibrated maps is illustrated in Fig. 5. Fig. 6 shows online-adapted IID maps on several positions (0°, 30°, $-30°$ and $-90°$) and in different update steps (10th, 100th and 500th step), as well as offline-calibrated maps on the corresponding positions. Every 100 update steps of our approach need about 7 minutes. The plot in Fig. 5 shows that a good similarity is achieved after about 400 steps, which takes about 30 minutes, while offline calibration requires more than 2 hours.

C. Natural communication

Our approach was tested in three scenarios where additional persons ($N > 1$) dynamically entered and vacated the scene. The results were then compared with a heuristic method which considers the last seen face as the matched visual position to the current sound source. If more than one face appears in the camera image, the heuristic method randomly chooses one. The heuristic method is similar to methods in [8], [9] for linking auditory and visual information. The three scenarios and the corresponding results are described as follows.

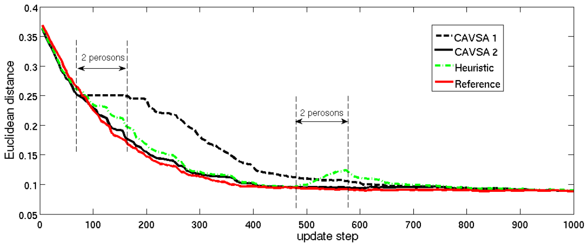


Fig. 7. Scenario 1: one additional person entered the room in the 70th adaptation step and vacated in the 170th adaptation step. He entered then in the 480th step and vacated in the 580th step again.

The difference between the first scenario and the basic online scenario in section IV-B is that an additional person entered the room during the online adaptation in the first scenario, stood 1m away, faced the robot for a while, did not speak and then vacated. After some update steps the person entered the room and vacated again in the same manner. The only sound source was the loudspeaker at 0°, since the additional person did not speak. For this scenario, CAVSA which computes the association uncertainty with H in equation (3), CAVSA with consideration of both H and H' in equation (6), the heuristic method and the method using known sound source position (0°) as reference are compared. To simplify the description let us denote these four methods as “CAVSA 1”, “CAVSA 2”, “Heuristic” and “Reference”, respectively. As shown in Fig. 7, the quality of audio-motor maps was still poor between the 70th and 170th adaptation step, when the additional person was in the room for the first time. “CAVSA 1“ did not update the maps due to the high entropy H in equation (3). ”Heuristic” selected sometimes the wrong position for adaptation, but improved the performance of audio-motor maps to some degree because the quality of the maps was still poor. “CAVSA 2“ nearly reached the performance of “Reference” which used true sound source position. Between the 480th and 580th step the maps were refined. “CAVSA 1“ and “CAVSA 2“ were almost not influenced by the additional person because of their good performance on audiovisual integration, while the error in “Heuristic” increased due to using wrong positions.

The only difference between the first and second scenario is that two additional persons instead of one dynamically entered the room. Fig. 8 shows the comparison of the four methods with Euclidean distance to offline-calibrated maps over time. In comparison to Fig. 7 the four methods performed similarly except that “Heuristic” got much worse when two additional persons appeared than in the first scenario when only one additional person appeared.

In the third scenario the loudspeaker was not used. Instead two speakers stood 1m away from the robot, faced the robot and talked to it alternatingly. After some steps one person vacated the room and only one spoke to the robot. The adaptation began with an audio-motor map which had been adapted for 80 steps. Fig. 9 illustrates the comparison

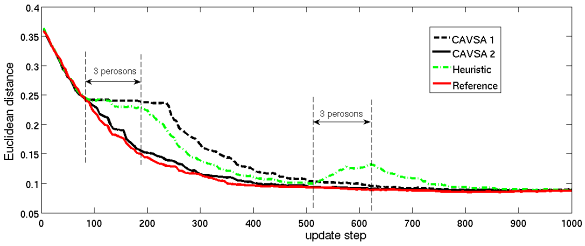


Fig. 8. Scenario 2: two additional persons entered the room in the 80th adaptation step and vacated in the 190th step. They entered then in the 515th step and vacated in the 620th step again.

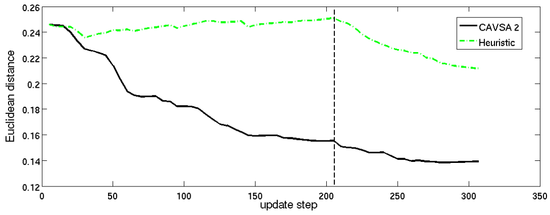


Fig. 9. Scenario 3: from start two speakers talked to the robot alternatingly. In the 205th step one person vacated the room and only one spoke to the robot.

of "CAVSA 2" and "Heuristic" up to the 310th adaptation step. It was shown that "CAVSA 2" performed much better than "Heuristic" when two speakers talked to the robot alternatingly.

The results in these three scenarios showed that the adaptation process with CAVSA was more robust in situations where additional persons dynamically entered and vacated the scene.

V. Summary and Outlook

We have suggested an approach for Computational AudioVisual Scene Analysis (CAVSA) with a focus on human-robot interaction in multi-person environments. In CAVSA the scene is represented with audio and visual proto-objects. Audio and visual Proto-objects for the same speaker are then grouped together in their STMs respectively. Finally, audio and visual proto-objects are matched based on position information. We have shown that our system can correctly determine the number and position of speakers in typically human-robot dialog scenarios. This was demonstrated by the online adaptation of audio-motor maps. Comparing our online adaptation of audio-motor maps using CAVSA with prior online adaptation methods, our approach is more robust in situations with more than one speaker and when speakers dynamically enter and leave the scene. Only spatial coincidence is so far used to group audio and visual proto-objects in their STMs, which fails for instance when a person moves quickly or several persons stand very close to each other. Hence we plan to employ more grouping features such as spectral energies for auditory STM and color or size for visual STM.

VI. Acknowledgements

Sincere thanks to Bram Bolder and Stefan Kirstein for much helpful support with the vision system. This work is supported by the Honda Research Institute Europe.

References

[1] D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*. IEEE Press/Wiley-Interscience, 2006.

[2] H. Finger, P. Ruvolo, S. Liu, and J. R. Movellan, "Approaches and databases for online calibration of binaural sound localization for robotic heads," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2010.

[3] T. Rodemann, K. Karova, F. Joublin, and C. Goerick, "Purely auditory online-adaptation of auditory-motor maps," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2007.

[4] M. M. V. Wanrooij and A. J. V. Opstal, "Relearning sound localization with a new ear," *Neuroscience*, vol. 25, pp. 5413–5424, June 2005.

[5] M. Zwiers, A. V. Opstal, and J. Cruysberg, "A spatial hearing deficit in early-blind humans," *Neuroscience*, vol. 21, pp. 1–5, 2001.

[6] E. I. Knudsen, "Instructed learning in the auditory localization pathway of the barn owl," *Nature*, vol. 417, pp. 322–328, 2002.

[7] ——, "Early blindness results in a degraded auditory map of space in the optic tectum of the barn owl," *Proc Natl Acad Sci USA*, vol. 85, pp. 6211–6214, 1998.

[8] J. Hoernstein, M. Lopes, J. Santos-Victor, and F. Lacerda, "Sound localization for humanoid robots-building audio-motor maps based on the hrtf," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2006.

[9] H. Nakashima and N. Ohnishi, "Acquiring localization ability by interaction between motion and sensing," *IEEE International Conference on Systems, Man and Cybernetics*, October 1999.

[10] B. Bolder, M. Dunn, M. Gienger, H. Janssen, H. Sugiura, and C. Goerick, "Visually guided whole body interaction," *IEEE International Conference on Robotics and Automation (ICRA)*, 2007.

[11] T. Rodemann, F. Joublin, and C. Goerick, "Audio proto objects for improved sound localization," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2009.

[12] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Audio-visual speaker recognition using time-varying stream reliability prediction," *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. vol. V, pp. 712–715, April 2003.

[13] H. Kim, K. Komatani, T. Ogata, and G. Okuno, "Auditory and visual integration based localization and tracking of humans in daily-life environments," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2007.

[14] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for humanoids," *Proc. of 17th International Joint Conference on Artificial Intelligence (IJCAI)*, August 2001.

[15] K. Bernardin and R. Stiefelhagen, "Audio-visual multi-person tracking and identification for smart environments," *Proceedings of the 15th international conference on Multimedia*, 2007.

[16] H. Hung and G. Friedl, "Towards audio-visual on-line diarization of participants in group meetings," *Proceedings of European Conference on Computer Vision (ECCV)*, October 2008.

[17] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," *Advances in Neural Information Processing Systems*, vol. 12, pp. 813–819, 2000.

[18] H. Liu and M. Shen, "Continuous sound source localization based on microphone array for mobile robots," *Proceedings of the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems(IROS)*, October 2010.

[19] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[20] T. Rodemann, F. Joublin, and C. Goerick, "Continuous and robust saccade adaptation in a real-world environment," *KI-Kuenstliche Intelligenz*, March 2006.

[21] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP Journal on Applied Signal Processing*, January 2002.

Pos-1

Restricted Boltzmann Machine with Transformation Units in a Mirror Neuron System Architecture

Junpei Zhong, Cornelius Weber and Stefan Wermter
Department of Computer Science,
University of Hamburg,
Vogt-Koelln-Str. 30,
22527 Hamburg, Germany
Email: {zhong,weber,wermter}@informatik.uni-hamburg.de

***Abstract*—In the mirror neuron system, the canonical neurons play a role in object shape and observer-object relation recognition. However, there are almost no functional models of canonical neurons towards the integration of these two functions. We attempt to represent the relative position between the object and the robot in a neural network model. Although at present some generative models based on the Restricted Boltzmann Machine can code the image transformation in continuous images, what we need to accomplish in canonical neuron modeling is different from the requirements of modeling transformation in video frames. As a result, we propose a novel model called "Restricted Boltzmann Machine with Transformation Units", which can represent the relative object positions based on laser images. The laser sensor provides binary and accurate images and can further be connected with other models to construct a unified architecture of the mirror neuron system.**

I. Introduction

Since Rizzolatti and his colleagues found that some neurons in the F5 area of macaque monkeys' premotor cortex fired when the monkeys did actions like reaching for something or biting a peanut, the so-called mirror neurons have become a significant research topic that explains many social behaviors of human beings. A number of computational models ([1]–[8]) have been designed to model different functions of the mirror neuron system.

Moreover, a lot of research points out that the object affordance, the relative position between an observer and an object, should be considered in an integrated mirror neuron system model, because actions cannot be understood without understanding object affordances (e.g. [9], [10]). This function may be realized by the canonical neurons which are active when the object that can be grasped by movements is observed, as if the brain is foreseeing a possible interaction with this object and preparing itself accordingly. These neurons act with a mechanism of recognizing the object affordance with visual or other stimuli [11]. This does not lead to the motor action itself, but to the semantic knowledge about the actions. These kinds of canonical neurons also exist in the ventral premotor area F5 [12], [13].

As a result, if we judge the property of action understanding with consideration of the object affordance of the whole mirror neuron system, a proper computational model is necessary to consider both the object shape and the observer-object relation. For instance, in [14], [15] and [16] of the MNS and MNS2 model, the authors addressed this problem by manually calculating it in a geometric way.

In this paper, we propose an ongoing model which emphasizes the representation of the relative position of the object. Specifically, this position information is represented in a distributed manner in units called Transformation Units.

In the next section, we introduce the main architecture of the Restricted Boltzmann Machine and its modified version of coding object transformations. In section 3, the Restricted Boltzmann Machine with transformation units is presented. Then the experiment of the novel Restricted Boltzmann Machine is described in section 4. At the end we close with discussions and conclusions.

II. Restricted Boltzmann Machine and Related Works

In the neural networks community, the problem of representing relative positions can be simply converted into a similar problem of coding the object transformation in a distributed representation of a network. In this way the observer-object relation problem can be considered as a modified version of the transformation problem if we fix the observer at the origin and regard the image transformation as a representation of the observer-object relation.

A. Restricted Boltzmann Machine

A binary Restricted Boltzmann Machine (RBM) consists of a layer of visible units $v \in \{0,1\}$ and hidden units $h \in \{0,1\}$. The connections between the hidden units and the visible units are symmetric. Without interconnections between the visible units or the hidden units, the hidden units are conditionally independent. The probability distribution over the hidden units and the visible units is defined by

$$P(v,h) = P(V=v, H=h) = exp(v^T b + h^T c + v^T W h)/Z \tag{1}$$

where b is the vector of biases for the visible units, c is the vector of biases for the hidden units, W is the matrix of connection weights, and $Z(v,h) = \sum_{v,h} exp(v^T b + h^T c + v^T W h)$.

The RBM's definition ensures that all the distributions of values in hidden units $P(h|v)$ are only dependent on the values of visible units; the values in visible units are determined in the same way.

The updating rule for parameters W is:

$$\Delta W = \langle vh \rangle_{p(h|v,W)} - \langle vh \rangle_{p(v,h|w)} \tag{2}$$

the notation $\langle \cdot \rangle_{P()}$ is the expectation over the probability distribution $P()$.

B. RBM for Image Transformation

The first related paper about this topic of image transformation was advocated by Memisevic and Hinton [17]. They established a gated Restricted Boltzmann Machine which is able to extract distributed, domain-specific representations of image patch transformation. The model developed a mechanism to represent domain-specific motion features in the hidden units[1]. It works in two pathways: in the gated regression, from a given pair of two observed images, the hidden units are coded as a transformation. While in the modulated filters pathway, given the codes for the transformations as latent variables, the model can subsequently predict transformations of the next successive image. However, due to its tensor parameterization, the computational load of the training is quite high. Later Memisevic and Hinton developed a factored version of GRBM that uses three low-rank matrices as a factor instead of three-way parameters [18] to reduce the computational effort.

Based on the older version of GRBM, Taylor et al. proposed another method to reduce the complexity [19] by adding the convolution with a compact parameterization so that the invariance of relative transformations does not need to be re-learned.

A temporal Restricted Boltzmann Machine [20], [21] can be used as well to model sequences where the decision at each step requires some context information from the past. It contains one RBM model in each time step and these models are connected with latent to latent connections between the time steps. Therefore, it can also be used as a kind of transformation representation when the previous image provides a type of "temporal bias" to the hidden variables. Recently, Sutskever et al. [22] also proposed an improved type of TRBM by adding recurrent connections between the hidden units and visible units.

Recent research by Hinton [23] introduces the concept of "local capsules" which represents different kinds of transformations (poses, angles, lightings) so that it is possible to encapsulate the results with the inputs. The basic form of this model is an explicit representation of the pose. Furthermore, more levels of these capsules will be able to model different transformations.

III. Restricted Boltzmann Machine with Transformation Units

In this section we propose the Restricted Boltzmann Machine with Transformation Units (RBM-TU). As part of the mirror neuron system model, we expect that the information in the transformation units of the object's relative position can be further linked with PB units of the Simple Recurrent Network with Parametric Bias (SRN-PB) [24] so that these two units can interact as in Fig. 1. The RBM-TU is a modified version of the RBM. It has the same architecture as RBM except the full connections between transformation units and hidden layer (Fig. 2), so that this network can be regarded as a hybrid architecture consisting of a generative model and a discriminative model. The RBM recognizes and reconstructs the learned images, while the transformation units represent the relation between the objects and the sensor.

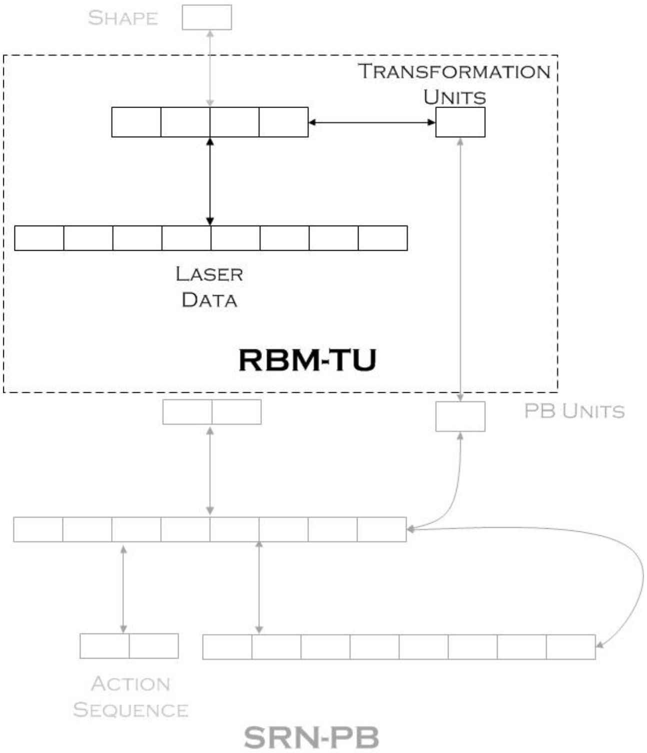


Fig. 1. The proposed computational model of a mirror neuron system towards action understanding: the RBM-TU represents information on the object shape and position, and the SRN-PB processes the action sequence together with the parametric biases units so that the structure of the input sequence can be characterized and recognized. The network within the dashed rectangle is the one we are proposing here.

Training is done by placing an object within the laser sensor area while the robot and the object stay still. The algorithm for the training is introduced in the next section. After training, the property of transformation units enables us to recognize the transformation of the input compared to the training sets. Comparing our approach to the GRBM, it is more straightforward to implement and reduce the computational effort

[1] These hidden units are not exactly the same as hidden units in generic RBM, but a kind of latent variables.

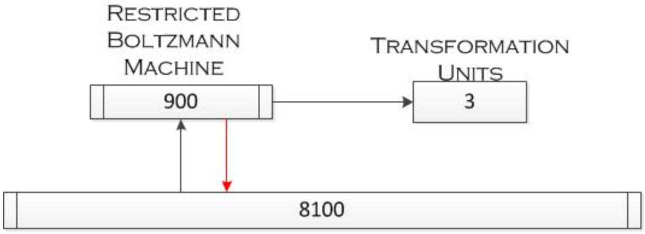


Fig. 2. Restricted Boltzmann Machine with transformation units Model: the laser information input into the visible units as a $90 * 90$ image. The hidden layer contains 900 units, which are fully connected with three transformation units. We simply use a linear function in the transformation units.

substantially because the updating of the RBM weights and the transformation units' weights are separate algorithms.

A. RBM Training

In the training mode, during each epoch, the weights between hidden units and visible units are updated in the same way as the generic RBM model by Contrastive Divergence. After that, the output values of the hidden units with the sigmoid function become inputs to the transformation units. In this way, the hidden units and the transformation units form an independent two-layer network, in which the first layer consists of the hidden units and the second layer consists of the transformation units. The connection weights between the hidden units and the transformation units are updated by back-propagating the error from the transformation units. For the position of the object during training, which acts as a reference position, the target values are set to zero. The activation function of hidden units can be expressed as:

$$H_h = \frac{1}{1+e^{(-s_h)}} \tag{3}$$

where s_h is the internal value of the hidden layer, and H_h is the output value of the hth hidden unit.

The activation function of transformation units is a linear function:

$$TU_i = \sum_h H_h W_{h,i} \tag{4}$$

where W is the connection weight matrix between hidden units and transformation units. The updating rule of connection weights between transformation units and hidden units is:

$$\Delta W_{h,i} = \eta \delta_i H_h \tag{5}$$

where η is the learning rate of transformation weights, and δ is the output error for the transformation units. The algorithm for a complete epoch can be depicted as follows:

Algorithm 1 RBM-TU Training

```
Extract a mini-batch of data sets from the training set.[2]
Update the connection weights between the hidden units and the visible units using Contrastive Divergence.
Calculate the values of hidden units given the updated weights and input.
Update the connection weights between the hidden units and the transformation units by Eq. 5.
```

In the recognition mode, the connection weights are fixed and the values in the transformation units are calculated with laser images of different relative positions of the object. We expect that the difference of dense coding in the hidden layer due to the relative position will lead to variations of the transformation units. In the next section, an experiment of representing untrained positions will be conducted to examine the plausibility of this model.

In the experiment, an Aldebaran NAO [26] is equipped with a laser head (Fig. 3). The laser head is built based on a URG-04LX laser sensor by Hokuyo [27], with the angle coverage of 240 degrees. The reason why we use a laser sensor is that understanding the object distance often needs calibration of vision processing, which takes great efforts in the robot perception process and needs to be re-calibrated in different environments. However, if we consider the perception development of human beings, the tactile sensation plays an important role in order to modulate the vision reference of a human [28]. Therefore, in this experiment, we attempt to simulate the tactile sensing via the laser sensor, which provides a more accurate position information than by only using a vision system. Further experiments may use it as an automatically calibration tool of robot vision system.

IV. Experiment Descriptions

A. Experiment Setup

Our environment is a normal office and an object is placed in front of the NAO for the training. The range sensor produces 500 images as data sets. These images have small variations because of angle variations of the laser beams and noise of the sensor, although we keep the positions of robot and object constant (Fig. 4). The two-dimensional dotted images of the laser head's surrounding are captured and learned (Fig. 6). During the RBM training, we select 100 images as one mini-batch to update the weights once after the whole mini-batch learning is finished. As soon as the training is over, the same object is placed in different positions from the trained position as long as they can be scanned by the laser sensor (the coordinates of these 20 example positions are depicted in Fig. 5).

The laser sensor in this experiment can be considered as a "measurement" method to survey the surrounding objects, and each part of the robot remains in the same position during training and recognition so that the transformation of the objects equals the relative positions of the objects with respect to the robot.

[2] The word "batch" here means dividing the training set into small sets, which is very advantageous for parallel computing on GPU boards or in Matlab [25]. When obtaining each patch of training data, the object remains in the origin position, and the laser range sensor scans the whole area 100 times.

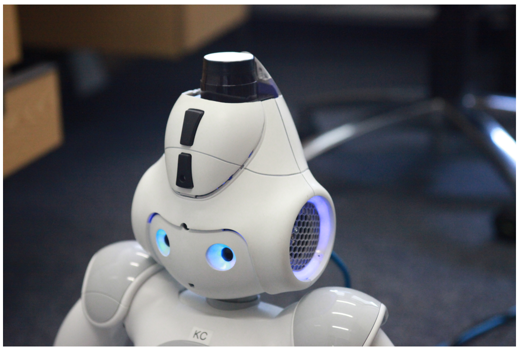
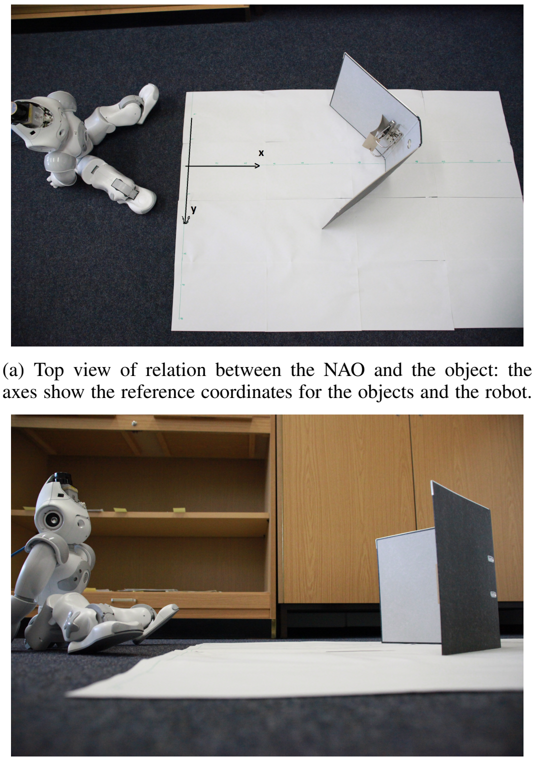


Fig. 3. Laser sensor in the NAO head



(a) Top view of relation between the NAO and the object: the axes show the reference coordinates for the objects and the robot.

(b) Side view of relation between the NAO and the object

Fig. 4. The NAO (left) senses different positions of the object (right, a folder) through the laser sensor. Markers on the ground are used to indicate the relative position between the robot and the object.

Binary images obtained from the laser sensor contain the information on whether the laser beam encounters an object and on the distance between sensor and object. These images are processed and reconstructed in the Restricted Boltzmann Machine. At the same time, we examine the relation between the representations in the transformation units and the actual transformation.

B. Experiment Results

We compare the values in transformation units with the selected untrained relative positions. Fig. 7 shows the relation

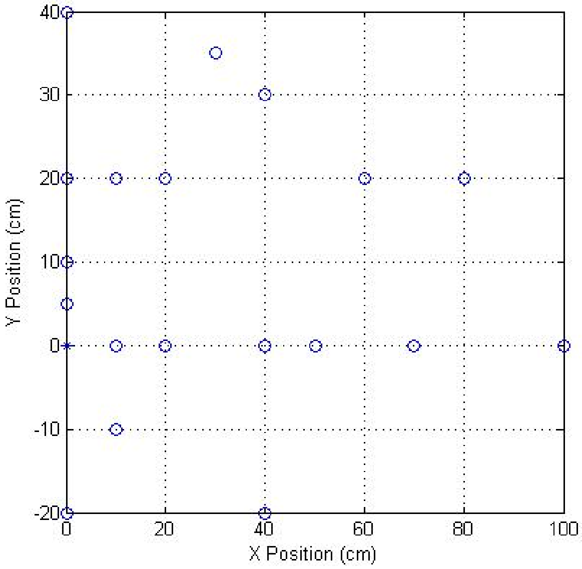


Fig. 5. Locations of 20 sample points: the object remains in the same position $(0, 0)$ during training (star mark). Then in recognition, 20 different sample points are picked to locate the object in order to test the values in transformation units (circle marks). The spine of the folder is always along the x axis.

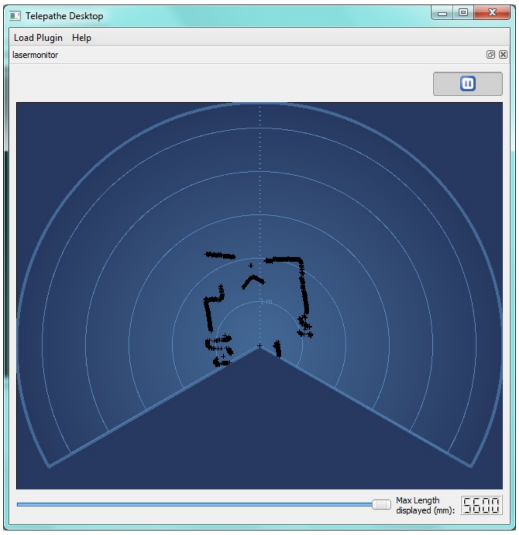
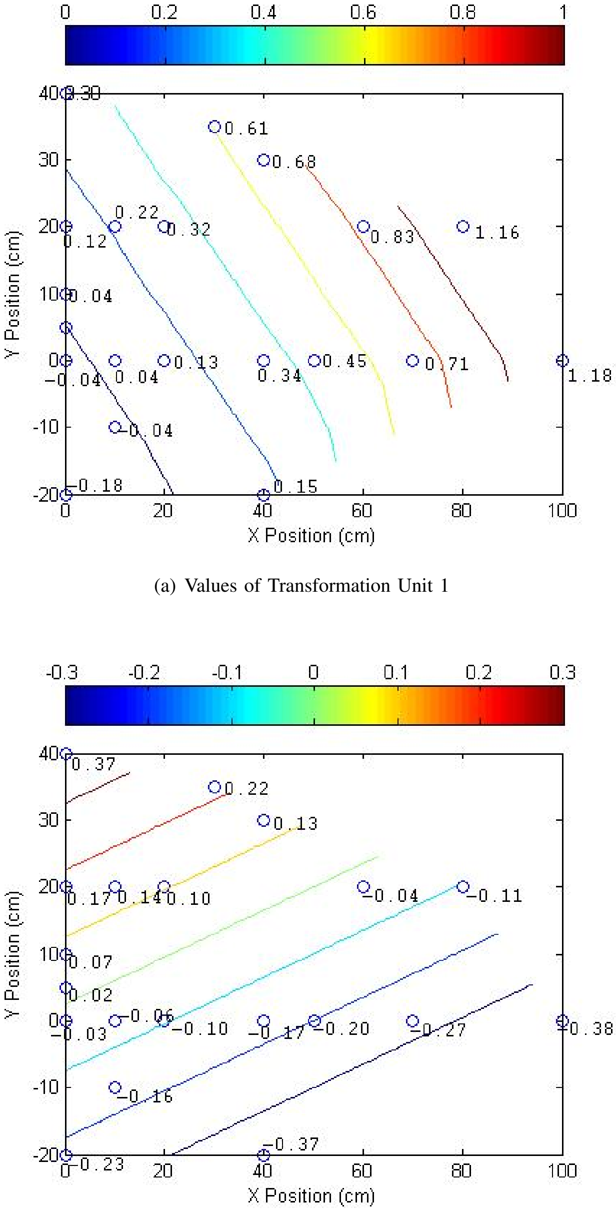


Fig. 6. Image obtained from NAO laser head: the binary image is obtained through laser beams reflected by obstacles. By the calculation of the time difference between emitting and receiving we can obtain the distance of the obstacle which that beam encounters.



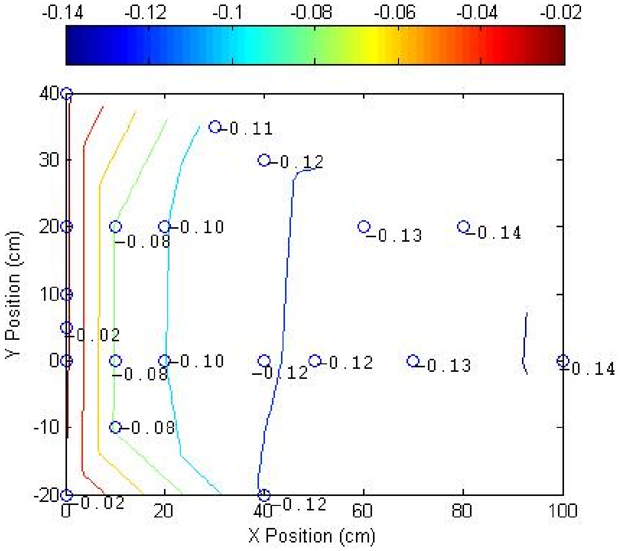
(a) Values of Transformation Unit 1

(b) Values of Transformation Unit 2

Fig. 7. Relation between Values in Transformation Units and Positions

between them.

Although we only select twenty positions as representatives within the range of the laser sensor, we can still identify the relation between transformation values and object positions. Fig. 7 uses interpolation to estimate the values in-between to better illustrate the transformation units. We can interpret this combination of the transformation units as the ability to locate a unique position of object; in this way we can regard



(c) Values of Transformation Unit 3

Fig. 7. Relation between Transformation Values and Positions (cont.): The circles represent the values obtained from transformation units (values are accurate to 0.01; some values are omitted due to the limited figure space). The contour curves are obtained by linear interpolation. We can pinpoint the relative position of the object along the contours in these images by associating the values in the transformation units.

the transformation units as a representation of the relative positions between the robot and the object.

V. Discussions and Conclusions

In this paper, we propose a novel hybrid architecture of a Restricted Boltzmann Machine with Transformation Units to model the functional model of observer-object relation representation in canonical neurons.

The RBM-TU model is a hybrid model which consists of a generative model (RBM) and a discriminative model (transformation units) to represent image transformation. Because of the independent learning processes between these two parts, the energy function is not as complex as for GRBM and other RBM-based image transformation models. Although it is not enough to apply as a video analysis for the transformation of consecutive images as GRBM does, it is straight-forward to be implemented because the independent RBM is the same as the original one.

The training of RBM and RBM-TU is almost the same as the generic RBM, except that the updating of weights between hidden units and transformation units are back-propagated by the target value of zeros when the object is placed in the original position. Due to the architecture of the separate generative and discriminative models, the training of RBM and transformation units are independent. This saves computational efforts compared to other RBM-based image transformation coding. The simple representation is suitable to process precise image representation, e.g. binary laser images.

In terms of coding the object affordance like canonical neurons in the mirror neuron system, this computational model of the canonical neurons towards object affordance is only partially developed. An ideal functional model should represent both:

1) representing the shape of the object;
2) representing the relation of observer to the object.

The two requirements above can be related to the recognition and representation of "what" and "where" the object is. Although this problem is not yet fully solved, our model with laser sensing could become a reasonable foundational architecture of the multi-modal sensing to better perceive the information. Although we only use one object here to justify the transformation units, in further experiments RBM-TU with a soft-max label layer will be used to identify the object shapes as well.

The current experiment is done in the fixed context of a room during both learning and recognizing the object position. In the next experiment we will attempt to combine the robot vision and laser sensor as multi-modal input so as to eliminate the room context and concentrate on the relative position of robot and object. Moreover, as presented in the first section, this model can be expected to be further integrated with SRNPB [24] to build up a complete mirror neuron system model.

Acknowledgment

This research has been partially supported by the EU projects RobotDoc under 235065 from the FP7, Marie Curie Action ITN and KSERA by the European Commission under n°2010-248085. We thank Katja Kösters for language suggestions.

References

[1] Y. Demiris and B. Khadhouri, "Hierarchical attentive multiple models for execution and recognition of actions," *Robotics and Autonomous Systems*, vol. 54, no. 5, pp. 361–369, 2006.

[2] Y. Demiris and M. Johnson, "Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning," *Connection Science*, vol. 15, no. 4, pp. 231–243, 2003.

[3] D. Wolpert and M. Kawato, "Multiple paired forward and inverse models for motor control," *Neural Networks*, vol. 11, no. 7-8, pp. 1317–1329, 1998.

[4] M. Haruno, D. Wolpert, and M. Kawato, "Hierarchical mosaic for movement generation," in *International Congress Series*, vol. 1250. Elsevier, 2003, pp. 575–590.

[5] J. Tani, M. Ito, and Y. Sugita, "Self-organization of distributedly represented multiple behavior schemata in a mirror system: reviews of robot experiments using RNNPB," *Neural Networks*, vol. 17, no. 8-9, pp. 1273–1289, 2004.

[6] E. Borenstein and E. Ruppin, "The evolution of imitation and mirror neurons in adaptive agents," *Cognitive Systems Research*, vol. 6, no. 3, pp. 229–242, 2005.

[7] M. Elshaw, C. Weber, A. Zochios, and S. Wermter, "A mirror neuron inspired hierarchical network for action selection," *Proc. NeuroBotics*, pp. 89–97, 2004.

[8] S. Wermter, C. Weber, M. Elshaw, V. Gallese, and F. Pulvermuller, "A mirror neuron inspired hierarchical network for action selection," *Biomimetic neural learning for intelligent robots*, pp. 162–181, 2005.

[9] G. Metta, G. Sandini, L. Natale, L. Craighero, and L. Fadiga, "Understanding mirror neurons: a bio-robotic approach," *Interaction studies*, vol. 7, no. 2, pp. 197–232, 2006.

[10] E. Sahin and S. Erdogan, "Towards linking affordances with mirror/canonical neurons," in *24th International Symposium on Computer and Information Sciences, 2009. ISCIS 2009.* IEEE, 2009, pp. 397–404.

[11] A. Murata, L. Fadiga, L. Fogassi, V. Gallese, V. Raos, and G. Rizzolatti, "Object representation in the ventral premotor cortex (area f5) of the monkey," *Journal of Neurophysiology*, vol. 78, no. 4, p. 2226, 1997.

[12] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. C. Mazziotta, and G. Rizzolatti, "Grasping the intentions of others with one's own mirror neuron system," *PLoS Biol*, vol. 3, no. 3, p. e79, 02 2005.

[13] J. Grezes, J. L. Armony, J. Rowe, and R. E. Passingham, "Activations related to and neurones in the human brain: an fmri study," *NeuroImage*, vol. 18, no. 4, pp. 928 – 937, 2003.

[14] E. Oztop and M. Arbib, "Schema design and implementation of the grasp-related mirror neuron system," *Biological cybernetics*, vol. 87, no. 2, pp. 116–140, 2002.

[15] J. Bonaiuto, E. Rosta, and M. Arbib, "Extending the mirror neuron system model, i," *Biological cybernetics*, vol. 96, no. 1, pp. 9–38, 2007.

[16] J. Bonaiuto and M. Arbib, "Extending the mirror neuron system model, ii: what did i just do? a new role for mirror neurons," *Biological cybernetics*, vol. 102, no. 4, pp. 341–359, 2010.

[17] R. Memisevic and G. Hinton, "Unsupervised learning of image transformations," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[18] R. Memisevic and G. E. Hinton, "Learning to represent spatial transformations with factored higher-order boltzmann machines," *Neural Comput.*, vol. 22, pp. 1473–1492, 2010.

[19] G. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," *Computer Vision–ECCV 2010*, pp. 140–153, 2010.

[20] G. Taylor, G. Hinton, and S. Roweis, "Modeling human motion using binary latent variables," *Advances in neural information processing systems*, vol. 19, p. 1345, 2007.

[21] I. Sutskever and G. Hinton, "Learning multilevel distributed representations for high-dimensional sequences," in *Proceeding of the Eleventh International Conference on Artificial Intelligence and Statistics*. Citeseer, 2007, pp. 544–551.

[22] I. Sutskever, G. Hinton, and G. Taylor, "The recurrent temporal restricted boltzmann machine," *Advances in Neural Information Processing Systems*, vol. 21, 2009.

[23] G. Hinton, A. Krizhevsky, and S. Wang, "Transforming auto-encoders," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, ser. Lecture Notes in Computer Science, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds. Springer Berlin / Heidelberg, 2011, vol. 6791, pp. 44–51.

[24] J. Zhong, C. Weber, and S. Wermter, "Robot trajectory prediction and recognition based on a computational mirror neurons model," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, ser. Lecture Notes in Computer Science, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds. Springer Berlin / Heidelberg, 2011, vol. 6792, pp. 333–340.

[25] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, p. 1, 2010.

[26] N. Aldebaran, "Robotics http://www. aldebaran-robotics. com/eng," *NaoRobocup. php [Consulta Julio 2009]*.

[27] L. Kneip, F. Tâche, G. Caprari, and R. Siegwart, "Characterization of the compact hokuyo urg-04lx 2d laser range scanner," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 1447–1454.

[28] E. Azañón, K. Camacho, and S. Soto-Faraco, "Tactile remapping beyond space," *European Journal of Neuroscience*, vol. 31, no. 10, pp. 1858–1867, 2010.

Towards Robust Speech Recognition for Human-Robot Interaction

Stefan Heinrich and Stefan Wermter
Knowledge Technology Group, Department of Informatics, University of Hamburg, Hamburg, Germany
Email: {heinrich,wermter}@informatik.uni-hamburg.de

Abstract—**Robust speech recognition under noisy conditions like in human-robot interaction (HRI) in a natural environment often can only be achieved by relying on a headset and restricting the available set of utterances or the set of different speakers. Current automatic speech recognition (ASR) systems are commonly based on finite-state grammars (FSG) or statistical language models like Tri-grams, which achieve good recognition rates but have specific limitations such as a high rate of false positives or insufficient rates for the sentence accuracy. In this paper we present an investigation of comparing different forms of spoken human-robot interaction including a ceiling boundary microphone and microphones of the humanoid robot NAO with a headset. We describe and evaluate an ASR system using a multi-pass decoder – which combines the advantages of an FSG and a Tri-gram decoder – and show its usefulness in HRI.**

I. Introduction

With current speech recognition systems it is possible to reach an acceptable word recognition rate if the system has been adapted to a user, or if the system works under low-noise conditions. However, on the one hand in *human-robot interaction* (HRI) or in *ambient intelligence environments* (AmIE), the need for robust and automatic speech recognition is still immanent [1], [2]. On the other hand research in *cognitive neuroscience robotics* (CNR) and multimodal communication benefits from a robust and functioning speech recognition as a basis [3]. Headsets and other user-bound microphones are not convenient in an natural environment in which, for instance, a robot is supposed to interact with an elderly person. A microphone built into the robot or placed at the ceiling, a wall, or a table allows for free movement but reduces the quality of speech signals substantially because of larger distances to the person and therefore more background noise.

One method to deal with the additional problems is of course a further adaptation of the speech recogniser towards a domain-specific vocabulary and grammar. Enhancing recognised speech with a grammar-based decoder (*finite state grammar*, FSG) can lead to improved results in terms of recognised sentences, but it also leads to a high rate of false positives, since an FSG decoder tries to map the recognised utterances to legal sentences. To deal with this problem, one can combine the FSG with the classical Tri-gram decoder to reject unlikely results. Such a multi-pass decoder can be applied also to noisy sound sources like a ceiling boundary microphone or microphones, installed on a robot.

In the past research has been done on combining FSG and N-grams decoding processes: In 1997 Lin et. al. used an FSG and an N-gram decoder for spotting key-phrases in longer sentences [4]. Based on the assumption that sentences of interest are usually surrounded by carrier phrases, they employed N-gram decoding to cover those surrounding phrases on the one hand and FSG decoding on the other hand if a start word of the grammar was found by the N-gram decoder. Furthermore, with their approach they rejected FSG-hypotheses if the average word score exceeded a preset threshold. However, this approach combined FSG and N-grams while modifying and fine-tuning the decoding processes on a very low-level, preventing to switch to another FSG or N-gram model easily. Therefore it would be interesting to exploit the dynamical result of an N-gram hypotheses list for the rating of an FSG-hypothesis instead of a fixed threshold.

Levit et. al. combined 2009 an FSG decoder and a second different decoder in a complimentary manner for the use in small devices [5]. In their approach they used an FSG decoder as a fast and efficient baseline recogniser, capable of recognising only a limited number of utterances. The second decoder, used for augmenting the first decoder, was also FSG-based but according to the authors could be replaced by a statistical language model like N-grams, too. An augmentation for the first decoder could be a 'decoy', which is a sentence with a similar meaning, similar to an already included sentence. However, those decoys can only be trained off-line. In this approach the result of the first decoder was not rated or rejected afterwards, but the search space was shaped to avoid the appearance of false positives.

Doostdar et. al. proposed 2008 an approach where an FSG and a Tri-gram decoder processed speech data independently based on a common acoustic model [6]. The best hypothesis of the FSG decoder was compared with the n-best list of hypotheses of the Tri-gram decoder. Without modifying essential parts of the underlying system they achieved a high false positive reduction and overall a good recognition rate, while they restricted the domain to 36 words and a command grammar. Although aiming for applying their system on service robots, they limited their investigation to the use of a headset. Yet it would be interesting to test such an approach far-field in a real environment using the service robots' microphones or other user-independent microphones.

In contrast, Sasaki et. al. investigated 2008 the usability of a command recognition system using a ceiling microphone array [7]. After detecting and separating a sound source an extracted sound was fed to a speech recogniser. The used open source speech recognition engine was configured for the use of 30 words and a very simple grammar allowing only 4 different sentence types like GO TO X or COME HERE. With their experiments, the authors have shown that using a ceiling microphone in combination with a limited dictionary leads to a moderate word accuracy rate. Also they claim that their approach is applicable to a robot, which uses an embedded microphone array. A crucial open question is the effect on the sentence accuracy if a more natural interaction and therefore a larger vocabulary and grammar is being used. Based on the presented moderate word accuracy the sentence accuracy is likely to be small for sentences with more than three words, leading to many false positives.

In this paper we present a speech recognition approach with a multi-pass decoder in a home environment addressing the research question of the effect of the decoder in the far-field. We test the usability of HRI and investigate the effect of different microphones, including the microphones of the NAO humanoid robot and a boundary microphone, placed at the ceiling, compared to a standard headset. After analysing the background of speech recognition we will detail the description of a multi-pass decoder in section 2. Then we will describe the scenario for the empirical evaluation in section 3, present the results of our experiments in section 4, and draw a conclusion in section 5.

II. The Approach

Before explaining the multi-pass decoder in detail, we first outline some relevant fundamentals of a statistical speech recognition system and the architecture of a common single-pass decoder (see also [8]).

A. Speech Recognition Background

The input of a speech recogniser is a complex series of changes in air pressure, which through sampling and quantisation can be digitalised to a pulse-code-modulated audio stream. From an audio stream the features or the characteristics of specific phones can be extracted. A statistical speech recogniser, which uses a Hidden Markov Model (HMM), can determine the likelihoods of those acoustic observations.

With a finite grammar or a statistical language model, a search space can be constructed, which consists of HMMs determined by the acoustic model. Both, grammar and language model, are based on a dictionary, defining which sequence of phones constitute which words. The grammar defines a state automaton of predefined transitions between words, including the transition probabilities. Language models in contrast are trained statistically based on the measured frequency of a word preceding another word. With so-called N-grams, dependencies between a word and the $(N-1)$ preceding words can be determined. Since N-grams of higher order need substantially more training data Bi-Grams or Tri-grams are often used in current *automatic speech recognition* (ASR) systems.

During processing of an utterance, a statistical speech recogniser searches the generated graph for the best fitting hypothesis. In every time frame, the possible hypotheses are scored. With a best-first search, or a specialised search algorithm like the Viterbi Algorithm, hypotheses with bad scores are pruned.

In principle it is possible to adapt ASR for improving the recognition rate with two different approaches:

1) The acoustic model is trained for a single specific speaker. This method leads to precise HMM's for phones, which allows for a larger vocabulary.
2) The domain is restricted in terms of a limited vocabulary. This restricted approach reaches good recognition rates even with an acoustical model trained for many different speakers.

B. Multi-Pass Decoder

Both introduced methods, the *finite state grammar* (FSG) based decoder as well as the Tri-gram decoder, have specific advantages and limitations.

- The FSG decoder can be very strict, allowing valid sentences without fillers only. Unfortunately, such an FSG decoder maps every input to a path in the search space, which is spanned from all valid starting words to all valid finishing words. For example if the speaker is using a sentence like NAO *EHM* PICK PHONE, the decoder may map it to a most likely sentence like NAO WHERE IS PHONE. Even if the speaker is just randomly putting words together, the decoder may often produce a valid sentence and therefore – very often – a false positive.
- With a Tri-Gram decoder an ASR system is more flexible and can get decent results if the quality of the audio signal is high and the data set for training the language model is sufficiently large. However, since Tri-grams mainly take into account the last two most probable words, they cannot deal with long-range dependencies. Therefore even if the word accuracy is reasonably high, the sentence accuracy as a cumulative product is fairly moderate [8].

To overcome the limitations of both single decoders, we can combine them to a multi-pass decoder. First, we use the FSG decoder to produce the most likely hypothesis. Second, we use the Tri-gram decoder – which is able to backoff to Bi-grams or Uni-grams – to produce a reasonably large list of best hypotheses. Even if the best hypothesis of the Tri-gram decoder is not appropriate there is a good chance that one of the similar sentences is. In the next step, we compare the best hypothesis of the FSG decoder with the list of n-best hypotheses of the Tri-gram decoder. If we find a match we can accept this sentence, otherwise we reject the sentence. Figure 1 illustrates the HMM-based ASR system using the multi-pass decoder.

C. Speech Recogniser and its Adaptation

In this study, we use the ASR framework *Pocketsphinx*, because it is open source and has been ported and optimised for hand-held devices [9]. In comparison to other promising systems [10], [11] it provides the advantage of being an effective research tool on the one hand and being applicable to devices and robots with moderate computing power on the other hand. Pocketsphinx comes with a speaker-independent acoustic-model 'HUB4' based on English broadcast news. Also available is a language model trained on the same data.

Since it is our aim to keep the system speaker independent, we decided to limit the vocabulary and to reduce the format

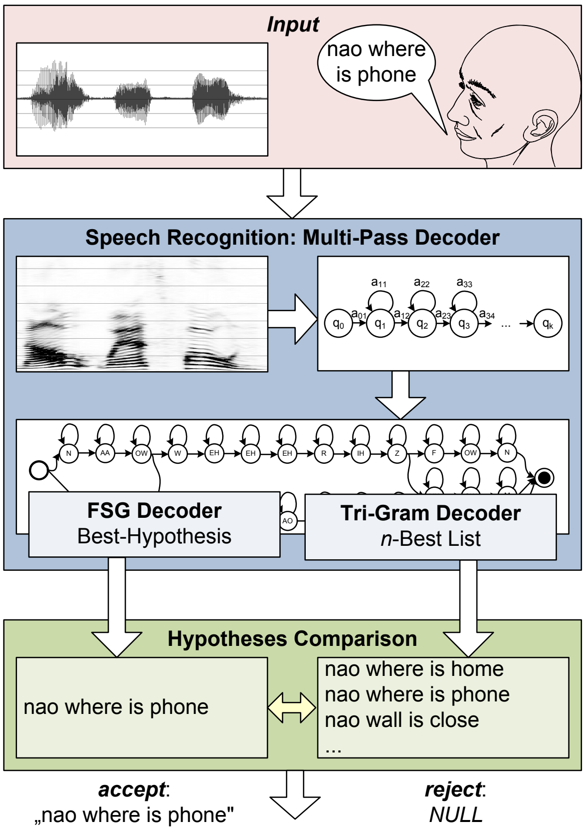


Fig. 1. Architecture of a multi-pass decoder

```
public <utterance>  = <confirmation> |(nao <communication>);

<communication>  = <information> | <instruction> | <question>;
<instruction>    = <command> | <action>;
<information>    = ((<object> | <agent>) close to (<object>
                     | <agent> | <place>))
                   | (<object> can be <affordance>)
                   | (<object> has color <color>);
<question>       = (what can <object>)
                   | (which color has <object>)
                   | (where is (<object> | <agent>));
<confirmation>   = yes | correct | right | (well done) | no
                   | wrong | incorrect;
<command>        = abort | help | reset | (shut down) | stop;
<action>         = <head_action> | <hand_action> | <body_action>;
<hand_action>    = (<affordance> <object>)
                   | (show (<object> | <agent>) );
<body_action>    = (turn body <direction>) | (sit down)
                   | (walk <number>) | (bring <object>)
                   | (go to (<agent> | <object>) ) | (come here);
<head_action>    = (turn head <direction>)
                   | ((find | look at) (<object> | <agent>))
                   | (follow <agent>);
<agent>          = nao | i | patient;
<object>         = apple | banana | ball
                   | dice | phone | oximeter;
<direction>      = left | straight | right;
<number>         = one | two | three;
<affordance>     = pick | drop | push;
<color>          = yellow | orange | red | purple | blue | green;
<place>          = home | desk | sofa | chair | floor | wall;
```

Fig. 2. Grammar for the scenario

of a sentence to a simpler situated grammar or command grammar, as it can be useful in HRI. Devices and robots in our AmIE are supposed to be used for a specific set of tasks, while the scenario can have different human interactors. The acoustic-model HUB4 was trained over a very large set of data (140 hours) including different English speakers [12]. With a vocabulary reduction to 100 words and the new grammar, as outlined in figure 2, we generated an own FSG automaton on the one hand and trained an own language model on the other hand. For the training of the language model, we used the complete set of sentences which can be produced with our grammar. The grammar allows for short answers like YES or INCORRECT as well as for more complex descriptions of the environment like NAO BANANA HAS COLOR YELLOW.

In summary we adapted Pocketsphinx to recognise instruction, information, and question sentences in English.

III. Our Scenario

The scenario of this study is an ambient intelligent home environment. To investigate opportunities and chances of technical devices and humanoid robots in home environments, those scenarios are of increasing relevance [13], [14]. In particular EU research projects like KSERA aim to develop a socially assistive robot that helps elderly people [15]. Such a scenario consists of a home environment including interactive devices and a humanoid robot.

A. Environment

Our AmIE is a lab room of 7x4 meters, which is furnished like a standard home without specific equipment to reduce noise or echoes, and is equipped with technical devices like a ceiling boundary microphone and a NAO H25 humanoid robot. A human user is supposed to interact with the environment and the NAO robot and therefore should be able to communicate in natural language. For this study the human user is wearing a headset as a reference microphone. The scenario is presented in detail in figure 3. The details of the used microphones are as follows:

a) Ceiling Microphone: The ceiling boundary microphone is a condenser microphone of 85 mm width, placed three meter above the ground. It is using an omni-directional polar pattern and has a frequency response of 30Hz - 18kHz.

b) NAO: The NAO robot is a 58 cm tall robot with 25 degrees of freedom (*DOF*), two VGA cameras, and four microphones, developed for academic purposes [16]. Besides his physical robustness, the robot provides some basic integrated functionalities like an initial set of prepared movements, a detection system for visual markers, and a text-to-speech module. Controllable over WLAN with a mounted C++ API namely NaoQi, the NAO can be used as a completely autonomously agent or as a remotely controlled machine. The microphones are placed around the head and have an electrical bandpass of 300Hz - 8kHz. In its current version the NAO uses a basic noise reduction technique to improve the quality of processed sounds.

c) Headset: The used headset is a mid-segment headset specialised for communication. The frequency response of the microphone is between 100Hz - 10kHz.

To allow reliable comparison, the location of the speaker is at a distance of 2m meter to the ceiling microphone as well as to the NAO robot.

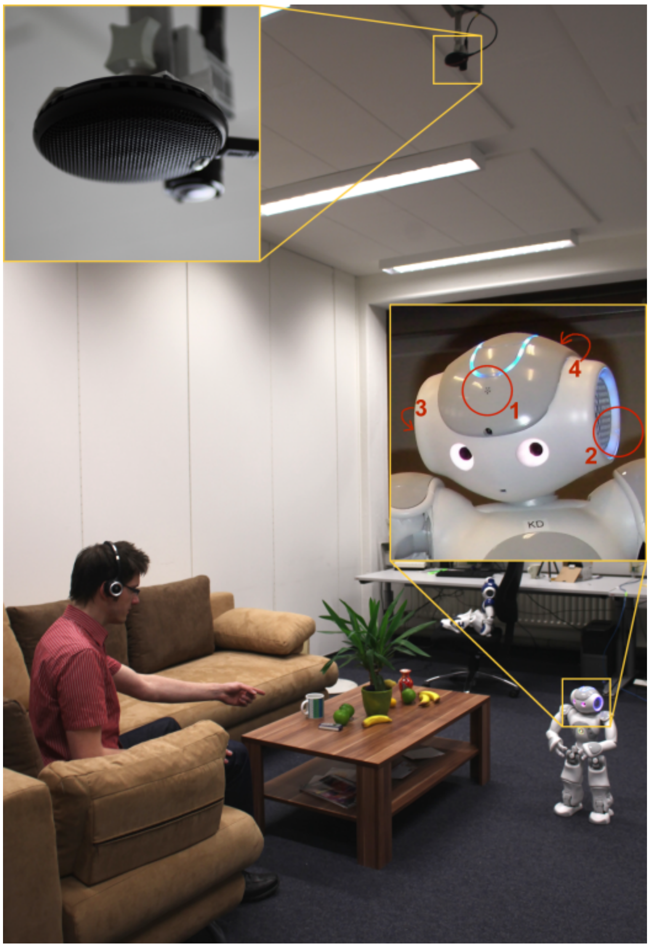


Fig. 3. Scenario environment

B. Dataset

The set of data to test the approach was collected under natural conditions within our AmIE. Different non-native English mixed male and female test subjects were asked to read a random sentence, produced from our grammar. All sentences were recorded in parallel with the headset, the ceiling microphone and the NAO robot in a 16 bit format and a sample rate of 48.000 Hz. In summary we collected 592 recorded sentences each, which led to 1776 audio files.

C. Evaluation Method

For the empirical validation, we converted all files to the monaural, little-endian, unheadered 16-bit signed PCM audio format sampled at 16000 Hz, which is the standard audio input stream for Pocketsphinx.

With Pocketsphinx we run a speech recognition test on every recorded sentence. Since it is not the focus of this study to test for false negatives and true negatives, we did not include incorrect sentences or empty recordings in the test. The result of the speech recogniser was compared with the whole desired sentence to check for the sentence accuracy as means of comparability. If the sentence was completely correct it was counted as true positive, otherwise a false positive. For example if the correct sentence is NAO WHAT COLOR HAS BALL, then NAO WHAT COLOR HAS WALL as well as NAO WHAT COLOR IS BALL are incorrect.

To test for statistical significance of the false positive reduction with the multi-pass decoder, we calculated the *chi-square* (χ^2) score over the true-positives/false-positives ratios. If, for example, the χ^2 score over the tp/fp ratio of the multi-pass against the tp/fp ratio of the FSG decoder is very high, then we have evidence for a high degree of dissimilarity [17].

IV. Empirical Results

The empirical investigation of our approach consists of two parts. First, we analysed the overall rate of true and false positives of the multi-pass decoder in comparison to specific single-pass decoders. Second, we determined the influence of the size n of the list of best hypotheses. Every investigation has been carried out in parallel for every microphone type as described above.

A. Effect of Different Decoders

With the 592 recorded sentences we tested the speech recognition using the FSG-decoder and the Tri-gram decoder in a single-pass fashion and combined them in a multi-pass fashion, using n-best list size of 10. In table I the results are presented where every row contains the number of correctly recognised sentences (true positives) and incorrectly recognised sentences (false positives).

TABLE I
Comparison of different decoders

(a) FSG decoder

	True positives	False positives	Tp/fp ratio
Headset	458 (77.4%)	101 (17.1%)	81.93%
Ceiling mic.	251 (42.4%)	251 (50.3%)	45.72%
NAO robot	39 (6.6%)	447 (75.5%)	8.02%

(b) Tri-gram decoder

	True positives	False positives	Tp/fp ratio
Headset	380 (64.2%)	212 (35.8%)	64.19%
Ceiling mic.	133 (22.5%)	459 (77.5%)	22.47%
NAO robot	14 (2.4%)	322 (54.4%)	4.17%

(c) Multi-pass decoder, $n = 10$

	True positives	False positives	Tp/fp ratio
Headset	378 (63.9%)	24 (4.1%)	94.03%
Ceiling mic.	160 (27.0%)	76 (12.8%)	67.80%
NAO robot	31 (5.2%)	130 (22.0%)	19.25%

tp/fp ratio = tp / (tp + fp) * 100

The data shows that for a headset every decoder led to a relatively high rate of correct sentences, counting 458 (77.4%) with the FSG, 380 (64.2%) with the Tri-gram, and 378 (63.9%) with the multi-pass decoder. The single-pass decoder produced 101 false positives (tp/fp ratio of 81.93%) with FSG and 212 false positives (tp/fp ratio of 64.19%) with Tri-gram, while the multi-pass decoder produced 24 false positives (tp/fp ratio of 94.03%).

For the ceiling microphone the rate of correct sentences was fairly moderate, reaching 251 (42.4%) with the FSG, 133 (22.5%) with the Tri-gram, and 160 (27.0%) with the multi-pass decoder. The number of produced false positives was relativly high for the single-pass decoder reaching 298 (tp/fp ratio of 45.72%) with FSG and 459 false positives (tp/fp ratio of 22.47%) with Tri-gram, whereas the multi-pass decoder produced 76 false positives (tp/fp ratio of 67.80%).

The rate of correct sentences for the NAO robot microphones was very low, getting only 39 (6.6%) with the FSG, 14 (2.4%) with the Tri-gram, and 31 (5.2%) with the multi-pass decoder. However, the single-pass decoder produced 447 false positives (tp/fp ratio of 8.02%) with the FSG and 322 false positives (tp/fp ratio of 4.17%) with the Tri-gram, while the multi-pass decoder produced 130 false positives (tp/fp ratio of 19.25%).

In table II some examples for the the recognition results with different decoder and microphones are presented. The results indicate that in many cases where sentences could not be recognised correctly, some specific single words like APPLE were recognised incorrectly. In some cases valid but incorrect sentences were recognised by both decoders, but were successfully rejected by the multi-pass decoder. Furthermore, with the NAO robot often only single words were recognised.

TABLE II
EXAMPLES OF RECOGNISED SENTENCES

🟩 True positive	🟨 Rejected	🟥 False positive

(a) "NAO GO TO OXIMETER"

	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	🟩 NAO GO TO OXIMETER	🟥 NAO WHAT COLOR OXIMETER	🟩 NAO GO TO OXIMETER
Ceiling mic.	🟥 NAO SIT DOWN	🟥 NAO SIT DOWN	🟥 NAO SIT DOWN
NAO robot	🟩 NAO GO TO OXIMETER	🟥 NAO BE	🟨

(b) "NAO APPLE CLOSE TO PATIENT"

	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	🟨	🟥 NAO APPLE HAS CLOSE TO PATIENT	🟨
Ceiling mic.	🟥 NAO I CLOSE TO PATIENT	🟥 NAO HEAD CLOSE TO PATIENT	🟨
NAO robot	🟥 NAO FIND PATIENT	🟥 NAO TO PATIENT	🟨

(c) "NAO WHICH COLOR HAS BALL"

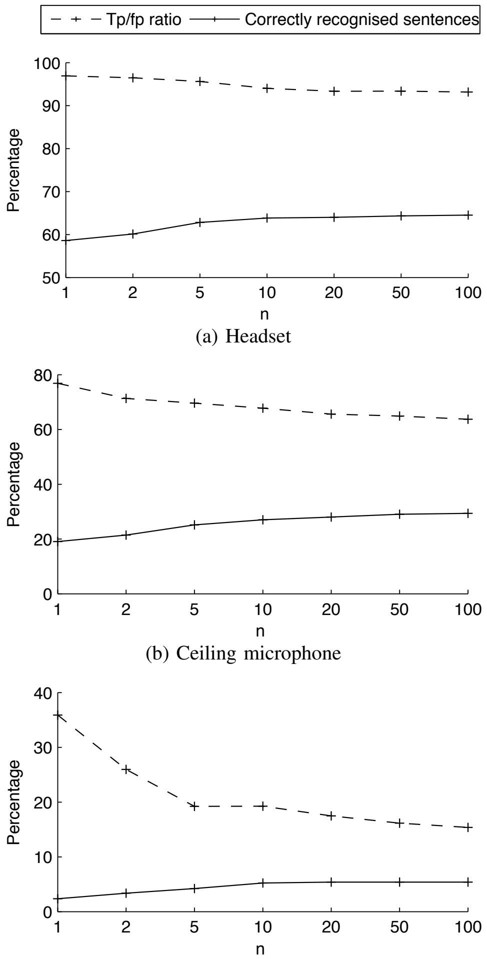
	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	🟩 NAO WHICH COLOR HAS BALL	🟩 NAO WHICH COLOR HAS BALL	🟩 NAO WHICH COLOR HAS BALL
Ceiling mic.	🟥 NAO WHERE IS PHONE	🟥 NAO WHERE IS HEAD AT PHONE	🟨
NAO robot	🟥 NO	🟨	🟨

(d) "WELL DONE"

	FSG decoder	Tri-gram dec.	Multi-pass dec.
Headset	🟩 WELL DONE	🟩 WELL DONE	🟩 WELL DONE
Ceiling mic.	🟩 WELL DONE	🟩 WELL DONE	🟩 WELL DONE
NAO robot	🟥 YES	🟨	🟨

B. *Influence of Parameter n*

To determine the influence of the size of the n-best list, we varied n over $\{1, 2, 5, 10, 20, 50, 100\}$. Figure 4 displays the ratio of true positives and false positives in comparison to the rate of correctly recognised sentences for every microphone type as described above.



(a) Headset

(b) Ceiling microphone

(c) NAO robot

Fig. 4. Comparison of true positives/false positives ratio and correctly recognised sentences

On the one hand, for small n the percentage of false positives is smaller for every microphone type. On the other hand, a small n results in a more frequent rejection of sentences.

Finding an optimal n seems to strongly depend on the microphone used and therefore on the expected quality of the speech signals. In our scenario a larger n around 20 is sufficient for the use of headsets, in terms of getting a good true positives to false positives ratio while not rejecting too many good candidates. For a moderate microphone like the ceiling microphone, a smaller n around 5 is sufficient. With low-quality microphones like in the NAO robot the variance of n does not point to an optimal configuration. Smaller n result in very few correctly recognised sentences, while larger n result in a very low tp/fs rate.

C. Result Summary

In summary, we observed that using a multi-pass decoder reduced the number of produced false positives significantly. For a low-noise headset as well as for boundary microphones and inexpensive microphones installed on a mobile robot, the experiment has shown that reducing the false positives to a good degree does not lead to a substantial reduction of true positives. The overall recognition rates with the NAO were insufficient, while the ceiling microphone worked with a reasonable rate using the multi-pass decoder. A good value for n depends on the hypotheses space and the microphone used. For our scenario, overall using $n = 10$ best hypotheses was sufficient. If the expected quality is moderate and the number of different words and possible sentences are high, then a larger value for n is likely to lead to better results.

V. Conclusion

In this paper we presented a study of speech recognition using a multi-pass FSG and Tri-gram decoder comparing a ceiling microphone and the microphones of a humanoid robot with a standard headset. The results of our approach are in line with [6], showing that a multi-pass decoder can successfully be used to reduce false positives and to obtain robust speech recognition. Furthermore we can state that using a multi-pass decoder in combination with a ceiling boundary microphone is useful for HRI: Adapting to domain-specific vocabulary and grammar on the one hand and combining the advantages of an FSG and a Tri-gram decoder leads to acceptable speech recognition rates. The size of the n-best list is not very crucial and depends on the search space to some extent. Build-in microphones of humanoid robots such as the NAO still come with a low SRN due to noisy fans or motors, and need intensive preprocessing to allow for speech recognition.

In the future the proposed method can be improved in various ways. First, one could improve the quality of the speech recorded by a (ceiling) microphone itself. Using for example a sophisticated noise filter or integrating a large number of microphones could lead to a more reliable result [18]. Second, one could not only integrate different decoding methods but also the context information into one ASR system to accept or reject recognised utterances. For example vision could provide information about lip movement and therefore provide probabilities for silence or a specific phoneme [19]. Speech recognition serves as a starting ground for research in HRI and CNR and as a driving force for a better understanding of language itself. In this context we have shown that using a multi-pass decoder and environmental microphones is a viable approach.

Acknowledgment

The authors would like to thank Arne Köhn, Carolin Mönter, and Sebastian Schneegans for the support in automatically collecting a large set of data. We also thank our collaborating partners of the KSERA project funded by the European Commission under n° 2010-248085 and of the RobotDoC project funded by Marie Curie ITN under 235065.

References

[1] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "A communication robot in a shopping mall," *IEEE Robotics and Automation Society*, vol. 26, no. 5, pp. 897–913, 2010.

[2] K. K. Paliwal and K. Yao, "Robust speech recognition under noisy ambient conditions," in *Human-Centric Interfaces for Ambient Intelligence*. Academic Press, Elsevier, 2009, ch. 6.

[3] S. Wermter, M. Page, M. Knowles, V. Gallese, F. Pulvermüller, and J. G. Taylor, "Multimodal communication in animals, humans and robots: An introduction to perspectives in brain-inspired informatics," *Neural Networks*, vol. 22, no. 2, pp. 111–115, 2009.

[4] Q. Lin, D. Lubensky, M. Picheny, and P. S. Rao, "Key-phrase spotting using an integrated language model of n-grams and finite-state grammar," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*. Rhodes, Greece: ISCA Archive, Sep. 1997, pp. 255–258.

[5] M. Levit, S. Chang, and B. Buntschuh, "Garbage modeling with decoys for a sequential recognition scenario," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU 2009)*. Merano, Italy: IEEE Xplore, Dec. 2009, pp. 468–473.

[6] M. Doostdar, S. Schiffer, and G. Lakemeyer, "Robust speech recognition for service robotics applications," in *Proceedings of the Int. RoboCup Symposium 2008 (RoboCup 2008)*, ser. Lecture Notes in Computer Science, vol. 5399. Suzhou, China: Springer, Jul. 2008, pp. 1–12.

[7] Y. Sasaki, S. Kagami, H. Mizoguchi, and T. Enomoto, "A predefined command recognition system using a ceiling microphone array in noisy housing environments," in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2008)*. Nice, France: IEEE Xplore, Sep. 2008, pp. 2178–2184.

[8] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Prentice Hall, 2009.

[9] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. (ICASSP 2006)*. Toulouse, France: IEEE Xplore, May 2006.

[10] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proceedings of the 2009 APSIPA Annual Summit and Conference (APSIPA ASC 2009)*. Sapporo, Japan: APSIPA, Oct. 2009, pp. 131–137.

[11] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney, "The RWTH Aachen University open source speech recognition system," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, Brighton, U.K., Sep. 2009, pp. 2111–2114.

[12] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "English broadcast news speech (HUB4)," Linguistic Data Consortium, Philadelphia, 1997.

[13] S. Wermter, G. Palm, and M. Elshaw, *Biomimetic Neural Learning for Intelligent Robots*. Springer, Heidelberg, 2005.

[14] H. Nakashima, H. Aghajan, and J. C. Augusto, *Handbook of Ambient Intelligence and Smart Environments*. Springer Publishing Company, Incorporated, 2009.

[15] D. van der Pol, J. Juola, L. Meesters, C. Weber, A. Yan, and S. Wermter, "Knowledgeable service robots for aging: Human robot interaction," KSERA consortium, Deliverable D3.1, October 2010.

[16] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "The NAO humanoid: A combination of performance and affordability," *CoRR*, 2008. [Online]. Available: http://arxiv.org/abs/0807.3223

[17] C. D. Manning and H. Schuetze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[18] H. Nakajima, K. Kikuchi, T. Daigo, Y. Kaneda, K. Nakadai, and Y. Hasegawa, "Real-time sound source orientation estimation using a 96 channel microphone array," in *Proceedings of the 2009 IEEE/RSJ Int. Conference on Intelligent Robots and Systems (IROS 2009)*. St. Louis, USA: IEEE Xplore, October 11-15 2009, pp. 676–683.

[19] T. Yoshida, K. Nakadai, and H. G. Okuno, "Two-layered audio-visual speech recognition for robots in noisy environments," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2010)*. Taipei, Taiwan: IEEE Xplore, October 18-22 2010, pp. 988–993.

On-line learning of sensorimotor transformations on a humanoid robot

Marco Antonelli*, Eris Chinellato*†, Angel P. del Pobil*‡
*Robotic Intelligence Laboratory, Jaume I University, 12071 Castellón de la Plana, Spain
Email: {antonell,pobil}@icc.uji.es
† Department of Electrical and Electronic Engineering, Imperial College London, UK
Email: e.chinellato@imperial.ac.uk
‡ Department of Interaction Science, Sungkyunkwan University, Seoul, South Korea

***Abstract*—In infant primates, the combination of looking and reaching to the same target is used to establish an implicit sensorimotor representation of the peripersonal space. This representation is created incrementally by linking together correlated signals. Also, such a map is not learned all at once, but following an order established by the temporal dependences between different modalities. Inspired by these observations we have developed a framework for building and maintaining an implicit sensorimotor map of the environment. In this work we present how this framework can be extended to allow the robot to update on-line the sensorimotor transformations among visual, oculomotor and arm-motor cues.**

I. Introduction

Manipulation tasks, such as reaching or grasping, require knowledge about the position of the target in the peripersonal space. In human beings, this information is obtained by different sensorimotor signals that have to be linked together in order to create an integrated representation of the environment [1]. These cues are represented in different frames of reference (FoR), as for examples retinotopic, head-centered and arm-centered. The brain encodes these transformations between one frame to another in populations of neurons that exhibit gain fields. Such neural populations are found in posterior parietal cortex (PPC) areas V6A and MIP. Both these areas receive their main input from V6 and project to the dorsal premotor cortex [2]–[5]. The information regarding eye position and gaze direction is employed by area V6A in order to estimate the position of surrounding objects and guide reaching movements towards them [2], [6], [7].

The computational model we propose for modeling the kind of sensorimotor transformations performed by the PPC [8] is based on radial basis function networks (RBFNs) [9]–[11]. Networks of suitable basis functions are able to naturally reproduce the gain-field effects often observed in parietal neurons [12] and are particularly suitable for maintaining sensorimotor associations [9].

Even if, in principle, a large neural network could learn all the necessary sensorimotor transformations, in practice, given the high dimensionality of the input space, this is not computationally feasible, not even for the brain itself. Our conceptual framework [8], and its first implementation on a humanoid robot [13], show that the problem can be tackled by decomposing the learning task into incremental phases, where some fundamental competences are learned before more advanced ones [14]. This task decomposition can be observed also in infant development, where a child learns to direct his gaze to an object before learning how to reach to it [15].

In the literature, several works have tackled the problem of the hand-eye coordination with the use of Self-Organizing Maps (SOM) [16]–[18], RBFNs [19], [20] or other types of neural networks [21], [22]. Hovewer, some major differences can be pointed out between the previous works and our approach. For example, Marjanovic at al. do not use stereo vision [19], while other approaches do not consider eye movements [16], [17], [20]. In general, the sensorimotor association is created directly between the visuo/oculomotor space and the arm-motor space without consider the advantage (and the biological plausibility) of using the visuo to oculomotor transformation [16]–[18], [20]. When this transformation is taken into account, either it is not learned on-line [20], [21] or it is not learned at all [22].

In our approach, the sensorimotor framework was built using three RBFNs: one for converting the visual position of a stimulus into an oculomotor position, and the other two for hand-eye movement coordination. The aim of this work is to extend the proposed sensorimotor framework [8], [13] to allow the robot to incrementally learn to gaze and reach objects in the peripersonal space. The system is based on two key concepts: *adaptation* and *task decomposition*. The described framework is implemented on an agent based architecture and tested on *Tombatossals* humanoid torso.

In this paper, we describe the sensorimotor framework (Section II) and how it can be updated during the exploration of the peripersonal space (Section III). Experiments (SectionIV show that the robot is able to learn the transformations between the different FoR and create, in such a way, an implicit mapping of the environment.

II. Encoding of the sensorimotor transformations

Saccades and reaching movements constitute the basic behaviors employed to build a sensorimotor representation of the peripersonal space. Such a representation is built incrementally, through subsequent, increasingly complex interactions. The learning sequence in our system is inspired by infant

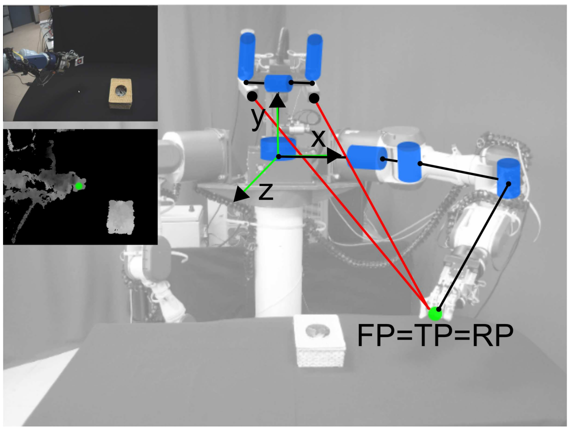


Fig. 1. Association of the oculomotor and arm-motor signals. When the robot move its hand (RP) and gaze towards the same point (FP), it can update its sensorimotor representation to locally reduce the sensorimotor transformation error for both the direct and inverse transformations. TP is the target point.

development [15]. As a first step, the system learns to associate retinal information and gaze direction (i.e. proprioceptive eye position). This can be done by successive foveations to salient points of the visual space. The subject looks around and focuses the eyes on certain stimuli, thus learning the association between retinal information and eye position.
Then, gaze direction is associated to arm position, as exemplified in Figure 1. This can be done by moving the arm randomly and following it with the gaze, so that each motor configuration of the arm joints is associated to a corresponding configuration of the eye motor control, and vice versa. This process allows learning a bidirectional link between different sensorimotor systems. Thus, the robot can look where its hand is but also reach the point in space he is looking at. The learning of the bidirectional oculomotor/arm-motor association between hand and eye positions (O↔A) requires that the robot is able to fixate its hand. Thus the visual to oculomotor association (V→O) need to be learned before.

The representation of the peripersonal space is maintained both by limb sensorimotor signals on the one hand and visual and oculomotor signals on the other hand. In this way, different representations of the same target can be maintained contextually, and used to act on the target if required [8]. The relations between these representations, i.e. sensorimotor transformations, are accessed and modified by a conjunction of gaze and reach movements to the same target (left and right sides of Fig. 2).
The location of a visual target is expressed with its position in the cyclopean visual field accompanied by information on binocular disparity. The output of the visual to oculomotor transformation is the correspondent potential vergence/version movement required to foveate the target (left side of Fig. 2). This transformation has been implemented with an RBF network, described below. The movement computed by the network, together with the current eye position, provide the head centered representation of the target.

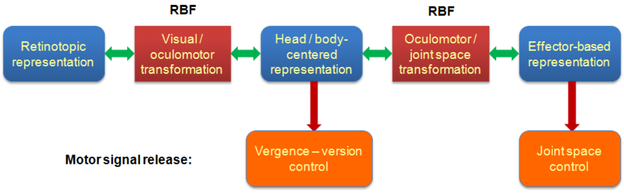

RBF
RBF
Retinotopic representation
Visual / oculomotor transformation
Head / body-centered representation
Oculomotor / joint space transformation
Effector-based representation
Motor signal release:
Vergence – version control
Joint space control

Fig. 2. Computational framework of the sensorimotor model. Two transformations permit to represent a stimulus contextually in visual, oculomotor and arm-motor frames of reference. These representations allow the robot to perform reaching and gazing movement on demand.

The right hand of Fig. 2 schema shows the oculomotor/arm-motor mapping, required to code arm movements. This mapping is modified by proprioceptive feedback, and allows reaching towards visual or remembered targets. The inverse mapping is also stored, so that the robot is able to look at its hand without the need of new visual processing. We use two populations of radial basis function neurons for mapping the vergence/version space (representing oculomotor neurons) into the arm joint space (representing arm-motor neurons) and vice versa.

III. On-line learning

A *radial basis function network* is a single-layer feed-forward neural network where the hidden layer is composed of a set of radial basis functions[1] that can differ for the center location or for the shape of the activation. The hidden units employ a non-linear transfer function on the input data. The output of the network is computed by means of a linear combination of the hidden units:

$$\mathbf{y} = \mathbf{h}(\mathbf{x}) \cdot \mathbf{W} \tag{1}$$

where $\mathbf{y}$ is the output vector, $\mathbf{h}$ is the activation of the hidden layer given the input $\mathbf{x}$ and $\mathbf{W}$ is the matrix of the weights. The weights can be learned by any of the standard iterative methods for neural networks, as the *delta rule*. The delta rule is a gradient descent learning rule that can be used to adapt the weights of a single-layer neural networks. The weight update equation is the following:

$$\Delta\mathbf{W} = \alpha \cdot \mathbf{h}(\mathbf{x})^T \cdot (\mathbf{t} - \mathbf{y}) \tag{2}$$

where $\mathbf{y}$ is the actual output of the network and $\mathbf{t}$ is the expected output. The parameter α is a small positive value called *learning rate* which can be constant or variable.
The application of the algorithm to a real robot raises the problem of how to calculate or estimate the error of the map. Usually, this error is provided by external agents as human users or other systems that control the environment. This type of learning is called supervised. However, given that our model is inspired to primate behavior, we want the robot itself to calculate or estimate the output error. We call this learning method *self-supervised* because the error term is provided by the robot itself.

[1] A radial basis function is a function whose value depends only on the distance of the input space from a point, called *center of activation*.

Algorithm 1 V→O: Learning with ground truth

```
$\mathbf{o}(t) \leftarrow$ oculomotor position (vergence/version)
$\mathbf{v}(t) \leftarrow$ stimulus position
Assuming the robot is fixating the stimulus: $\mathbf{v}(t) = 0$
$\Delta\mathbf{o}(t) \leftarrow$ random value
$\mathbf{o}(t+1) \leftarrow \mathbf{o}(t) + \Delta\mathbf{o}(t)$
Move head to position $\mathbf{o}(t+1)$
$\mathbf{v}(t+1) \leftarrow$ new stimulus position
$\Delta\mathbf{o}(t+1) \leftarrow T_{V \to O}(\mathbf{v}(t+1))$ (movement computed by the network)
$\mathbf{e}(t+1) \leftarrow -(\Delta\mathbf{o}(t) + \Delta\mathbf{o}(t+1))$
$\Delta\mathbf{W}(t+1) \leftarrow \alpha \cdot \mathbf{e}(t+1) \cdot \mathbf{H}(\mathbf{v}(t+1))$
```

Algorithm 2 V→O: Learning with linear approximation

```
$\mathbf{o}(t) \leftarrow$ oculomotor position (vergence/version)
$\mathbf{v}(t) \leftarrow$ stimulus position
$\Delta\mathbf{o}(t) \leftarrow T_{V \to O}(\mathbf{v}(t))$
$\mathbf{o}(t+1) \leftarrow \mathbf{o}(t) + \Delta\mathbf{o}(t)$
Move head to position $\mathbf{o}(t+1)$
$\mathbf{v}(t+1) \leftarrow$ new stimulus position
$\hat{\mathbf{v}}(t) \leftarrow \mathbf{v}(t) - \mathbf{v}(t+1)$
$\Delta\hat{\mathbf{o}}(t) \leftarrow T_{V \to O}(\hat{\mathbf{v}}(t))$
$\mathbf{e}(t) \leftarrow \Delta\mathbf{o}(t) - \Delta\hat{\mathbf{o}}(t)$
$\Delta\mathbf{W}(t) \leftarrow \alpha \cdot \mathbf{e}(t) \cdot \mathbf{H}(\hat{\mathbf{v}}(t))$
```

A. Visual to oculomotor transformation

For the visual to oculomotor (V→O) transformation, the application of the delta rule is not straightforward in the context of self-supervised learning. The visual to oculomotor transformation uses the retinotopic position of the stimulus to update the oculomotor position in order to keep the stimulus at the center of the retina. A stimulus position different from zero indicates that the gaze position needs to be adjusted.

However, the error is observed on the retina, i.e., in the input space, while the application of the delta rule requires the error in the output space, i.e. in terms of vergence and version angles. This error can not be calculated exactly because its computation requires the knowledge of the transformation V→O itself, that is what the robot tries to learn. Therefore, the output error has to be known a-priori or it has to be estimated from the input.

1) Learning with ground truth: A possible solution to the described problem is to provide the robot with an ad-hoc behavior for the training phase. When the robot is fixating a stimulus, it can perform a random ocular movement towards a different location. After the movement, the position of the visual stimulus can be used to estimate the vergence/version components of the movement that would allow to return to the starting point. If there were no errors, the sum of the two movements would make the robot fixate again the original stimulus. The actual residual error is used to train the network according to Eq. 2, as explained in Algorithm 1. This algorithm has the advantage of training the network with the exact movement (ground truth) that it is expected to perform. Also, this solution allows learning the visual to oculomotor mapping without any previous knowledge. On the other hand, the adaption process can be applied only during the training phase and not in the normal behavior of the robot, unless the visual environment is so rich that visual stimuli are always available close to the fixation point.

2) Learning with linear approximation: Another way to estimate the error is to consider that the transformation can be approximated locally by a linear function, and exploit the property $f(v_1) - f(v_2) = f(v_1 - v_2)$. In this case, the executed movement $\Delta\mathbf{o}(t) = T_{V \to O}(\mathbf{x}(t))$, that brings the stimulus position from $\mathbf{v}(t)$ to $\mathbf{v}(t+1)$, can be seen as a target movement for a hypothetical stimulus $\hat{\mathbf{v}}(t) = \mathbf{v}(t) - \mathbf{v}(t+1)$ (see Algorithm 2). Ideally, stimulus $\hat{\mathbf{v}}(t)$ should correspond to $\mathbf{v}(t)$ but practically they differ by $\mathbf{v}(t+1)$ which is small but not zero. This method exploits the linear assumption to adapt the network by using the performed movement. Even if V→O is not linear, it is monotonic under the assumption of having convex lenses, i.e. increasing the retinotopic distance of the stimulus requires a greater motor movement. Thus, if the learning rate is small enough, the linear approximation allows approaching the function from the right direction. Moreover, the smaller the error of the performed movement is, the closer the functioning of the algorithm is to the *ground truth method*. It is important to note that even if the adaptation procedure is based on a linear assumption, the learned transformation is not linear. In this method, learning is applied to the current activation of the hidden layer, but the system has to memorize the previous input-output pair in order to compute $\hat{\mathbf{v}}(t)$ and the error $\mathbf{e}(t)$.

B. Oculomotor and arm-motor transformations

As in the V→O case, on-line training of the robot is achieved by means of the *delta rule*. In order to create the eye-arm association it is necessary that the robot fixates a point that it has reached with the hand. This can be done using vision as master signal, under the assumption that the vision system is able to recognize the robot hand. We addressed this problem by placing a marker on the robot's hand. Once the robot is able to fixate the hand, the application of the delta rule becomes straightforward, as the input $\mathbf{x}$ and the target $\mathbf{t}$ are given by oculomotor and arm-motor positions, while $\mathbf{y}$ can be obtained by applying the direct and inverse networks to the input.

The transformations are learned during a free exploration of the peripersonal space (Algorithm 3). This phase consists of random arm movements and subsequent saccades towards the final hand position, which allows learning the transformation from joint space to oculomotor space and vice versa. To generate the training points, a correct gazing behavior is required. Therefore, the second transformation can only be trained after the first one has a reasonably accurate performance.

After the training phase, the robot can perform its normal behavior, and work in the goal-oriented exploration mode, where a target object in space has to be foveated and reached.

Algorithm 3 O↔A: Exploration-based phase

```
Move randomly the arm to position $\mathbf{a}(t)$
Find the arm in the visual space
Fixate the arm using $T_{V\to O}$
$\mathbf{o}(t) \leftarrow$ head position
$\hat{\mathbf{o}}(t) \leftarrow T_{A\to O}(\mathbf{a}(t))$
$\Delta\mathbf{W}_{A\to O}(t) \leftarrow \alpha \cdot (\mathbf{o}(t) - \hat{\mathbf{o}}(t)) \cdot \mathbf{H}(\mathbf{a}(t))$
$\hat{\mathbf{a}}(t) \leftarrow T_{O\to A}(\mathbf{o}(t))$
$\Delta\mathbf{W}_{O\to A}(t) \leftarrow \alpha \cdot (\mathbf{a}(t) - \hat{\mathbf{a}}(t)) \cdot \mathbf{H}(\mathbf{o}(t))$
```

IV. Experimental results

A. Hardware and software architecture

Tombatossals is a humanoid torso endowed with a pan-tilt-vergence stereo head and two multi-joint arms (Figure 1). Both wrists mount force/torque sensors that can be used to stop arm movements in case of collisions. The head mounts two cameras with a resolution of 1024x768 pixels that can acquire color images at 30 Hz. The baseline between cameras is 270mm. Both the head and the arms are equipped with encoders that allow to gain access to the motor position with high precision. During the training of the RBFNs the left hand was equipped with a marker that can be recognized by means of the ARToolkit library. This marker can be used by the robot to identify its own hand, and it represents also a visual stimulus that can be moved very precisely in the three dimensional space. We employed three degrees of freedom (d.o.f.) for both the head (tilt, left pan and right pan) and the left arm (first, second and forth joint).

The control system of the robot consists of a collection of modules that can exchange data among them through software ports. The connection among modules was managed using YARP (*Yet Another Robotic Platform*) [23]. The system architecture is shown in Figure 3. Modules can be classified according to their nature (left: device interfaces; center: sensorimotor transformations; right: memories) or to their target domain (yellow: image domain; blue: gaze domain; red: arm-motor domain). The two exceptions are the visuomotor memory, which store visual features together with their oculomotor position, and the O↔A module, which manages both oculo- and arm-motor FoR. Finally, the *task manager* takes care of controlling the data flow between the modules in order to accomplish the desired task, such as reaching an object, looking at the hand and so on.

B. Results

In this section we describe the experiments conducted in order to validate the framework. The structure and parameters of the networks were chosen using a heuristic search on a simulated model of the robot. We decided to employ fixed centers, whose receptive fields can not move according to the input data, favoring biological plausibility over potentially better performance [8].

The centers of the RBFN that computes the V→O transformation were distributed according to a retinotopic-like criterion (input to V6A is, at least partly, retinotopic), following a logarithmic distribution. A logarithmic organization of the neural receptive fields is suitable for modeling foveal magnification, while affording greater sensitivity to disparity, as observed in the primate visual cortex [24]. A reasonably good performance was achieved by distributing the neuron centers on a 7x7x7 neural lattice in the cyclopean (x,y) and disparity space. Neural activation functions were Gaussians with a spread proportional to the distance between adjacent centers.

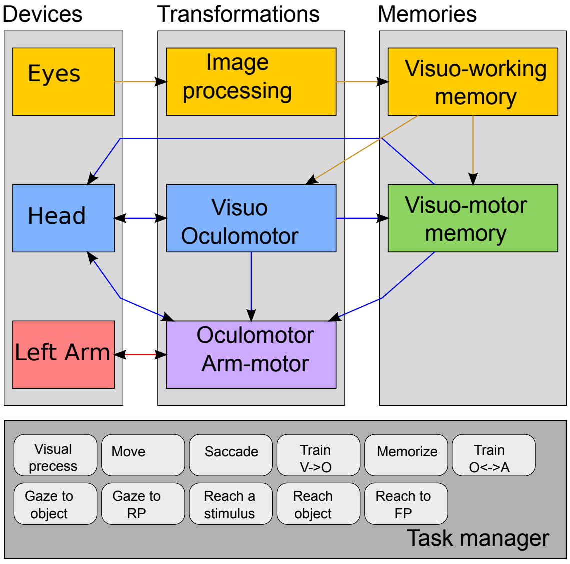


Fig. 3. Cognitive architecture of robot, with modules classified according to role and domain.

The centers of the O↔A networks were distributed uniformly on a 7x7x7 lattice in the input space, i.e. the vergence - version(x,y) space for the direct transformation, and the arm joint space for the inverse one. Again, the activation functions were Gaussians, but the spread was the same for all neurons. The weights of the networks were not learned from scratch, but were bootstrapped using those provided by the simulated model of the robot. However, due to the approximation of the latter, the weights had to be adapted to fit the parameters of the robot's body. At each interaction with the environment, the robot autonomously estimated the local error of the sensorimotor transformations and corrected the weights using the *delta rule*, as described in the previous section. The value of α, the learning rate, is of critical importance. If it is too high, the robot will overfit the training points and will perform poorly in presence of noise. On the other hand, a small value of α produces a slow convergence of the network. The learning rate was set empirically to 0.01 for both V→O and O↔A transformations.

The two parts of Fig. 2 were learned in two different phases. In the first phase the robot learns to properly foveate the target stimulus. In the second phase, the robot foveates its hand in order to align eye and arm position for learning the O↔A transformation.

TABLE I

VISUAL TO OCULOMOTOR NETWORK, ERROR MEANS AND STANDARD DEVIATIONS [DEGREES].

Simulation		Grd. truth		Linear approx.		SSE	
μ	σ	μ	σ	μ	σ	μ	σ
0.91	0.57	0.40	0.26	0.42	0.34	0.02	0.014

1) Visual to oculomotor transformation: The V→O learning process (Section III-A) can be summarized as follows. The vision system locates saliency points in the left and right images. The task manager selects the target stimulus and executes a sequence of saccades until the target is foveated. After each saccade the new position of the stimulus is used to train the V→O network. Once the target is foveated, the head is moved randomly and another stimulus is chosen as target. The choice of the target can be random or can be driven by the information stored in *visuo-motor memory* [14]. In the current setup, markers were used instead of real objects in order to minimize visual errors. The visual target was considered foveated when its distance from the centers of both images was smaller then 10 pixels. The foveating behavior was performed 500 times. The *linear approximation method* (see Algorithm 2) was used to estimate the error of the network and its performance was compared with the *ground truth method* (Algorithm 1).

At the end of the experiment, the performance of the system was evaluated. To that aim, the head was placed in an initial position where it fixated the visual marker. Thus, the head was moved on a lattice on the vergence/version space and the visual position of the marker was recorded. The maximum amplitude of the movement was $\pm 0.3\,rad$ for both vergence and version (x, y) angles. The pairs *visual position - performed movement* were used as testing set for the system. Table I reports the errors of the networks with the weights obtained during the simulations, and successively adapted with the *delta rule*, using both the *ground truth* and the *linear approximation* methods. The error of the simulated model is relatively small but the online adaptation mechanism allows improving the performance of the system by a factor of 2. The linear approximation method provides results close to those obtained with the ground truth and allows the system to adjust the weights during the normal exploration of the environment. Results obtained with a Sum of Squared Error (SSE) batch learning method are also displayed, and will be commented upon below.

An even better analysis of the error can be performed by plotting the mean of the error as a function of the distance from the fovea (Fig. 4). The weights found in the simulation (black line) produce an error that depends on the distance. This is due to the approximation of the model that did not take into account the focal distortion of the lens. The delta rule allows correcting for this error either using the ground truth method (red line) or the linear approximation one (blue line). On the periphery, the correction factor is reduced due to the smaller number of training points for that region. However, the network is potentially able to approximate the right transformation as shown by the result obtained training the network by minimizing the sum of squared error (SSE, green line, see also Table I). It is not surprising that the minimization of the SSE outperforms the on-line method as it uses the global information of the training set, while the delta rule uses only the information of the new observed point. On the other hand, the delta rule does not require to store the whole dataset to learn the sensorimotor transformation, which is rather implausible from a biological point of view.

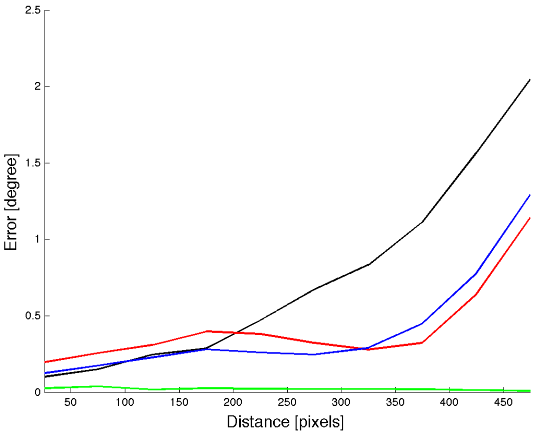


Fig. 4. V→O: Mean error as a function of the distance from the fovea.

2) Oculomotor⇔arm-motor transformations: The networks that compute the direct and inverse transformations from oculomotor to arm-motor signals require a training set composed of pairs of eye and arm positions. The training set was generated by moving the arm randomly.

Once the arm movement is completed, the *task manager* releases a sequence of saccades to foveate the marker placed on the hand. However, this procedure generates training points that are inherently noisy, due to small imprecisions in identifying the center of the marker in the two images and due to the size of the “fovea”, which was set to 10 pixels. The on-line training was iterated for 791 trials. The misalignment between the oculo- and arm-motor position was estimated to be on average 2.08 pixels, as viewed by the robot’s cameras. For each transformation, Table II reports the mean and standard deviation of the error of the RBFNs bootstrapped from the weights learned during the simulations and successively adapted on-line. Again, the table shows also the results obtained with the weights learned using the minimization of SSE to show the potential performances of the networks.

The results indicate that the difference between the simulated model and the robot is remarkable. The weights obtained during the simulations are not good enough for the robot to execute the reaching task. Therefore, the on-line adaptation is necessary. By training the network with 791 points, the system was able to improve its performance by a factor of 10 for

TABLE II
OCULOMOTOR⇔ARM-MOTOR NETWORKS, ERROR MEANS AND STANDARD DEVIATIONS [DEGREES].

Transf.	Simulation		Delta Rule		SSE	
	μ	σ	μ	σ	μ	σ
A→O	3.98	0.89	0.36	0.26	0.33	0.25
O→A	13.20	6.20	3.89	3.05	0.83	0.84

A→O and of 3 for O→A. The off-line training performed by minimizing the SSE method shows that the performance of the O→A can improve with more training points while the A→O seems to have reached its best performance. In general, as observed during the simulations, O→A is harder to learn than A→O. This should not be surprising, as the former computes the inverse kinematics while the latter computes the direct kinematics.

V. Conclusions

In this work we have described how the transformations necessary to maintain a sensorimotor representation of the environment can be updated by the robot itself during the exploration of the peripersonal space. First, the robot learns to associate visual cues with the oculomotor movements (V→O) and then produces arm movements to learn the association between the fixation point and the reaching point (O↔A). During such exploration, the weights of the RBFNs that implement the V→O and O↔A transformations are refined using the delta rule. This learning process is the normal behavior of the agent, constituting the most fundamental component of its basic capability of interacting with the world, and contextually updating its representation of it.

The proposed framework works on the assumption that problems such as object recognition and disparity extraction have already been solved. We are currently working on improving the visual system by introducing models, inspired by the primate visual cortex, which will allow the system to deal with real objects instead of visual markers.

VI. Acknowledgements

This work was supported by the European Commissions Seventh Framework Programme FP7/2007-2013, under Grant 217077 (EYESHOTS project), by the Ministerio de Ciencia y Innovación (FPI Grant BES-2009-027151), by the Fundaciò Caixa-Castello-Bancaixa (P1-1B2008-51), and by the World Class University Program through the National Research Foundation of Korea, funded by the Ministry of Education, Science, and Technology (Grant R31-2008-000-10062-0).

References

[1] E. Chinellato, B. J. Grzyb, N. Marzocchi, A. Bosco, P. Fattori, and A. P. del Pobil, "The dorso-medial visual stream: From neural activation to sensorimotor interaction," *Neurocomputing*, vol. 74, no. 8, pp. 1203 – 1212, 2011.

[2] C. Galletti, D. Kutz, M. Gamberini, R. Breveglieri, and P. Fattori, "Role of the medial parieto-occipital cortex in the control of reaching and grasping movements," *Experimental Brain Research*, vol. 153, no. 2, pp. 158–170, 2003.

[3] P. Fattori, M. Gamberini, D. Kutz, and C. Galletti, "Arm-reaching neurons in the parietal area V6A of the macaque monkey," *European Journal of Neuroscience*, vol. 13, pp. 2309–2313, 2001.

[4] P. Dechent and J. Frahm, "Characterization of the human visual V6 complex by functional magnetic resonance imaging," *European Journal of Neuroscience*, vol. 17, no. 10, pp. 2201–2211, 2003.

[5] R. Caminiti, S. Ferraina, and A. B. Mayer, "Visuomotor transformations: early cortical mechanisms of reaching," *Current opinion in neurobiology*, vol. 8, no. 6, pp. 753—761, Jan 1998.

[6] P. Fattori, D. Kutz, R. Breveglieri, N. Marzocchi, and C. Galletti, "Spatial tuning of reaching activity in the medial parieto-occipital cortex (area V6A) of macaque monkey," *European Journal of Neuroscience*, vol. 22, no. 4, pp. 956–972, 2005.

[7] N. Marzocchi, R. Breveglieri, C. Galletti, and P. Fattori, "Reaching activity in parietal area V6A of macaque: eye influence on arm activity or retinocentric coding of reaching movements?" *European Journal of Neuroscience*, vol. 27, no. 3, pp. 775–789, 2008.

[8] E. Chinellato, M. Antonelli, B. Grzyb, and A. delPobil, "Implicit sensorimotor mapping of the peripersonal space by gazing and reaching," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, pp. 45–53, Jan 2011.

[9] A. Pouget and T. J. Sejnowski, "Spatial transformations in the parietal cortex using basis functions," *Journal of Cognitive Neuroscience*, vol. 9, no. 2, pp. 222–237, Mar 1997.

[10] A. Pouget and L. Snyder, "Computational approaches to sensorimotor transformations," *Nature neuroscience*, vol. 3, pp. 1192–1198, 2000.

[11] A. Pouget, S. Deneve, and J. Duhamel, "A computational perspective on the neural basis of multisensory spatial representations," *Nature Reviews Neuroscience*, vol. 3, no. 9, pp. 741–747, 2002.

[12] E. Salinas and P. Thier, "Gain modulation: A major meeting report computational principle of the central nervous system," *Neuron*, vol. 27, pp. 15–21, 2000.

[13] M. Antonelli, E. Chinellato, and A. delPobil, "Implicit mapping of the peripersonal space of a humanoid robot," *IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain*, pp. 1–8, Feb 2011.

[14] S. McBride, J. Law, and M. Lee, "Integration of active vision and reaching from a developmental robotics perspective," *Autonomous Mental Development, IEEE Transactions on*, vol. 2, no. 4, pp. 355—366, 2010.

[15] K. E. Adolph and A. S. Joh., "Motor development: How infants get into the act," in *Introduction to Infant Development*, A. Slater and M. Lewis, Eds. Oxford University Press, 2007, pp. 63–80.

[16] T. M. Martinetz, H. J. Ritter, and K. J. Schulten, "Three-dimensional neural net for learning visuomotor coordination of a robot arm," *IEEE T Neural Networ*, vol. 1, no. 1, pp. 131–136, Mar. 1990.

[17] M. Jones and D. Vernon, "Using neural networks to learn hand-eye coordination," *Neural Computing and Applications*, vol. 2, no. 1, pp. 2–12, 1994.

[18] S. Fuke, M. Ogino, and M. Asada, "Acquisition of the head-centered peri-personal spatial representation found in vip neuron," *Autonomous Mental Development, IEEE Transactions on*, vol. 1, no. 2, pp. 131–140, 2009.

[19] M. Marjanovic, B. Scassellati, and M. Williamson, "Self-taught visually guided pointing for a humanoid robot," *From Animals to Animats 4: Proc. Fourth Int l Conf. Simulation of Adaptive Behavior*, pp. 35—44, 1996.

[20] G. Sun and B. Scassellati, "A fast and efficient model for learning to reach," *International Journal of Humanoid Robotics*, vol. 2, no. 4, pp. 391–414, 2005.

[21] W. Schenck, H. Hoffmann, and R. Möller, "Learning internal models for eye-hand coordination in reaching and grasping," *Proceedings of the European Cognitive Science Conference, Osnabrück, Germany*, p. 289, 2003.

[22] F. Nori, L. Natale, G. Sandini, and G. Metta, "Autonomous learning of 3d reaching in a humanoid robot," *IEEE/RSJ IROS, International Conference on Intelligent Robots and Systems*, pp. 1142–1147, 2007.

[23] G. Metta, P. Fitzpatrick, and L. Natale, "Yarp: Yet another robot platform," *International Journal of Advanced Robotics Systems, special issue on Software Development and Integration in Robotics*, vol. 3, no. 1, 2006.

[24] G. Poggio, F. Gonzalez, and F. Krause, "Stereoscopic mechanisms in monkey visual cortex: binocular correlation and disparity selectivity," *The Journal of neuroscience*, vol. 8, no. 12, p. 4531, 1988.

Pos-4

Our robot is more human than yours: Effects of group membership on anthropomorphic judgments of social robots

Friederike Eyssel and Dieta Kuchenbrandt

Abstract— **In an experiment we empirically tested effects of social category membership on the evaluation of humanoid robots. To do so, German participants rated a humanoid robot that either belonged to their German ingroup or to a national outgroup with regard to anthropomorphism. That is, we asked participants to provide mind attribution, warmth ratings, perceptions of shared reality with the robot and evaluation of robot design. We operationalized alleged robot group membership by manipulating the robots name and by means of providing additional information about where the robot prototype had ostensibly been developed. This resulted in a "German" and a "Turkish" robot prototype that participants had to evaluate In line with social psychological findings from intergroup research, we predicted that participants would anthropomorphize the German robot more than the Turkish robot. That is, we hypothesized that they would attribute more mind and experience more shared reality with the ingroup (vs. the outgroup) robot. Finally, we predicted that participants would prefer the ingroup over the outgroup robot even with regard to its design. Our results support all experimental hypotheses: Generally, on all dependent measures, participants anthropomorphized the German robot significantly more strongly than the Turkish prototype. Analogously, they experienced more shared reality with it and found its design more aesthetically pleasing than compared to the outgroup robot. Implications of these results are discussed.**

I. Introduction

Research by Epley and his colleagues [1] has shown that people tend to ascribe typically human attributes (e., g., traits, intentions, mind, emotions) even to nonhuman entities, such as technical gadgets, gods, pets or robots. That is, people imbue "the imagined or real behavior of nonhuman agents with humanlike characteristics, motivations, intentions, and emotions" (pp. 864-865).

Epley and colleagues have emphasized the fact that anthropomorphism should not be equated with or reduced to attributions of "lifelikeness", "naturalness", "humanlikeness" [2], or mere animism [3, 4]. Instead, they have proposed that anthropomorphic inferences go well beyond merely observable behavior [1, 5].

In their theoretical framework, these researchers have introduced three core determinants of anthropomorphism, two motivational factors and a cognitive one. The two motivational determinants of anthropomorphism are *sociality motivation* and *effectance motivation* [1, 5]:

Sociality motivation refers to the desire for social connection and a sense of belonging. Epley et al. [1] have observed that people who lack social connection appear to anthropomorphize pets, inanimate agents or other nonhuman entities more strongly than those who are socially connected. *Effectance motivation*, on the other hand describes the desire to master our environment and to interact with it as competent social actors [6]. Finally, they propose that a cognitive determinant drives people's anthropomorphic judgments. That is, when confronted with an unfamiliar agent, people draw on knowledge about themselves and the category of "humans". This way, *elicited agent knowledge* is activated.

II. Related Work

Social psychological research on intergroup processes has shown that people readily categorize others in terms of their age, sex or ethnicity [7, 8]. That is, people like to think about others in terms of "us" vs. "them". From this results the robust finding that people tend to evaluate or behave more positively toward members of their own social than towards a social group outgroup, and this is even the case if group memberships have been assigned randomly (e.g., ostensibly based on an individual's preference for something). This phenomenon is called "ingroup favoritism" [for an overview of social psychological intergroup theories, refer to 9, 10].

Would such an effect even go beyond judgments of fellow social group members and extend to products or technical devices such as robots? That is, would group membership of

This research was funded by the DFG-Grant COE 277.

F. Eyssel is with the Center of Excellence in Cognitive Interaction Technology, University of Bielefeld, Germany (phone: +49-521-106-12044; email: feyssel@cit-ec.uni-bielefeld.de).

D. Kuchenbrandt is with the Center of Excellence in Cognitive Interaction Technology, University of Bielefeld, Germany (email: dkuchenb@cit-ec.uni-bielefeld.de)

a robot prototype affect users' perception and anthropomorphic inferences about the robot?

The present research sought to provide answers to these research questions: Specifically, we investigated whether participants would a) attribute more typically human attributes, b) feel psychologically closer and, c) even rate the design of the robot more positively when the robot allegedly belonged to their own social group vs. a social outgroup.

With regard to the issue of the operationalization of ingroup vs. outgroup membership of the robot, we relied on a simple manipulation that will be described in detail in Section C.

As documented in our previous research on effects of social categorization in robots [11], participants readily applied gender categories and the respective stereotypes that go along with them to humanoid robot that appeared male vs. female based on its hairstyle. That is, in line with Epley's notion of elicited agent knowledge that serves as a heuristic cue when judging an unfamiliar entity, hairstyle as a visual gender cue was sufficient to trigger subsequent stereotypical evaluations of robot prototypes. Thus, participants used the knowledge that was most readily accessible about the categories "male" or "female", and this in turn biased their judgments about the "gendered" robots. In a similar vein, [12] and [13] have demonstrated that such categorization effects are not restricted to the gender category: To illustrate, Powers and colleagues [12, 13] have shown that a robot's behavior, its tone of voice or its appearance (e. g., in terms of babyfacedness) constituted key cues for subsequent judgments about the robots and its "persona". These authors have suggested that people do not "approach the robot *tabula rasa* but rather develop a default model of the robot's knowledge" [12, p. 159].

Moreover, Lee and colleagues [14] have demonstrated that the appearance and language of a robot affected the perceived knowledge of the robot when it was described in terms of its ethnicity, either ostensibly stemming from China or the US. Taking this information into account, participants assumed that the Chinese-speaking robot of "Asian ethnicity" would know more about landmarks in China compared to an "American" robot that had been developed in the United States. This exemplifies that participants apparently used accessible social category information to form their judgments about the robot prototypes. In other words, social categorization processes influenced perceptions of the robot types depending on their alleged "ethnicity".

Our goal was to broaden the theoretical and empirical scope of the previous work by testing whether participants would show biased anthropopmorphic inferences about an ingroup vs. outgroup robot. This notion is based on research on intergroup processes and discrimination [15-17]. In this work, it has repeatedly been shown that people reserve typically human traits for their own group, whereas they are hesitant to attribute these features to another social group [15]. In a similar vein, consumer research has found that people tend to integrate brands less into their self-concept if the brand represents a dissociative (i.e. disliked) reference group [18, 19]. To illustrate, White and Dahl [19] have shown that consumers avoided brands associated with a negatively viewed outgroup more than a brand that was associated with outgroups in general. Taken together, previous research has provided evidence for the fact that social category information matters, independent of whether the social category had been activated in a subtle or more direct manner. Would social category membership of a robot thus influence the extent to which participants anthropomorphized it? To shed light on this research question, we conducted an experiment.

III. Method

A. Participants, Design, and Procedure

Participants were 78 German university students (37 male, 40 female, one person did not report his/her gender), with a mean age of 23.27 years (*SD* = 3.29). They were on average in their 4th semester of study (*SD* = 3.46). Participants were randomly assigned to one of two experimental conditions. That is, they either evaluated an ostensibly German robot or a Turkish robot with regard to anthropomorphic inferences.

As a cover story, participants were asked to take part in a survey study on the evaluation of an allegedly newly developed robot. They were told that with their ratings, they would help to improve the prototype before the robot would launch the market.

Thus, participants completed a questionnaire package that was used to measure participants' anthropomorphic inferences about Flobi [20]. Finally, they were reimbursed, debriefed and dismissed.

B. Hardware

In the present research, we used the social robot Flobi as a research platform [20]. Flobi's head has 18 degrees of freedom, in order to convey emotional states, such as happiness, sadness, fear, surprise, and anger. Two actuators rotate the eyebrows, three actuators move both eyes, four actuators serve to move the upper and lower eyelids, three actuators move the neck, and finally, six actuators are used to animate Flobi's lips. By means of four LEDs, red or white light can be projected onto Flobi's cheek surfaces to indicate either shame or health [see 21 for details].

Importantly, since the present research sought to focus on the effects of group membership on the evaluation of the robot platform, we only showed pictures of the robot target to participants and no actual human-robot interaction took place. Figure 1 shows the picture we used to familiarize participants with the robot target.

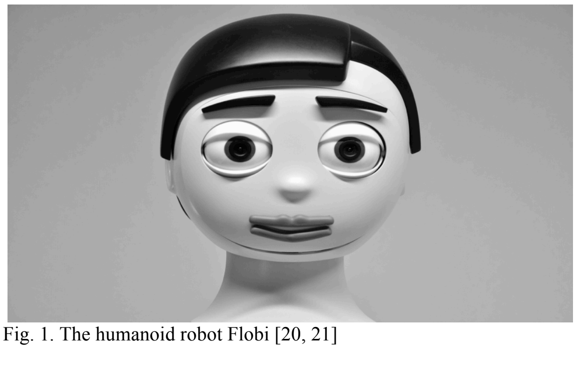


Fig. 1. The humanoid robot Flobi [20, 21]

C. Experimental Manipulation

Because our aim was to test effects of social categorization in the context of robots, we operationalized the robot's group membership by varying two aspects: On the one hand, we manipulated group membership by selecting a German vs. a Turkish first name for the robot prototype. That is, in one condition, participants learned that we would be interested in their evaluation of the newly developed robot "ARMIN" (a German first name), whereas the other half of participants received the same instructions, but they were asked to provide evaluations of the robot "ARMAN" (a Turkish first name). Secondly, to make group membership of both robots more salient, partipants were also informed that "ARMIN" had allegedly been developed at a German university, whereas "ARMAN" had ostensibly been developed at a Turkish university. The German-Turkish intergroup context was chosen because data were collected in Germany, where Turks represent the biggest socially relevant minority group.

D. Dependent Variables

To collect participants' responses regarding the dependent measures we used 7-point Likert scales. For subsequent data analyses, average scores were computed to form indices of the relevant constructs, with higher values reflecting greater agreement with the assessed dimension.

First, participants were asked to report how interpersonally warm they rated the robot prototype [22]. Warmth was assessed using five traits (helpful, sensitive, polite, generous, humble). These items were averaged to form a reliable index of *Warmth*, Cronbach's α = .81.

Second, we measured the extent to which participants attributed mind to the robot. We did so by asking them to rate the robot prototype with regard to its intelligence and eight further items: For instance "To what extent is the robot capable of feeling hungry / joy / pain / fear?"; "To what extent is the robot capable of hoping for things?"; How likely is it that the robot has a personality?"; "To what extent is the robot capable of being aware of things?" " How likely is it that the robot has a soul?". These items were adapted from Gray, Gray and Wegner's mind survey [23] and translated into German for the purpose of this experiments. The items formed a highly reliable *Mind* scale with Cronbach's α = .89.

Third, we administered four items to assess participants' degree of shared reality with the robot [24]. The *Shared Reality* index reflects perceptions of similarity and experienced psychological closeness to the robot. Furthermore, it covers aspects of human-robot acceptance, because participants had to indicate how much they would like to talk to the robot and to what extent they would be willing to live with it. The index of shared reality with the robot was reliable, given Cronbach's α = .74.

Finally, we were interested in a more subtle measure of favoritism for the ingroup robot. Therefore, we asked participants two questions related to their *Aesthetic Preferences*. The items read: "To what extent do you think the robot's design is well-developed?" and "Do you think the robot's design is aesthetically pleasing." This measure proved highly reliable, Cronbach's α = .87

IV. RESULTS

To test the experimental hypothesis that the ingroup robot would be evaluated more favorably than the outgroup robot, we conducted *t*-tests on the focal dependent measures. In line with our predictions, we found that on all dependent measures, participants evaluated the ingroup robot "ARMIN" that had been developed at a German university more favorably than the outgroup robot "ARMAN" that had ostensibly been developed at a Turkish university.
Figure 2 depicts our findings:

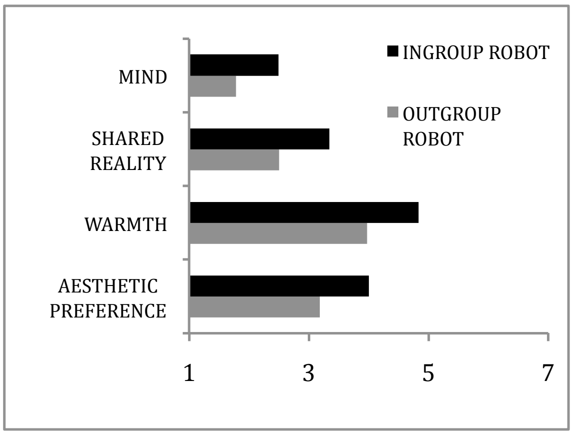


Fig. 2. Mean attribution of anthropomorphism and judgments of aesthetic preference as a function of *Robot Type*.

Figure 2 illustrates that participants anthropomorphized the German robot "ARMIN" significantly more than the Turkish equivalent named "ARMAN". This pattern of results was obtained on various dimensions: For example, participants perceived "ARMIN" as warmer (*M* = 4.84, *SD* = 1.15) than "ARMAN" (*M* = 3.97, *SD* = 1.31), *t*(76) = 3.04, *p* = .003. Most importantly, participants attributed more mind to the ingroup robot "ARMIN" (*M* = 2.65, *SD* = 1.40) than to the outgroup robot "ARMAN" (*M* = 1.99, *SD* = 0.72), *t*(76) = 2.75, *p* = .007. Going beyond mere judgments of the robot's sociality and mind, participants also reported significantly more shared reality with the ingroup robot (*M* = 3.34, *SD* =1.39) than with the outgroup robot (*M* = 2.51, *SD* = 1.11), *t*(76) = 2.95 , *p* < .001. With regard to participants' aesthetic preference, results show that they clearly favored the design of the ingroup robot (*M* = 4.00, *SD* = 1.62) over the outgroup robot (*M* = 3.18, *SD* = 1.48). This mean difference, too, was statistically significant, *t*(76) = 2.33, *p* = .02, despite the fact that both groups had been presented with exactly the same pictures of the allegedly new robot protoypes.

V. DISCUSSION

Our research experimentally investigated processes of social categorization in the context of social robotics. From findings on human-human intergroup research we know that commonly, individuals prefer their own group over another social group. That is, people tend to attribute more positive traits to their own social group relative to a social outgroup. Similar effects have also been observed with regard to typically human attributes. For example, Haslam and colleagues [16] have shown that people ascribe more typically human traits to themselves than to others, a phenomenon the authors coined “self-humanization”. We were interested in finding out whether participants would also ascribe more typically human attributes (e. g., warmth, mind) to a robot that ostensibly was “part” of their ingroup vs. an outgroup. We chose a German-Turkish inter-“robot” setting to test this assumption. That is, simply by labeling the robot target differently, an “intergroup” mindset was activated and participants biased their judgments of the robot prototypes accordingly. In other words, participants clearly evaluated one and the same robot prototype more positively, only because it had allegedly been developed in their home country and carried a name that signified ingroup membership. Because the current research was conducted with German participants only to establish and activate a German-Turkish intergroup context, it would be interesting to conduct the same study with Turkish participants. If ingroup favoritism were the underlying process, analogous results should be obtained in a Turkish sample. However, because it proved difficult to get access to Turkish-speaking participants, only data from a German convenience sample were presented here to make a case for social categorization effects of group membership using social robots. Critics might argue that Germany is better known for its engineering science and technological advancement as compared to a country like Turkey and this might explain the overall bias in favor of the German robot “ARMIN”.

However, we would counterargue that in the present study, no objective measures of the perceived technological standards in Germany vs. Turkey were assessed. That is, participants did not evaluate the products with regard to functionality and technical development. With regard to objective criteria including technical advancement, the argument could have been plausible. Interestingly, however, our dependent measures did by no means assess participants’ impression of the technical state of the art of the robot prototypes. Rather, the measures of anthropomorphism that were focal in the present research included judgments of sociality, mind attribution, perceived shared reality with the robots and finally, their ratings of the design.

On all dependent measures, a similar pattern of results emerged: The German robot “ARMIN” consistently and significantly outperformed the Turkish robot “ARMAN”. These results can be interpreted and explained from different angles: On the one hand, our findings are in line with Epley and colleagues’ [1, 5] notion of *elicited agent knowledge*, show-ing that the group membership manipulation used in the present experiment obviously triggered relatively more favorable judgments of the robot that represented participants’ ingroup. The heuristic thinking underlying such evaluations could be summarized as “whatever belongs to my group is better” – in the case of our experiment, this apparently mindless and automatic reasoning extended to robot platforms [25]. Moreover, the present findings can easily be reconciled with previous work on the dehumanization of social outgroups [15-17]. As research in the context of human-human intergroup contexts has repeatedly and robustly shown, in comparison to one’s own reference group, social outgroups are commonly perceived less human. Analogous patterns were obtained for the outgroup robot with regard to ratings of anthropomorphism and overall evaluation.

Besides the theoretical impact of the present results, they also bear clear practical significance: That is, our data indicate that with regard to user-centered robotics, developers should carefully reflect upon the naming of their product. In keeping with [18] and [19], marketers should bear in mind attributes of the consumers they want to target, because an association of the product with them and their ingroup affects subsequent product evaluations. According to a human-centered robotics approach, we would suggest to tailor the robot to the respective user group. Taking into account the current developments in technology and social robotics, social robots will sooner or later accompany humans in every day life. Our results show that perceived closeness to and shared reality with a robot depended on whether the robot was perceived in terms of an ingroup member. As such, the robot could be interpreted as a a self-extension and was subsequently rated much more favorably than the respective outgroup counterpart. Thus, if the robot depicted in Fig. 1 should ever launch the market, we would recommend to activate ingroup associations in the consumers, either by means of the product name or by emphasizing where the product originated. Future research is underway to examine whether the present findings generalize to other ingroup-outgroup contexts. The current findings, however, nicely exemplify a whole new meaning of the term “social robot” and the implications that go along with it.

ACKNOWLEDGMENT

The authors thank Jasmin Bernotat, Marieke Eickhoff, Natascha Neumann, Maria Neufeld, Rebecca Bröhl and Alexander Tombrink for their help in preparing and conducting the study.

REFERENCES

[1] N. Epley, A. Waytz, and J. T. Cacioppo, “On Seeing Human: A three-factor theory of anthropomorphism,” *Psychological Review*, vol. 114, pp. 864-886, 2007.

[2] C. Bartneck, E. Croft, and D. Kulic, “Measuring the anthropomorphism, animacy, likeability, perceived intelligence and safety of robots,” in *Proc. of the Metrics of Human-Robot Interaction Workshop*, Technical Report 471, pp. 37-41, 2008.

[3] B. J. Scholl, and P. D. Tremoulet, “Perceptual causality and animacy,” *Trends in Cognitive Science*, vol. 4, pp. 200-209, 2000.

[4] F. Heider and M. Simmel, “An experimental study of apparent behavior,” *American Journal of Psychology*, vol. 57, pp. 243-249, 1944.

[5] N. Epley, A. Waytz, S. Akalis, and J. T. Cacioppo, “When we need a human: Motivational determinants of anthropomorphism,” *Social Cognition*, vol. 26, pp. 143-155, 2008.

[6] R. W. White, “Motivation reconsidered: The concept of competence.” *Psychological Review*, vol. 66, pp. 297-333, 1959.

[7] M. B. Brewer, “A dual-process model of impression formation”, in R. S. Wyer Jr., and T. K. Srull (Eds.), *A Dual-Process Model of*

Impression Formation: Advances in Social Cognition (pp. 1-35). Hillsdale: Erlbaum, 1988.

[8] S. T. Fiske, “Stereotyping, prejudice, and discrimination,” in D. T. Gilbert, S. T. Fiske, and G. Lindzey (Eds.), *Handbook of social psychology* (pp. 357-411), New York: McGraw-Hill, 1998.

[9] H. Tajfel, and J. C. Turner, J., “An integrative theory of intergroup conflict,” in W. G. Austin and S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33-47). Oxford, England: Brooks/Cole, 1979.

[10] J. C. Turner, M. A. Hogg, P. J.. Oakes, P. J., S. D. Reicher, and M. S. Wetherell, “*Rediscovering the social group: a self-categorization theory”*. Cambridge, MA: Basil Blackwell.

[11] F. Eyssel and F. Hegel. “(S)he’s got the look: Gender stereotyping of social robots,” *Journal of Applied Social Psychology*, to be published.

[12] A. Powers, and S. Kiesler, “The advisor robot: Tracing people’s mental model from a robot’s physical attributes”s. *Proc. of the Conf. on Human-Robot Interaction (HRI 2006)*, Salt Lake City, UT, USA, pp. 218-225, 2006.

[13] A. Powers, A. D. I. Kramer, S. Lim, J. Kuo, S.-L. Lee, S-L., and S. Kiesler, “Eliciting information from people with a gendered humanoid robot,” in *Proc. of the 14th IEEE Int. Symp. on Robot and Human Interactive Communication*, Nashville, TN, USA, pp. 158-163, 2005.

[14] Lee, S.-L., Kiesler, S., Lau, I. Y., & Chiu, C. Y. (2005). Human mental models of humanoid robots. Proceedings of the *IEEE 2005 International Conference on Robotics and Automation*, *2767-2772, 2005.*

[15] J.-Ph. Leyens, A. P. Rodriguez, R. T. Rodriguez, R. Gaunt, P. M. Paladino, J., Vaes, and S. Demoulin, “Psychological essentialism and the attribution of uniquely human emotions to ingroups and outgroups,” *European Journal of Social Psychology*, vol. 31, pp. 395-411, 2001.

[16] N. Haslam, P. Bain, S. Loughnan, and Y. Kashima, “Attributing and denying humanness to others,” *European Review of Social Psychology*, vol. 19, pp. 55-85, 2008.

[17] N. Haslam, Y. Kashima, and S. Loughnan, “Subhuman, inhuman and superhuman: Contrasting humans with nonhumans in three cultures,” *Social Cognition*, vol. 26, pp. 248-258, 2008.

[18] J. E. Escalas, and J. R. Bettman, J. R. , “Self-construal, reference groups, and brand meaning,” *Journal of Consumer Research*, vol. 32, pp. 378-388, 2005.

[19] K. White, and D. W., Dahl, “Are all out-groups created equal? Consumer identity and dissociative influence. *Journal* of *Consumer Research*, vol. 34, pp. 525-535, 2007.

[20] F. Hegel, F. Eyssel, and B. Wrede, “The Social Robot Flobi: Key Concepts of Industrial Design,” in *Proc. of 19th IEEE Int. Symp. in Robot and Human Interactive Communicatio*n, pp. 120-125, 2010.

[21] F. Hegel, “*Gestalterisch-konstruktiver Entwurf eines sozialen Roboters*” *[Design and concept of a social robot]*, Der Andere Verlag, Tönning, Germany, 2010.

[22] A. J. C. Cuddy, S. T. Fiske, and P. Glick, “Universal dimensions of social cognition: Warmth and competence”, *Trends in Cognitive Science*, vol. *11*, pp. 77-82, 2006.

[23] H. M. Gray, K. Gray, and D. M. Wegner, “Dimensions of mind perception”, *Science*, vol. 315, p. 619, 2007.

[24] G. Echterhoff, E. T. Higgins, and S. Groll, “Audience-tuning effects on memory: The role of shared reality,“ *Journal of Personality and Social Psychology, vol.* 89*,* 257-276, 2005.

[25] C. Nass, C., and Y. Moon, “Machines and mindlessness: Social responses to computers, ” *Journal of Social Issues*, vol. 56, pp. 81-103, 2000.

PM1-1

Title: Mobility Training of Infants and Toddlers Using Novel Mobile Robots and Interfaces

Authors: Sunil K. Agrawal*, Xi Chen**, James C. Galloway***, C. Ragonesi****

*Professor, Dept of Mechanical Engineering (**Speaker**), ** PhD Student, Dept of Mechanical Eng., ***Associate Professor, Dept. of Physical Therapy, ****PhD Student, Biomechanics and Movement Science Program

University of Delaware, Newark, DE 19716.

Abstract: Mobility impaired children are limited in exploring their environment. This impacts their cognitive and social developments in important early years. This talk describes a series of studies conducted at the University of Delaware on mobility training of infants and toddlers to purposefully drive in an environment using novel prototypes of mobile robots with conventional joystick [1], force-feedback joysticks [2], mobility interfaces for crawling and walking [3]. These robots were used to teach children "purposeful driving", "motion primitives such as turning left and right", "navigation within an obstacles course", "exercising their arms and legs". Studies of mobility training for infants and toddlers have been performed with these devices in research laboratory, preschool classrooms, and homes [1-4]. The talk will summarize the design of the devices, algorithms, and results of these training studies. This research is supported by grants from *National Science Foundation* and *National Institute of Health*.

References:

[1] J. C. Galloway, J.C. Ryu, S. K. Agrawal, "Babies driving robots: Self-generated mobility in very young infants", *Intelligent Service Robotics*, Vol. 1, No. 2, 2008, pp. 123-134 .

[2] Chen, X., Ragonesi, C., Galloway, J. C., and Agrawal, S. K., "Training Toddlers Seated on Mobile Robots to Drive Indoors among Obstacles", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2011.

[3] Chen, X., Liang, S., Dolph, S., Ragonesi, C., Galloway, J. C., Agrawal, S. K., "Design of a Novel Mobility Interface for Infants on a Mobile Robot", *ASME Journal of Medical Devices*, Vol. 4, 2010.

[4] Ragonesi, C., Chen, X., Agrawal, S., Galloway, J.C., "Power Mobility and Socialization in Preschool: A Case Report on a Child with Cerebral Palsy", *Pediatric Physical Therapy*, 2010, 322-329.

http://www.udel.edu/udaily/2011/mar/joysticks-infants-robots-030911.html

http://memagazine.asme.org/Articles/2011/March/Robots_Infants.cfm

http://www.udel.edu/PR/UDaily/2008/nov/robot110907.html

PM1-2

Collecting A Developmental Dataset of Reaching Behaviors: First Steps.

Tingfan Wu*,Juan Artigas‡,Whitney Mattson†, Paul Ruvolo*, Javier Movellan*, Daniel Messinger†‡
Machine Perception Laboratory*, University of California San Diego
Psychology†, Electrical & Computer Engineering‡, University of Miami
{jatigas, wmattson}@psy.miami.edu, {ting,paul,movellan}@mplab.ucsd.edu, dmessinger@miami.edu

***Abstract*—At birth infants are faced with the difficult task of learning to control their bodies to interact with the physical and social world around them. From a motor control point of view the problem infants face is monumental, due to the high number of degrees of freedom, high compliance, and low-repeatability provided by the human musculoskeletal system. In fact the complexity of the control tasks that infants solve so effortlessly far outstrips the abilities of the most sophisticated approaches to motor control and artificial intelligence. Observing how infants develop such motor skills as they interact with objects and caregivers may provide insights for new approaches to robotics. One obstacle for progress in this area is the lack of datasets that simultaneously capture the motion of multiple limbs of infants and caregivers across the developmental process. Here we present our first steps towards the collection of one such datasets, focused on the development of reaching behaviors. We describe the technical and logistic problems we faced so far and the solutions we found. We also show preliminary analyses that illustrate how the collected data suggests new approaches to motor control in robotics, and to theories of motor development in infants.**

I. Introduction

The sensory-motor systems of the brain generate movements that are compliant and non-repeatable, yet remarkably well-adapted to an unstructured, uncertain and non-stationary world. Contrary to this, the fields of robotics and motor control has for the most part focused on simplifying the control problem by using stiff, highly geared actuators, emphasizing repeatability over compliance, and avoiding unstructured conditions. This approach worked well for industrial applications and revolutionized the assembly line. However, in order to develop robotic technology that could transform daily life, it is important to focus on robots that approximate the control properties of the human body. The problem is that due to the complexity, compliance, non-repeatability, and temporal dynamics of the human body, most of the control schemes used in current practice become inapplicable.

Infants face the very difficult task of learning to control their bodies to interact efficiently with the physical and social world around them. From a motor control point of view this problem is formidable, due to the high number of degrees of freedom, high compliance, and low-repeatability provided by the human musculoskeletal system. The solution to this problem eludes the most sophisticated approaches to robotics and artificial intelligence. Yet infants solve this problem seamlessly within a few years of life. Understanding how this is done and reproducing this process in robots may have profound scientific and technological consequences. One obstacle for progress in this line of work is the lack of datasets that jointly capture the joint development of infant and caregiver body motions at high temporal and spatial resolution. Such datasets may provide the key to reverse engineer human motor development and to synthesize it in new human-mimetic control algorithms for robots. Another obstacle for progress is the lack of robot systems that approximate the complexity, compliance and control dynamics of the human body. Based on these motivating ideas, we have been pursuing a project whose goal is to gain a computational understanding of how infants learn to control their bodies. One component of this project focuses on the development of a sophisticated humanoid robot named Diego San (see Fig.1), that approximates the complexity, compliance and control dynamics of the human body. The other component, which is the focus of this paper, focuses on the collection of a motion capture dataset to understand the development of reaching behaviors in the context of physical and social interactions with the world. While the use of motion capture has recently become popular in the motor development literature, a unique aspect of our work is the attempt to simultaneously capture the motion of the entire body (arms, legs, trunk, head) in both the

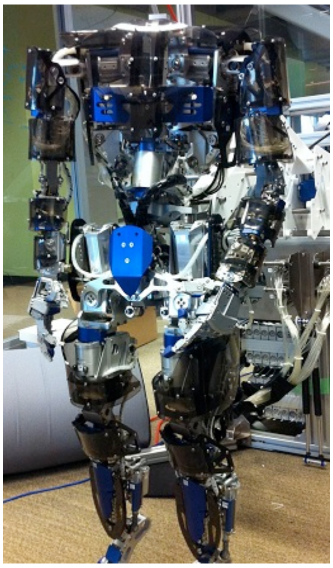


Fig. 1. Diego: a humanoid robot that approximates human body complexity and dynamics.

infant and caregiver. In this paper, we present our initial steps towards capturing such data. We describe the technological and logistic problems we are facing and the solutions we are finding to these problems. We also describe our preliminary steps analyzing the data obtained so far.

II. Setup and Data Collection Procedure

Although motion capture technology is now quite mature, most modern systems are optimized for single adults in standing postures with low levels of occlusions. However, our experiment focuses on infants in a natural lying posture interacting with their caregivers. Based on pilot work with several motion capture systems, we decided to use the PhaseSpace Impulse, because of the fact that it uses active LED markers. In contrast to systems using passive markers (such as the Vicon system), active LED markers have individual digital signatures so that the system can easily tell the identity of each marker at each point in time. This is particularly important in setups, like ours, with a large number of occlusions. Another alternative would have been to use magnetic based motion capture. Unfortunately due to the magnetic characteristics of our experimental room this was not an option we could use. A disadvantage of active markers is that they need to be powered via cables attached to a small wireless driver (see Figure 2). Our final system utilized 10 infrared cameras. The cameras were installed around the perimeter of a 3.3m×3.3m sound-attenuated playroom for recording. Additionally, four fixed video cameras are used to supplement the infrared system. This includes a small headband camera worn by the mother to record the infant's expressions during the sessions.

A. Mocap Suit Design

As there are no commercially available motion capture suits for infants we had to develop our own. This apparently mundane task turned out to be surprisingly difficult. While we have made significant progress, every capture session teaches us new lessons on how to improve the design. Originally we had 4 design criteria:

- **Scalability** the system needed to allow for the addition and removal of LED markers depending on task demands.
- **Low profile** the wiring and placement of the suit markers should not interfere with infant movement, or distract the infant during sessions. Additionally, the markers could not touch the infant's skin.
- **Durability** the longitudinal design involved multiple observation sessions a week with each infant being seen over the course of 8-10 weeks. The suit needed to be robust enough to functionally endure these sessions.
- **Redundancy** capturing the motion of an infant lying in the prone position with his or her mother leaning over the infant presented significant line of sight problems that are not typically encountered in traditional motion capture on adults. Thus, the suit had to be designed with a high degree of marker redundancy to optimize the amount of time each part of the infant's body was able to be tracked by the PhaseSpace system.

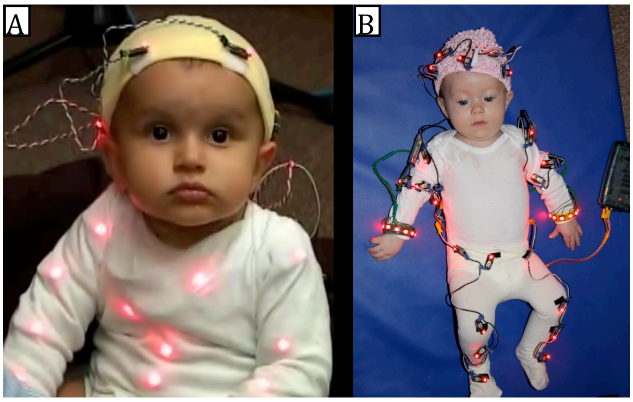


Fig. 2. Versions of the infant suit. A) An early suit prototype, note the diffusion of light from the LEDs on the infant's arms and body created by the external onesie. B) The current suit design, note the absence of the external onesie, and the additional markers

Each system of markers connects to an individual wireless LED driver, which supplies power for the markers and sends data from the markers to the server. Initially, the 24-gauge ribbon cable that connects the LED driver was spliced and an array of parallel LEDs was connected to each port in the LED driver. The markers were then attached using Velcro adhesive strips connected to an infant onesie. The procedure was to dress the infant in the onesie, attach the markers, and then place another onesie over the markers. (See Fig. 2A). After several pilot sessions with using this configuration, we found that the additional onesie caused diffusion in the LEDs, causing inconsistent tracking of markers.

A modified suit was designed (see Fig.2B), this time using series connections from the LED driver with connectors that could be crimped onto the wire directly. The series connections both reduced the amount of wire and made for a more logical arrangement of the LED strings. Each LED string could accommodate up to 7 markers per string. Five LED strings were constructed, two for the arms, two legs, and a head/body string. Given earlier tracking problems caused by the outermost onesie (see Fig.2A), we piloted a session with only the inner onesie and the now external wiring. Surprisingly, the pilot infant did not appear to notice the external wiring or the LED markers. We are currently using this configuration for all data acquisition. To date, no infants have spent a noticeable amount of time gazing at or touching the markers or other components of the suit.

B. Experimental Procedure

On each motion capture session infant and caregiver are brought into the motion capture playroom and the infant is placed on his/her back with caregiver facing the child (see Fig. 3). Caregiver is instructed to interact with infant in the following conditions:

1) Face to face without any toys for 2 minutes.
2) Caregiver offers a series of toys for 10 minutes with the goal of initiating reaching.

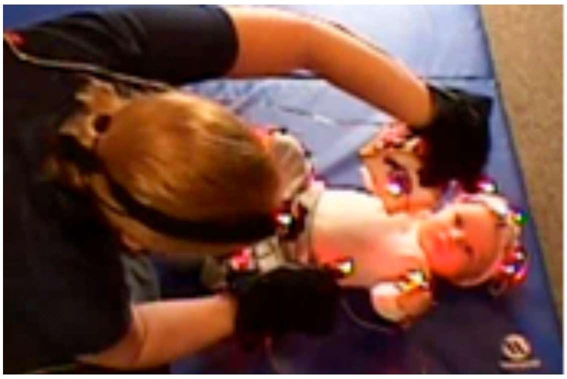


Fig. 3. An example of our experiment scenario. The infant is looking at the toy shown by the mother.

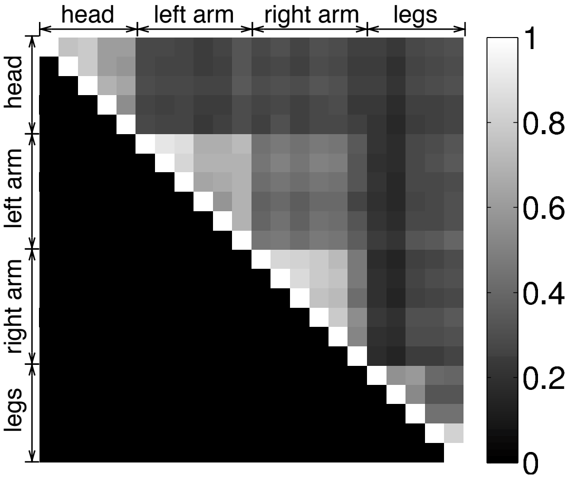


Fig. 4. correlation of movement magnitude between body regions by marker

3) Unstructured infant play with mobile for 1 minute.

The session ends with the examiner presenting toys to the infant in predetermined series of positions.

III. Preliminary Results

A. *The Nature of the Developmental Process*

The classic robot control literature typically measures the difficulty of a control problem in terms of the number of degrees of freedom and joint compliance. A robot with a large number of compliant joints, results on a highly dimensional problem with tightly coupled non-linear dynamics. Such problems are very difficult to solve with current approaches. A classical solution is to reduce the effective number of degrees of freedom by temporarily or permanently stiffening some joints, so that the control problem decouples into a small set of independent equations. Perhaps influenced by classic robotics theory, the literature in developmental psychology has adopted a similar point of view. For example, it has been proposed that infants' early attempts to reach engage a low number of degrees of freedom. From this point of view development proceeds from engagement of less to engagement of more degrees of freedom.

Unfortunately much of the behavioral evidence supporting this less-to-more view of motor development has focused on laboratory experiments that capture the motion of a single infant arm as it reaches towards an object in very restricted conditions [1], [2].

Thus our dataset will offer a unique opportunity to test current theories of motor control by analyzing how the different parts of the body, not just an individual arm, coordinate throughout development. To begin exploring this issue we performed some preliminary analysis on the data obtained so far, focusing on how hands and legs move while mom offers baby a toy for him/her to reach. The analysis was performed on a single motion capture session of an 18 week old infant. First, the instantaneous motion energy of each marker was estimated by squaring the displacement between adjacent frames. A displacement was counted only when a marker is visible in both frames and the displacement is reasonable (speed $<$ 24m/s) so as to exclude missing markers and motion marker jitter. The displacement values were normalized per marker across the entire session. Second, the temporal correlations between the average group motion energy for marker located on two segments of the infant's body were computed over the entire session. If two limbs were moving at the same time regardless of the direction of movement, there will be a high correlation between the corresponding markers (Fig.4).

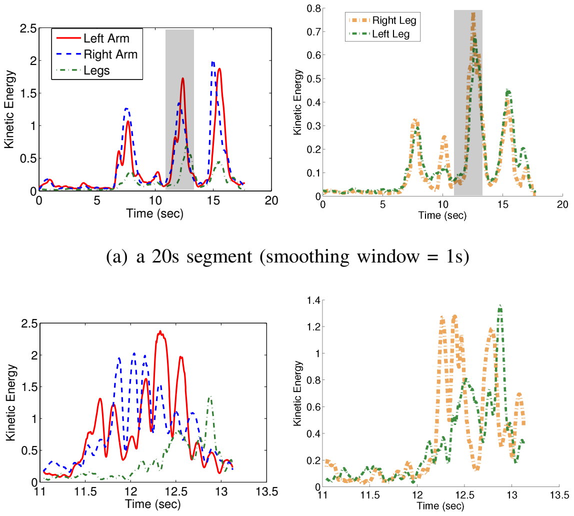
Not surprisingly, high correlations were observed for markers located within each of the following groups: head, left arm, right arm, left leg and right leg. Interestingly, high correlations were also observed between markers on different limbs (e.g., 0.4 Pearson correlation coefficient between the left and right arm). Basically it appears that infants were simultaneously using all the limbs (the two arms and legs). To explore the temporal unfolding of inter-limb correlation, we plotted the motion energy over time as the infant attempted to reach for an object (see Figure 5(a)).

Note that the two arms and legs move in synchrony bursts that lasts approximately 4 seconds. Looking closer at the second burst (see Fig.5(b)), we see a fine grain motion energy plot. Figure 6 shows the actual movement of all tracked groups in the horizontal plane.

Thus the evidence, while preliminary, does not appear to be consistent with a less-to-more view of motor development. If anything early in development children appear to use many more joints than necessary. Rather than reaching with a single arm, they appear to be simultaneously reaching with both their arms and limbs.

B. *Analyzing Mother Infant Contingency Structures*

An important component of the dataset we are collecting is the fact that we simultaneously motion capture infant and caregiver data. Our use of motion capture technology allows the analysis of contingencies between behavior of mother and infant at a very high spatial and temporal resolution with minimal human coding. The issue of which actions of a particular



(a) a 20s segment (smoothing window = 1s)

(b) zoom-in the second episode (shaded) in (a) (smoothing window = .05s)

Fig. 5. Kinetic energy for left arm, right arm, and legs during a 20 second segment

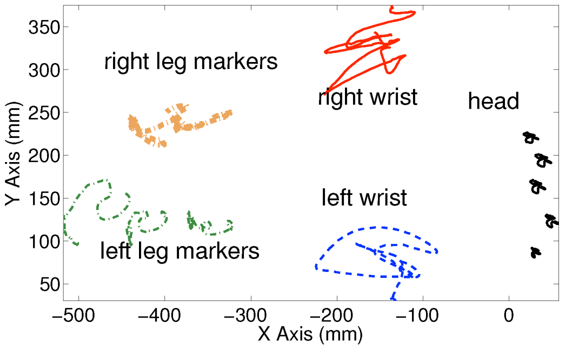


Fig. 6. Marker trajectories (projected on to the X-Y plane) of the movement episode shown in Fig. 5

partner within the context of a dyadic interaction engender predictable (contingent) responses of the other partner has long been a focus of developmental science [3], [4]. Most work in this area follows two basic steps: first use coarse-grained hand-coded variables to describe the dyadic interaction (e.g. coding infant gaze as being to one of several candidate loci of attention); second, use techniques such as cross-correlation to identify the temporal properties of influence between mother and infant [3]. Here we address this issue by examining to what extent the 3D position of a marker on the infant can be predicted by the 3D position of a marker on mother and vice-versa. Specifically, we are interested in the following questions:

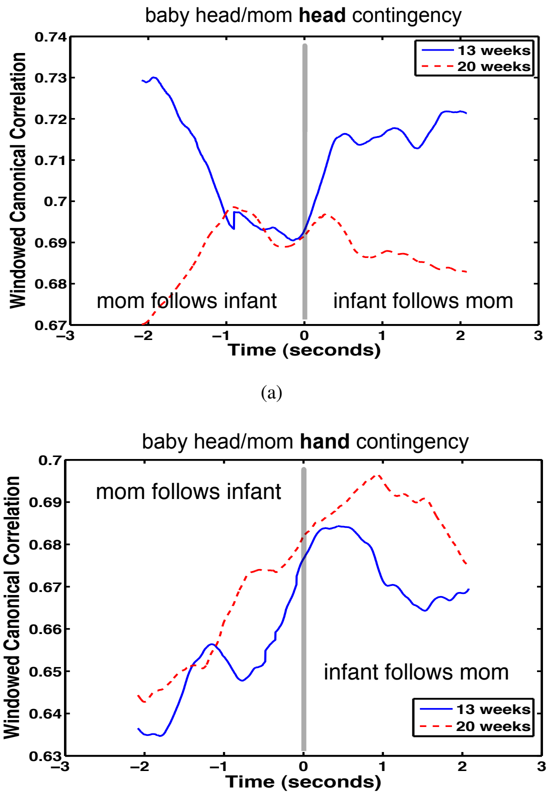
1) Movement of which parts of mother's body are most predictive of the orientation of the infant's head?
2) What is the temporal delay for each possible location on the mother's body that is maximally predictive of the orientation of the infant's head?

In this section we present our initial steps in developing computational strategies to address these questions. We used windowed Canonical Correlation Analysis [5] of the 3D position of one marker on the mother and the 3D position of the marker at the center of the infant's forehead.

Within the context of time series analysis, Canonical Correlation Analysis is a technique for projecting each of two multivariate time series to create two univariate time series that are maximally correlated. In the context of motion capture analysis we can view this as computing two directions of motion: one for marker 1 and one for marker 2 such that the motion of each of these markers projected onto the computed directions is maximally correlated. For example, Canonical Correlation might project the movement of mother's hand in the direction transverse to the infant's body and the infant's head motion direction along the same direction, indicating that motion of mother's hand across the infant's body is predictive of shaking of the infant's head. For each experiment one time series was composed of 3D positions of the marker on the center of the infant's forehead and the second time series was the 3D position of a particular marker on the mother. We used two marker locations on mother to investigate which was most informative of the infant's head orientation. Specifically, one marker was on mother's right hand and the second marker we investigated was at the center of mother's forehead.

We computed the canonical correlation as a function of time by using a running temporal window. We utilized two different window lenghts: one short window (2.08 seconds) that gives us fine temporal resolution but a somewhat noisy estimate of the local canonical correlation, and a longer window (16.67 seconds) that gives us a more stable characterization of the canonical correlation that is useful for characterizing coordination over entire sessions. For each session analyzed and for each time window length, the canonical correlation was computed using a sliding window with 50% temporal overlap between adjacent windows (e.g. for the 16.67 second window the first temporal window would be 0s to 16.67s and the second would be 8.34s to 25s, followed by 16.67s to 33.33s).

We examined one particular mother-infant dyad (Subject 10) at two time periods, 7 weeks apart. The Infant was 13 weeks old at the first session and 20 weeks old at the second. For each session we computed windowed canonical correlations between the two markers on mother and the marker at the center of the infant's forehead. The plots in Fig.7 show the mean canonical correlation values for each of the sessions as a function of delay. Peaks in the negative region indicate periods in which the movement of the mother's marker was predicted by a preceding movement of the infant's head, i.e., mom was following the infant. Peaks in the positive region indicate periods in which the infant is following mom. Figure 7 (a) suggests an approximately equal number of episodes in which mom's head motion follows infant's head motion and episodes in which infant head motion follows mom head motion. The time delay in these contingencies decreases with development. Figure 7 (b) shows that for the most part infant's



(a)

(b)

Fig. 7. Windowed Canonical Correlation for two sessions between infant head movement and mother head movement (top) and infant head movement and mother right hand movement (bottom).

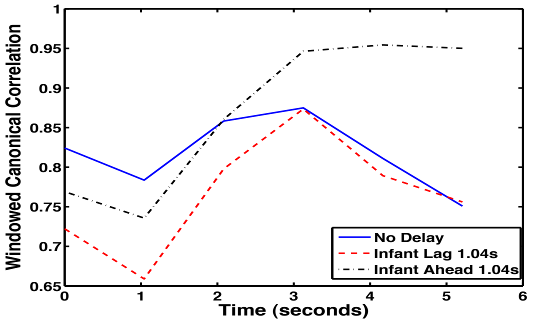


Fig. 8. Windowed Canonical Correlation for a segment where the infant shakes head and subsequently moves a toy that mother is holding (which causes her hand to move). This event happens at time = 4 seconds on the graph. The event causes the canonical correlation for Infant Ahead 1.04s to spike since the head motion precedes the motion of the marker on mother's right hand.

head follows mom's hand movement, rather than mom's hand movement following infant's head. Again the temporal delay of this contingency decreases with development.

In addition to analyzing the mean canonical correlation over the entire session, we also investigated whether short-time Canonical Correlation Analysis (window length of 2.08 seconds) could be used to to spotlight meaningful episodes of mother-infant interaction. The 3 regions with highest canonical correlation were as follows:

1) Subject 10 at 13 weeks: Mother struck a butterfly mobile that she was holding above the infant with her right hand and approximately 1 second later the infant shook its head. This segment had a high canonical correlation with a positive delay of 1.08 seconds whereas the canonical correlation for 0 delay and a delay of -1.08 seconds was lower.
2) Subject 10 at 13 weeks: While the infant grasped a toy that mother holding simultaneously, the infant shook his head and then moved the object with his hand. The movement of the infant's hand caused mother's hand to move as well. This episode is an example of a contingency between mother motion and infant motion that was entirely created by infant motor movements. The Windowed Canonical Correlations are given in Fig.8.
3) Subject 10 at 20 weeks: Mother tickled infant with her right hand, 1 second later the infant shook his head

Interestingly, most of the highest coupling events were not instances of the infant tracking the movement of mother's hand, but rather were composed of dynamic interactive events that incorporated both play with objects as well as social interaction.

While the present analysis is promising there are several extensions that we will pursue in the future:

- **Multiple Markers** Canonical correlation between multiple points on infant and multiple points on mother may give interesting contingency patterns that go beyond the single body part to single body part analyses presented here.
- **Canonical Correlation with Infant Kinematics** Instead of using the location of the center marker on the infant's head, it may be more informative to look at quantities derived from the markers such as infant head orientation or joint angles of various limbs of the infant.

IV. Discussion and Conclusion

The human musculoskeletal system is a sophisticated machine designed to support movements that are compliant, versatile, and adapted to the uncertain nature of the world. The price paid for this compliance and versatility is the fact that simple control approaches like the ones used in contemporary robotics, do not work. From an engineering point of view, infants face a formidable motor control problem when learning to control their own bodies. While controlling robots with the complexity and dynamics of the human body is currently beyond our most sophisticated algorithms and powerful computers, infants learn to do so seamlessly in less than two years. Uncovering how this is done will have profound scientific and technological consequences. A critical limitation of the literature on motor development is the fact that it is carved into isolated research niches e.g., reaching, facial expressions, crawling, walking, shared attention. Each of these niches typically studies one part of the body as it performs a single task, e.g., hands and arms reaching for an object. While this research strategy is reasonable, it may result in a distorted perspective of how infants really learn to control their bodies. Here we presented our first steps to address the current limitations in the developmental literature.

The preliminary data obtained so far is already presenting a different perspective about the development of motor control. For example, current work on the development of reaching emphasizes the fact that infants reach with very little movement in the elbow's joint when compared to adults. This is interpreted as evidence of a "less-to-more" developmental trajectory [1], [2]. This point of view seems plausible because we tend to assume that moving less degrees of freedom is easier than moving more degrees of freedom. Our experience with the compliant humanoid robot Diego San, being developed as part of this project is that this is not necessarily the case, i.e., due to the high compliance of the joints, getting Diego San to move one degree of freedom at a time is quite difficult. Interestingly our motion capture data suggests that during early reaching episodes infants not only move their arms and hands but also engage their legs, head and face. If anything the data suggests a developmental trajectory that progresses from engaging more degrees of freedom (reaching with two arms and two legs), to engaging less degrees of freedom (reaching with one arm). We are also finding that the physical and social contexts of motor development are tightly coupled. The result is that behavioral categories that are natural from an adult perspective may be artificial from an infant's perspective. For example, when a caregiver is present, making facial expressions, vocalizing, or moving the legs, may be as effective to make contact with an interesting object, as reaching with the arms and hands. Thus, it could be argued that moving the legs or making facial expressions should be a legitimate part of the literature on the development of reaching. This developmental approach is quite different to the standard way we get robots to reach: simplify the control problem by using as few degrees of freedom as possible. Thus it appears that our current notion of "simple" in robotics may not match well the notion of "simple" that infants appear to work with.

The results presented here are still preliminary and thus they should be taken only as an illustration of things to come. Yet the approach and the technologies we are exploring are already contributing a different perspective on motor development. This perspective may give us clues to develop a new generation of robots that learn to control their own bodies.

Acknowledgment

Thanks to all of the families who participated. This research is supported by the National Science Foundation under Grant No. IIS-INT2-0808767 and 0808653.

References

[1] N. Berthier, R. Clifton, D. McCall, and D. Robin. Proximodistal structure of early reaching in human infants. *Experimental Brain Research*, 127(3):259–269, 1999.

[2] A. Bhat, H. Lee, and J. Galloway. Toy-oriented changes in early arm movements ii–joint kinematics. *Infant Behavior and Development*, 30(2):307–324, 2007.

[3] A. Fogel, D. Messinger, K. Dickson, and H. Hsu. Posture and gaze in early mother–infant communication: synchronization of developmental trajectories. *Developmental Science*, 2(3):325–332, 1999.

[4] K. Kaye and A. Wells. Mothers' jiggling and the burst–pause pattern in neonatal feeding*. *Infant Behavior and Development*, 3:29–46, 1980.

[5] B. Thompson. Canonical correlation analysis. 2000.

Elevated activation of dopaminergic brain areas facilitates behavioral state transition

Beata J. Grzyb*†, Joschka Boedecker†, Minoru Asada†, and Angel P. del Pobil*
*Robotic Intelligence Laboratory, Jaume I University
Castellon de la Plana, 12071, Spain
Email: {grzyb,pobil}@icc.uji.es
† JST ERATO Asada Project, Graduate School of Eng., Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
Email: {joschka.boedecker,asada}@ams.eng.osaka-u.ac.jp

Abstract—**Dopamine neurons appear to code the discrepancy between a reward and its prediction, and as such play a key role in learning from positive and negative feedback. Although the traditional view stresses the role of errors in learning, we suggest that a temporary decrease in learning from negative feedback may in fact facilitate the process of finding more suitable behaviors that would reflect the change in behavioral competences of the agent. More in detail, omission of the errors enables selection of different behaviors in a context when they normally would not be selected providing more learning opportunities for fine-tuning these behaviors. Herein, we propose that omission of the errors is tightly related to an elevated level of dopamine that is caused by a high reward for gaining the control over the enviromnent. Our results with a robot simulator serve as a proof-of-concept for our approach.**

I. Introduction

Obtaining a desired objective is a satisfying experience for nearly everyone. The way of achieving the goal is not always easy and straightforward. Sometimes many different methods need to be explored until the goal can be accomplished. If none of the possible solutions is helpful in attaining the goal, the goal is memorized as impossible and later attempts abandoned. Sometimes, however, these efforts may be resumed when new competences appear. This can clearly be seen in infants. At the beginning nothing is reachable for them, as they do not have enough skills to coordinate eye and hand movements. Soon these first movements become successful, and infants learn that only close but not far objects are reachable for them. This, however, changes with the onset of crawling and later on with walking behavior. Consequently, infants need to relearn that attaining distant objects is possible only by locomotion, a behavior more appropriate in this context.

Our previous experiments revealed that infants during the transition phase to walking show a decreased ability to learn what lies within their reachable space [1], [2]. We suggested that the blocked ability to learn from negative outcome while reaching makes infants fine-tune their walking skill, as a primary motive for walking is to reach for something [3]. Thus, we proposed that a temporary decrease in learning from negative feedback could be an efficient mechanism behind infant learning new skills. Furthermore, we proposed that disregarding the errors is tightly connected to the sense of control, and results from an extremely high level of perceived self-efficacy. In this paper, we propose possible brain mechanisms that may lead to omission of errors during feedback processing.

An important theoretical framework underlying our proposal is the dynamic systems approach [4]. In this framework stable configurations of a system are referred to as attractors. In that sense, the decision not to reach for far objects is a stable attractor for the infants that are not able to locomote on their own. When a possibility of a new behavioral competence appears, for example walking, the system should change its decisions about appropriate behaviors. Therefore, the existing stable attractor needs to be destabilized in order for the system to accomodate a new one.

Another important role in our proposal is played by the neuromodulator dopamine. Phasic dopamine signals were suggested to trigger a switch in the current attractor state in the networks of prefrontal cortex, by transiently enhancing afferent input while potentiating local inhibitory signals thus gating new information into the prefrontal cortex [5]. Furthermore, dopamine in the amygdala was suggested to modulate behavioral transitions [6] that characterize development and progression in competences. The main focus in the literature, however, has been put on dopamine and its role in the prediction of rewards and determination of whether predictions about outcomes are violated or verified. A decrease of dopaminergic activity owing to the omission of rewards, was mainly interpreted as coding a prediction error or learning signal that is supposed to trigger learning and adaptation of future behavior [7]. Another interesting point is that the dopamine system may act at several different timescales in the brain from the fast, restricted signalling of reward and some attention-inducing stimuli to the slower processing of a range of positive and negative motivational events [8].

The main premise of our approach is that when an infant gains control over its environment, the reward circuitry in her brain will deliver a large reward to the executive brain areas facilitating repeated selection of actions that led to the gain of control. This reward is experienced even though behavior on a shorter time scale (e.g. reaching) fails, but progress is made on behavior spanning an extended time scale

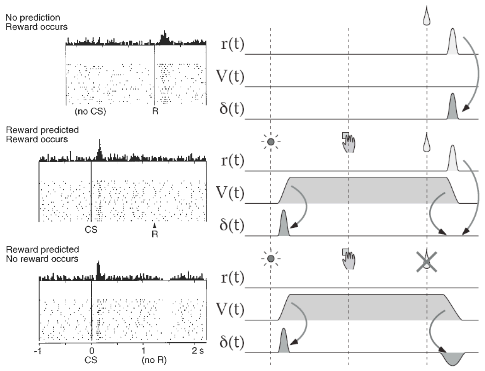


Fig. 1. On the left: Reward prediction error response of single dopamine neuron (taken from [9]). On the right: Interpretation of the responses of midbrain dopamine neurons in the TD model; r(t): reward; V(t): reward prediction; δ: dopamine response (taken from [10])

(e.g. reaching by walking). Omitting the errors in this way enables selection of different behaviors in a context when they normally would not be selected, thus destabilizing existing attractors and facilitating the formation of new ones.

We review the mechanism of the reward prediction error both in neuroscience, and in reinforcement learning theory in the next section. After that, we present cases where learning from negative outcome was significantly decreased. In Sec. IV we introduce our hypothesized neural circuitry facilitating behavioral state transition. The next section, presents results from our robot simulation study where the underlying assumptions of the model were tested. We close the paper with discussion and an outline of future work.

II. Prediction error learning

The response of dopamine neurons appears to code the discrepancy between a reward and its prediction [9]. A typical response of a single dopamine neuron is shown in Fig. 1 (on the left). During the acquisition process the dopamine neurons increase firing rates when reward (R) is received but not expected (no CS). Over time this increase in firing rate is back propagated to the earliest reliable stimulus (CS) for the reward. The dopamine cells no longer increase their firing rate upon presentation of the predicted reward. However, when rewards are expected but not received, the firing of dopamine neurons drops below tonic baseline levels.

The activity pattern of dopamine neurons represents the reward prediction error, which is central to the temporal difference (TD) learning model [11], [12]. The TD model calculates a prediction error $\delta(t)$ based on the temporal difference between the current discounted value function $\gamma V(t)$ and that of the previous time step, $V(t-1)$.

$$\delta_t = r(t) + \gamma V(t) - V(t-1), \tag{1}$$

where γ is a discount factor which allows rewards that arrive sooner to have a greater influence over delayed ones, and $r(t)$ represents the current reward [12], [10]. The interpretation of the dopamine neurons responses in the TD model is shown in Fig. 1 (on the right). Before learning, no reward is predicted, that is $V(t) \equiv 0$. Thus, the TD error $\delta(t)$ is the same as the reward itself. After learning has been completed, the predicted future reward $V(t)$ builds up immediately after the cue signal, causing the discounted temporal derivative to provide a positive pulse in the TD error even if there is no reward. At the time of reward delivery, $V(t)$ drops to zero and the negative temporal derivative of $V(t)$ cancels out the positive reward signal. However, when the reward is omitted, there is a negative response due to the drop in the predicted reward $V(t)$. By acting as a teaching signal, dopamine-mediated prediction errors are expected to gradually train learning mechanisms to improve their predictions in an incremental and trial-by-trial fashion [13].

III. Decreased ability of learning from errors

Although there may be some individual differences due to genetic variations affecting dopamine function, in general healthy people are equally good at learning to obtain positive outcomes and to avoid negative outcomes. People with Parkinson's disease, however, show specific deficits in trial-and-error learning from feedback. These effects were nicely explained by Frank's basal ganglia model [14]. Basal ganglia dopamine levels in these patients are severly depleted as a result of cell death. As the positive outcomes are signaled by a raise in the firing rate of dopamine neurons, the depleted overall dopamine levels in unmedicated patients results in a weaker reinforcement of the stimulus. On the other hand, the errors in reward prediction are signaled by a decrease in the firing rate of dopamine neurons. As a result of low dopamine levels, the errors in unmedicated patients have much stronger negative reinforcement of the stimulus. The dopaminergic medications, however, reverse these biases and medicated individuals with Parkinson's disease are better at learning from positive than from negative feedback. The dips of dopamine required to learn negative prediction errors are effectively filled in by the medication, and such blunting of negative prediction errors reduces learning from negative outcomes. Essentially, the medication prevents the brain from naturally and dynamically regulating its own dopamine levels, which has a detrimental effect on learning, particularly when dopamine levels should be low, as for negative decision outcomes.

The inability to learn from negative feedback was shown in healthy subjects during the trust game [15]. In this experiment information about the moral profile of the opponent was provided to the players before the game started. This information can create a prior belief, but feedback from the game should adjust this prior belief to reflect new evidence. However, the experiment showed the lack of differential responses between the positive and negative outcomes when playing with morally good or bad partners. More specifically the activation of the caudate nucleus differentiated between positive and negative feedback, but only for the 'neutral partner', and not for the 'good' one, and only weakly for the 'bad' one. The normal trial-and-error learning would predict a sharp decrease in the feedback response following violations of expectations. One

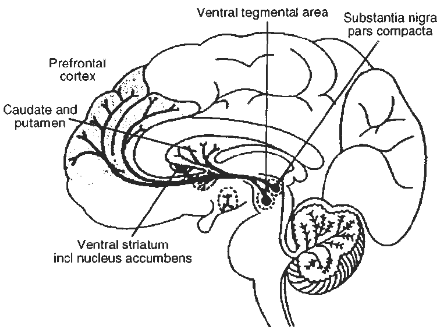

Ventral tegmental area
Substantia nigra pars compacta
Prefrontal cortex
Caudate and putamen
Ventral striatum incl nucleus accumbens

Fig. 2. The neurocircuitry of the ascending dopaminergic system (taken from [21]).

of the possible explanations suggested by the authors was that participants had a reward reaction to the presentation of the morally good partner, irrespective of decision.

In patients with bipolar disorder, failures in motor learning may result from the lack of striatal error signal during unsuccessful motor inhibition. Such deficits in motor regulation could be related to the emotional disregulation, as irritability and decreased motor inhibition may be linked mechanistically [16]. The impulsivity was suggested to represent a core characteristic of the disorder and to be responsible for symptoms like hyperactivation, excitability, and hasty decision making [17]. Patients with bipolar mania tend toward high goal setting, have unrealistically high success expectancies [18], and exhibit increased goal-directed activity and excessive involvement in pleasurable activities that have a high potential of risk [19]. Bipolar patients show elevated activation of dopaminergic brain areas when expecting high rewards compared to anticipation of no rewards, which could result from dysfunctional nucleus accumbens activation during prediction error processing [20]. When both, schizophrenia patients and healthy controls, showed lower nucleus accumbens activation upon omission rather than upon receipt of rewards as a potential correlate of such a learning signal, bipolar manic patients did not display a similar reduction in the activation of dopaminergic brain regions.

We have presented different cases where learning from negative outcome was significantly decreased. The first lesson from these examples is that elevated state of dopaminergic areas can lead to omission of the errors during learning like in the case of Parkinson's patients. The second lesson is that, abnormal activity in the striatum (dorsal or ventral) also causes decreased ability to learn from negative feedback. We believe that temporal omission of errors while learning a new skill may result from a similar mechanism. The next section introduces the details of our hypothesis.

IV. Sense of Control and Omitting the Errors

The principal assumption behind our approach is that a need for control is innate, and exercising control is extremely rewarding and beneficial for an individual's wellbeing [22], and people's ability to gain and maintain a sense of control is essential for their evolutionary survival [23]. The hypothesized neural circuitry that would explain the facilitation of behavioral state transition is depicted in Fig. 3.

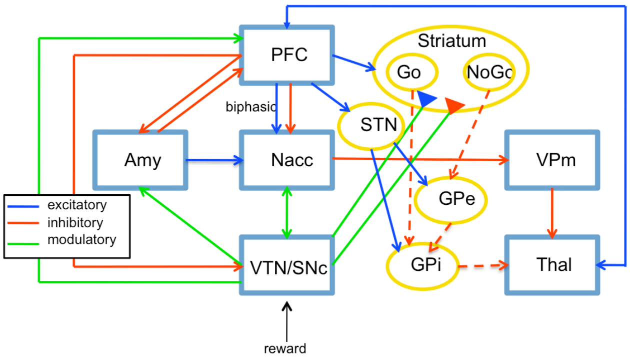


Fig. 3. The striato-cortical loops, including the direct ("Go") and indirect ("NoGo") pathways of the basal ganglia, and neural circuitry for perceiving control. PFC: prefrontal cortex; Amy: amygdala; Nacc: nuccleus accumbens; VTA: ventral tegmental area; SNc: substantia nigra pars compacta; GPe: internal segment of globus pallidus; GPe: external segment of globus pallidus; STN: subthalamic nucleus; Thal: thalamus; VPm: ventral pallidum.

Similary to existing BG models (eg. [24]), there are two BG pathways to selectively facilitate the execution of the most appropriate motor commands ("Go" pathway), while suppressing competing commands ("NoGo" pathway). The "Go" pathway depends on D1 receptors and supports learning from positive feedback, whereas the "NoGo" pathway depends on dopamine D2 receptors and supports learning from negative feedback. These two pathways compete with each other when the brain selects among possible actions, so that an adaptive action can be facilitated while at the same time competing actions are suppressed. More specifically, striatal "Go" neurons directly project to and inhibit the *internal* segment of the globus pallidus (GPi). The GPi in turn disinhibits the thalamus eventually facilitating the execution of the motor commands. Contrary, striatal "NoGo" neurons project to and inhibit the *external* segment of globus pallidus (GPe), releasing the inhibition of GPe onto GPi, and thus blocking the motor activity. Dopamine modulates the relative balance of these pathways by exciting synaptically-driven activity in Go cells via D1 receptors, while inhibiting NoGo activity via D2 receptors.

Prefrontal cortex (PFC) is constantly involved in the acquisition of new skills and knowledge, and may also play a role in organizing other parts of the cortex [25]. Increased activity in the medial PFC has been associated with perception of control [22]. The PFC and the amygdala have synergistic roles in regulating purposive behavior [26]. While the PFC guides a goal-directed behavior, the amygdala appears to extract the affective significance of stimuli. Communication between these two brain regions is bidirectional and appears to be essential in judging rewarding or aversive outcomes of actions. The PFC was shown to inversely correlate with amygdala during successful emotion regulation [27]. The inverse relationship reflects the inhibitory pathway from the dorsal and lateral regions of PFC to the amygdala. Furthermore, it was proposed that amygdala drives vmPFC in a bottom-up affective reactivity task but can be downregulated by more dorsal and lateral portions of the PFC via the vmPFC in a top-down reappraisal task. The optimal balance between such bottom-up and top-down influences in a given emotional situation was suggested to be crucial for the individual to respond adaptively [28].

The nucleus accumbens (Nacc) is a hub for information related to reward, motivation, and decision making [29]. The Nacc provides a ventral pathway by which the limbic system and prefrontal areas can influence the initiation of goal-directed behavior [30]. Dopamine D1 and D2 agonist when injected in the Nacc compared to the dorsal striatum facilitate the initiation, speed and vigor of locomotion, and markedly increase the frequency and duration of spontaneous exploratory activity. Suppression of ventral striatal activity when anticipated rewards were not obtained has been interpreted as a prediction error signal [20]. The Nacc receives strong, direct projection from the amygdala and prefrontal cortex. The PFC modulation of Nacc dopamine function appears to be biphasic [31]. Under normal activity PFC provides an inhibitory control over Nacc dopamine release. Electrical stimulation of PFC at $10Hz$, which closely corresponds to the firing rate of PFC neurons in animals engaged in cognitive tasks decreases dopamine release in the NAcc. However, electrical stimulation at $60Hz$ that is much higher then normal activity, caused an increase in NAcc dopamine levels. Activated Nacc neurons project to and inhibit pallidal neurons in the region called ventral pallidum (VPm). The suppression of tonic activity in the pallidum then disinhibits the thalamic nucleus [32].

The ventral tegmental area (VTA) dopamine cells play a crucial role in facilitating motivated behavior via its coordinated modulation of prefrontal and Nacc circuity, as well as its direct input to limbic structures which effects input to the Nacc at source [33]. Moreover, with simultaneous stimulation of both the amygdala and VTA, Nacc stimulation more readily produces initiation of forward locomotion and exploratory activity to novelty [30]. Dopaminergic input from the VTA modulates the activity of neurons within the nucleus accumbens, as well as within the PFC [34].

One possible explanation for decreased learning from negative feedback is that exercising control is highly rewarding itself and even if the outcome of the action is not as predicted, still the reward for gaining control is provided. That leads to high activity in the PFC. As discussed previously, the PFC modulates the Nacc dopamine function. This regulation is biphasic, and at normal activity the PFC provides an inhibitory control over Nacc dopamine release, but the PFC stimulation at much higher than normal levels increases nucleus accumbens dopamine. Herein, we assume that gaining control evokes such a high PFC response. Thus, high activity in Nacc leads to disinhibition of the VPm, and in turn dishinhibition of the thalamus. Simply speaking, that facilitates selection of the behaviors that led to the gain in control. This loop bypasses the dorsal striatal areas involved in action selection (colored yellow in Fig. 3). However, the dopamine prediction error that helps to improve the selected behavior still reaches these areas. Our hypothesized role of ignoring the errors is important only in the more executive areas responsible for action selections.

Fig. 4. The M3-neony robot simulator.

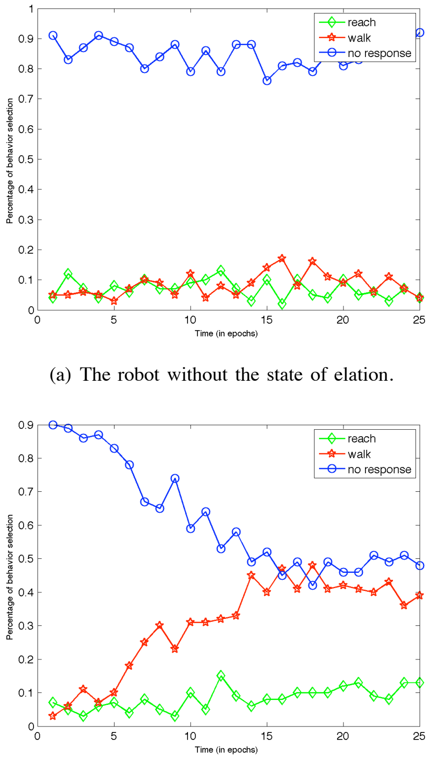
The details of this model are still to be verified, but its underlying assumptions about the role of ignoring the errors during hierarchical skill acquisition have been tested in a simulation study outlined in the next section.

V. Simulation

We investigated how ignoring the errors could help a robot (shown in Fig. 4) to learn new skills in an approximate optimal control framework. For the purpose of our study, the framework had a two-layer structure. The top layer, was a decision making layer, that was trained using standard Q-learning to select appropriately for a given context, one of the three possible behaviors, that is reaching, walking or no response. Herein, we made use of a standard inverse kinematics controller for the reaching action, and only the walking module was trained using standard Q-learning.

The state space of the decision making module was a discretized distance to the goal (6 states in our case changing by $2cm$). The goal of the modul was to select one of the possible sub-modules depending on their predicted action outcome. The module received a reward ($R = 60$) when the selected action was successful, and a punishment ($R = -30$) in the opposite case. The walking module had 6 different predefined states and actions, each state was described by 8 joint angles (4 for each leg). The goal of the module was to learn how to alternate from one state to another so that the robot does not loose balance, and it moves forward at the same time. The module received a partial reward for getting closer to the goal ($r = 10$), and negative reward for moving backwards ($r = -3$). When the robot reached the goal the module received additional reward ($r = 60$). Any action that resulted in loosing balance was punished ($r = -30$). In the simulations, epsilon greedy action selection was used with $\epsilon = 0.1$.

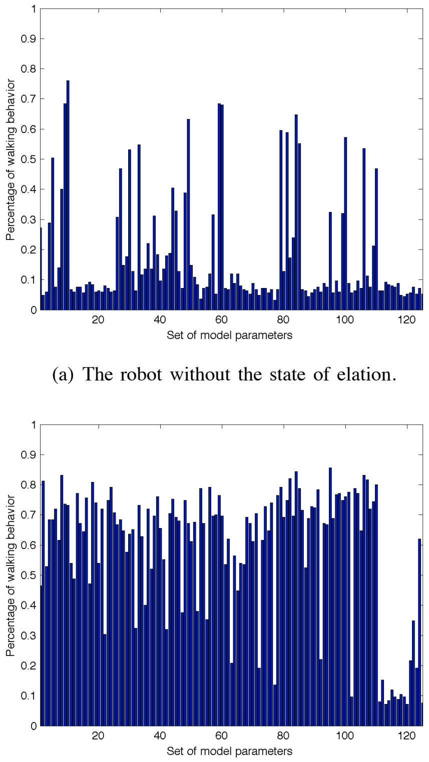
The simulation started with a robot not able to walk. The action of walking was available for selection, but its execution resulted in no movement. We simulated the onset of walking at



(a) The robot without the state of elation.

(b) The robot with the state of elation.

Fig. 5. The percentage of behavior selection.



(a) The robot without the state of elation.

(b) The robot with the state of elation.

Fig. 6. The percentage of walking behavior selection.

$w = 40$ epochs. Until the onset of walking the distance to the object (close or far distance) was changed randomly with 40% probability of change. After the onset of walking the object was placed only far away from the robot. We tested the robot in two different scenarios: without the state of elation, and with the state of elation. The state of elation was simulated by ignoring the negative outcomes of the actions in the decision making layer.

The settings and thus the behavior of the robot before the onset of walking was the same in both scenarios. Therefore, only the results of the simulations after this period are shown. As the robot chose actions with certain probability, the results of the simulations may vary across trials. We present the average results over 10 different trials. As it can easily be seen in Fig. 5(a), the robot without elation learned that the object is not reachable, and the probability of selecting the "no response" behavior was very high during the entire experiment. The robot had almost no opportunities to practice the walking behavior. On the other hand, the robot with the state of elation (shown in Fig. 5(b)), after 13 epochs started to select walking behavior more frequently making it possible for the walking module to improve.

As the results of the simulations may strongly depend on the values of reward, we repeated the simulation for different configurations of rewards. We varied the values for partial reward for getting closer to the goal in the walking module ($r \in \{5, 10, 15, 20, 25\}$), and the reward for successful action selection ($R \in \{30, 60, 90, 120, 240\}$) and punishment ($P \in \{-30, -60, -90, -120, -240\}$) for failure in reaching a goal in the decision making module. The results for total of 125 different configurations are shown in Fig. 6(a) and Fig. 6(b). As it can be seen, just a few configurations for the robot without elation allow the walking behavior to be selected more often. Thus, introducing the state of elation, facilitated in many cases the transition from selecting no response behavior towards selection of the walking behavior.

VI. Discussion

In terms of the dynamic systems approach [4], we may conceptualize the role of disregarding the error as follows. Assuming that the behavior of the infant is governed by a dynamic system component for decision making, and another one for execution of movement, the performance-dependent reward signal would be one of the control parameters of the decision making component. In the stable case where behaviors have been learned well (for instance to reach for near objects), negative rewards during exploratory actions would lead to further stabilization of the already learned attractors. If, however, the negative reward is ignored, i.e. the control parameter is changed, existing attractors might be destabilized.

This in turn would make it easier for the system to switch to other attractors, giving their corresponding movements more chance to be practiced in a new context where they would normally not be chosen. Over time, this practice might lead to new stable attractors even under consideration of the error signal once the effect of high dopamine state wears off.

As robots are expected to be active participants in humans' daily life, they need to be able to constantly learn and improve their abilities autonomously. The conceptual model and its simplified implementation in the simulation study of this paper offer one possible mechanism contributing to such adaptive behavior acquisition.

VII. Future work

As the preeliminary result with the robot simulator seems to confirm the viability of our approach, the next step in our research is to implement the conceptual model in more detail and evaluate its ability to account for the behavioral data in [1], [2]. Furthermore, we will perform a series of experiments with a real M3-neony humanoid robot, and study the dependence of the results on parameter settings in the simplified version of our model presented in Sec. IV.

VIII. Conclusion

The core idea behind the model was that the level of sense of control determines how much the negative outcome of the action is taken into account for decision making. Omission of the errors was suggested to enable selection of different behaviors in a context when they normally would not be selected providing more learning opportunities for fine-tuning these behaviors.

Acknowledgment

This research was supported by: Ministerio de Ciencia y Innovacion (FPU grant AP2007-02565); Grants-in-Aid for Scientific Research (Research Project Number: 22220002). The authors would like to thank Fabio Dalla Libera for providing the source code of the M3-neony robot simulator.

References

[1] B. J. Grzyb, A. P. del Pobil, and L. B. Smith, "Reaching for the unreachable: (mis)perception of body effectivity in older infants," manuscript in preparation.

[2] ——, "Reaching for the unreachable: the cause or the consequence of learning to walk," manuscript in preparation.

[3] B. J. Grzyb, J. Boedecker, M. Asada, A. P. del Pobil, and L. B. Smith, "Trying anyways: how ignoring the errors may help in learning new skills," in *First Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2011.

[4] E. Thelen and L. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, 1994.

[5] J. D. Cohen, T. S. Braver, and J. W. Brown, "Computational perspectives on dopamine function in prefrontal cortex," *Current Opinion in Neurobiology*, vol. 12(2), pp. 223–229, 2002.

[6] G. A. Barr, S. Moriceau, K. Shionoya, K. Muzny, P. Gao, S. Wang, and R. M. Sullivan, "Transitions in infant learning are modulated by dopamine in the amygdala," *Nature Neuroscience*, vol. 12(11), pp. 1367–1369, 2009.

[7] B. Abler, H. Walter, and S. Erk, "Neural correlates of frustration," *Neuroreport*, vol. 16(7), pp. 669–72, 2005.

[8] W. Schultz, "Multiple reward signals in the brain," *Nature Reviews Neuroscience*, vol. 1, pp. 199–206, 2000.

[9] W. Schultz, P. Dayan, and P. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275 (5306), pp. 1593–1599, 1997.

[10] K. Doya, "Metalearning and neuromodulation," *Neural Networks*, vol. 15(4-6), pp. 495–506, 2002.

[11] R. Sutton, "Learning to predict by the methods of temporal differences," *Machine Learning*, vol. 3(1), pp. 9–44, 1988.

[12] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[13] K. C. Berridge, "The debate over dopamine's role in reward: the case for incentive salience," *Psychopharmacology*, vol. 191, pp. 391–431, 2007.

[14] T. V. Maia and M. J. Frank, "From reinforcement learning models to psychiatric and neurological disorders," *Nature Neuroscience*, vol. 14, pp. 154–162, 2011.

[15] M. R. Delgado, R. H. Frank, and E. A. Phelps, "Perceptions of moral character modulate the neural systems of reward during the trust game," *Nature Neuroscience*, vol. 8, pp. 1611–1618, 2005.

[16] E. Leibenluft, B. A. Rich, D. T. Vinton, E. E. Nelson, S. J. Fromm, L. H. Berghorst, P. Joshi, A. Robb, R. J. Schachar, D. P. Dickstein, E. B. McClure, and D. S. Pine, "Neural circuitry engaged during unsuccessful motor inhibition in pediatric bipolar disorder," *Am J Psychiatry*, vol. 164, pp. 52–60, 2007.

[17] P. Najt, J. Perez, M. Sanches, M. Peluso, D. Glahn, and J. Soares, "Impulsivity and bipolar disorder," *European Neuropsychopharmacology*, vol. 17, pp. 313–320, 2007.

[18] S. Johnson, "Mania and dysregulation in goal pursuit: a review," *Clin Psychol Rev*, vol. 25, pp. 241–262, 2005.

[19] "American psychiatric association (2000). diagnostic and statistical manual of mental disorders. washington, dc."

[20] B. Abler, I. Greenhouse, D. Ongur, H. Walter, and S. Heckers, "Abnormal reward system activation in mania," *Neuropsychopharmacology*, vol. 33, pp. 2217–2227, 2008.

[21] W. Schultz, "The reward signal of midbrain dopamine neurons," *News Physiol. Sci.*, vol. 14, pp. 249–255, 1999.

[22] L. Leotti, S. Iyengar, and K. Ochsner, "Born to choose: the origins and value of the need for control," *Trends Cogn Sci.*, vol. 14(10), pp. 457–463, 2010.

[23] D. J. Shapiro, C. Schwartz, and J. Astin, "Controlling ourselves, controlling our world. psychology's role in understanding positive and negative consequences of seeking and gaining control." *Am Psychol.*, vol. 51(12), pp. 1213–30, 1996.

[24] M. J. Frank, "Hold your horses: A dynamic computational role for subthalamic nucleus in decision making," *Neural Networks*, vol. 19, pp. 1120–1136, 2006.

[25] M. H. Johnson, *Developmental psychology: an advanced textbook*. Psychology Press, 1999, ch. Developmental Neuroscience, pp. 199–230.

[26] H. T. Ghashghaeia, C. C. Hilgetag, and H. Barbasa, "Sequence of information processing for emotions based on the anatomic dialogue between prefrontal cortex and amygdala," *Neuroimage*, vol. 34(3), pp. 905–923, 2007.

[27] M. Milad and G. Quirk, "Neurons in medial prefrontal cortex signal memory for fear extinction," *Nature*, vol. 420(6911), pp. 70–74, 2002.

[28] T. Johnstone, C. M. van Reekum, H. L. Urry, N. H. Kalin, and R. J. Davidson, "Failure to regulate: Counterproductive recruitment of top-down prefrontal-subcortical circuitry in major depression," *The Journal of Neuroscience*, vol. 27(33), pp. 8877–8884, 2007.

[29] P. O'Donnell, *Handbook of basal ganglia structure and function*. Academic Press, 2010, ch. Gating of limbic input to the ventral striatum, pp. 367–37.

[30] R. A. Depue and P. F. Collins, "Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion," *Behavioral and Brain Sciences*, vol. 22, pp. 491–569, 1999.

[31] M. E. Jackson, A. S. Frost, and B. Moghaddam, "Stimulation of prefrontal cortex at physiologically relevant frequencies inhibits dopamine release in the nucleus accumbens," *Journal of Neurochemistry*, vol. 78, pp. 920–923, 2001.

[32] D. Purves, G. J. Augustine, D. Fitzpatrick, W. C. Hall, A.-S. LaMantia, J. O. McNamara, and L. E. White, Eds., *Neuroscience, Fourth Edition*. Sinauer Associates, Inc., 2007.

[33] R. A. Depue and W. G. Iacono, "Neurobehavioral aspects of affective disorders," *Annu Rev Psychol.*, vol. 40, pp. 457–92, 1989.

[34] A. D. Arco and F. Mora, "Neurotransmitters and prefrontal cortexlimbic system interactions: implications for plasticity and psychiatric disorders," *Journal of Neural Transmission*, vol. 116, pp. 941–952, 2009.

PM2-1

Mutual adaptive interaction between robots and human based on the dynamical systems approach

Tetsuya Ogata

Abstract – **This paper introduce a concept of "mutual adaptive interaction" with primitive experimental examples. Interaction manners between human and robots usually pre-designed precisely. On the other hand, signs used by humans continuously evolve in form and meaning through interactions. We presume the open-ended learning is one of the essential aspects of the sign (primitive language) and is caused by iteration of misunderstanding and re-adaptation in mutual adaptive interactions. Following the concepts, our studies treating human navigation and robot's multimodal interaction are shown as examples.**

I. Introduction

Recently, much attention is paid to studies of robot "programming by demonstration". It would relate to the concept of imitation learning. Though the concept of imitation itself is quite simple, it has various interesting and difficult problems robot system should overcome. One of them is that robot should judge what motions should and/or can be imitated and what motions should not and/or cannot. This interpretation depends not only on the ability of robot hardware but also on the context of learning process. Even if human shows same demonstration, robot has to change its interpretation, i.e. what should be learnt from the demonstration, according to the learning condition. This kind of learning often happens in the interaction between human's child and the caregiver. Children and their caregiver seem to change the interpretation by estimating the intention of the demonstration. If the estimation fails, they notice the miscommunication, and they modify the conjecture to re-adapt to the other's intention. The repetition of miscommunication and re-adaptation leads to further evolution of interactions.

We have tackled this evolutional interaction between humans and robot by taking three approaches, embodied system, mutual adaptation, and self-consistency. We actually use neuro-dynamical systems to deal with these three aspects. This paper introduces our previous studies especially concerning on the open-ended interactions treating human navigation [1] and robot's multimodal interaction [2] as examples.

Section II describes the study on a navigation robot with a human covered his/her eyes through mutual adaptation, Section III describes the study on sign evolution between small robots through multimodal interaction, Section IV discusses three concepts for designing the open-ended interaction system, and Section V concludes the paper.

II. Mutual Interaction in Human Navigation Task

This section introduces our studies on implicit mutual learning between a human and a robot. Concretely, we designed a navigation task in which a humanoid robot, Robovie and a human participant navigate together in a given workspace (Fig. 1). The experimental environment was L shaped course whose outside walls were marked red and blue for every block. The robot and the participant whose eyes are covered hold their arms together and navigate the workspace. We asked the participant to travel clock-wisely in the workspace as quickly as possible without hitting the wall. We also informed to the subject the following points: 1) the robot moved just avoiding the walls using its range sensor without any planning, 2) the subject could control the robot motion by changing the angles of robot's elbow joint, and 3) the robot changed the behavior by learning the environment and the subject's control manner. The robot movement is determined by adding two motor forces; one is the output from a neural network in the robot and the other is the participant's directional control. The performance is measured by the travel time period at each trial.

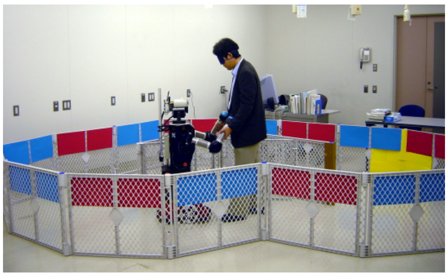


Figure 1. Navigation Task

An interesting point of this collaboration task is that the sensory information is quite limited for both the robot and the participant. The robot can access only local sensory information such as range sensors and a poor vision system only detecting vague color information. The participant's eyes are covered during the navigation task. But the participant is allowed to look around the workspace before the experiments. The participant has to guess his/her position by means of the interactive force felt between the robot and his/her arms utilizing his/her memorized image of the workspace geometry. Both sides attempt to acquire the for-

Tetsuya Ogata is with the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto, Japan, and also with PRESTO, Japan Science and Technology Agency, Japan, ogata@kuis.kyoto-u.ac.jp

ward model of anticipating the future sensory image from past experiences.

A. Recurrent neural network and consolidation learning

In our experiment, the current sensory inputs may not tell the exact position of the robot due to the *sensory aliasing* problems. To solve this problem, we use the recurrent neural network (RNN) storing/self-organizing the contextual information in the context units. Figure 2 shows the RNN architecture designed for our robot. It has two modes of operations. The first mode is the open-loop mode that is one-step prediction by using the inputs of the current sensory-motor values. The second mode is the close-loop mode in which the output re-entered the input layer through the feedback connection. By iterating this with the closed loop, the RNN can generate an arbitrary length of the look-ahead prediction for future sequences. The RNN was trained by using the back propagation through the time (BPTT) learning method [3].

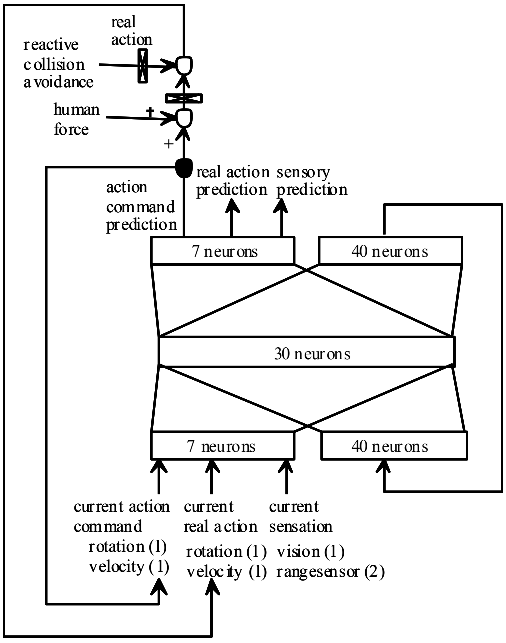


Figure 2. Recurrent neural network implemented in Robovie

B. Consolidation Learning

It is generally observed that if the neural net attempts to learn a new sequence, the contents of the current memory are severely damaged. To overcome this problem, we used the consolidation learning technique. In this method the newly obtained sequence pattern is stored in the "hippocampal" database. The RNN, which corresponds to the neocortex rehearses the various memory patterns by close-loop with random initial state, and these patterns are also saved in the database. The ensembles of such various rehearsed sequences actually represent the structure of the past memory in the dynamical systems sense. The RNN is trained using both the rehearsed sequential patterns and the new experience.

C. Development Process of the Interaction

The interactions were evaluated in 15-trial navigations with seven male subjects. The RNN with a usual learning method and the RNN with consolidation learning were compared in the experiments. In consolidation learning the teaching data were the current sequence pattern and the three rehearsal patterns.

As the experimental results, though the performance of the RNN with the usual learning method stagnated, the performance of the RNN with the consolidation-learning algorithm continued to improve. The results of the questionnaire showed that the RNN with the consolidation-learning algorithm gave the best mental impressions.

To analyze the effect of consolidation learning, we examined the robustness of the RNN dynamics by looking at its initial sensitivity characteristics. Both RNNs with the usual learning and consolidation learning were tested to generate the output sequences with the noise addition. It is observed that output trajectories of the RNN by the usual learning tend to diverge more than those by consolidation learning. This robustness characteristic of the RNN seems to be directly related to the Operability in the mental impressions.

We also investigated the development process of the RNN. We compared the rehearsed trajectories of the RNNs obtained in each trial in our experiments. Figure 3 shows a typical transition of the amount of change of the trajectories. It was confirmed that the transition with the consolidation-learning method had three peaks (1st, 4th, and 8th trial) and decreased gradually. This means that the phase transitions occurred three times in the development process and became stable. On the other hand, the transition with the usual-learning method had no clear peak and increased. This means that there was no clear phase transition in the process and it became unstable.

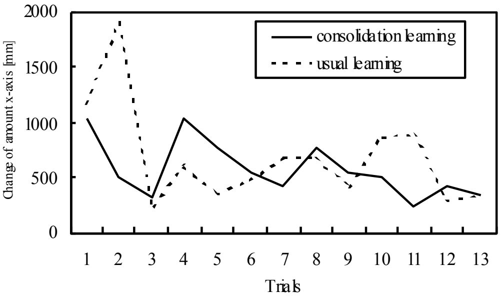


Figure 3. Transition of the Change of Rehearsal Trajectory

The trials in which the phase transition occurred with the consolidation-learning method corresponded to the trials in which the performance improved drastically. Also sharp improvement of mental impression could be recognized in the comments of the questionnaires of these trials. The

RNN with the consolidation-learning method might be useful to realize a robot-human open-ended interaction.

III. Sign Emergence in Robot Imitative Interaction

This section introduces another study on mutual adaptation between two robots conducting multi-modal interaction. In our target interaction, two robots try to convey a motion/voice pattern to the other robot using a voice/motion pattern as a sign. If the interpretation of the sign is shared between them, they have conveyed their intention to the other correctly, otherwise incorrectly.

A. *Interaction Model*

An overview of our interaction model is shown in Fig. 4. An agent robot Keepon has a pair of NNs to interpret signs and two RNNs called MTRNN (Multiple Time scale RNN) for voice and motion. For example, a robot interprets a voice sign as a motion pattern as follows. (1) Recognition: The observed voice is transformed into voice parameters by Voice MTRNN. (2) Interpretation: The voice parameters are transformed into motion parameters by Interpretation NN. (3) Generation: The motion parameters are transformed into a motion pattern by Motion MTRNN.

Through interchanging of voice and motion, two robots modify their interpretation of signs to adapt to the other by retraining their Interpretation NN.

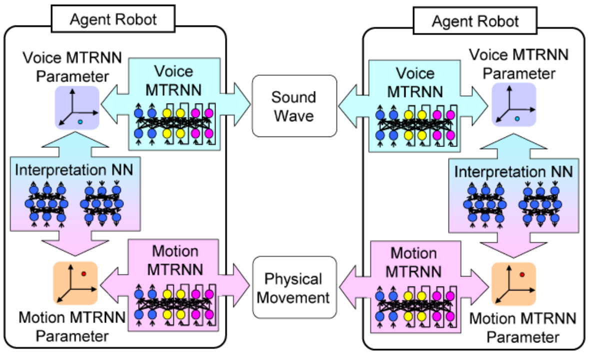


Fig. 4 Overview of Interaction Model

B. *MTRNN model*

We utilized MTRNN for the recognition and generation of signs (voice and motion). The model works as a recognizer and a generator of actions by learning motor information and sensory information simultaneously. Furthermore it recognizes and generates unknown actions with its generalization capability. This capability provides the diversity of actions used as signs that is essence of evolutionary interaction.

MTRNN, proposed by Yamashita et al. [4], is an extended RNN model (Fig. 5). This model deals with sequential data through calculating the next state $S(t+1)$ from the current state $S(t)$. The model is composed of three neuron groups, each with an associated time constant. The three groups in increasing order in the time constant, are input/output nodes (IO), fast context nodes (C_f) and slow context nodes (C_s).

C_{s0}, initial value of the C_s, is self-organized depending on a dynamical structure among training patterns through the process that connection weights, which are shared by all patterns are updated.

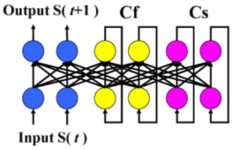


Figure 5. Composition of MTRNN

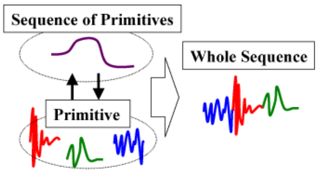


Figure 6. Dynamics Representation of MTRNN

We used a small life-like robot “Keepon” [5] for our experiments (Fig. 7). We set two Keepons facing each other at intervals of 230 mm, and place speakers beside them.

To synthesize sound for the Keepons’ voice, we used “Maeda model”: the vocal tract model. This model has seven parameters that determine the vocal tract shape. By using this Maeda model, we can apply the framework of sensori-motor integration to recognize and generate voice.

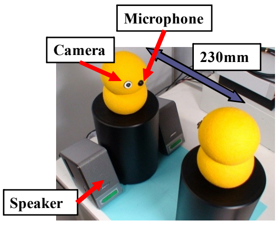


Figure 7. Experimental setup (top)

C. *Experimental Result*

One of the experimental results is shown in Fig. 8. The graph at the top of Fig. 8 shows the sequence of communication error for Keepon A and B. The others show voice and motion patterns generated in segment I, II, and III of the interaction period. The result of the experiment revealed the following facts. There is repetition of coherent states with low error and incoherent states with high error in the interaction. In coherent states, the two robots conveyed their intention to the other correctly, and interacted stably using similar voice-motion pairs (cf. segment I, III). On the other hand, in incoherent states, they failed to convey their intention and showed irregular behaviors (cf. segment II). The signs used for communication in coherent states (e.g. segment I and III) are different. Moreover the voice and motion patterns used as signs are different from the training patterns described in the section III-B.1. The communication error tended to decrease on the whole, but the interaction kept evolving without convergence.

IV. Discussion

In this section, I would like to discuss the important aspects of above two examples of dynamical interaction. In two interactions between robots and human, we found that the interaction repeats phase transitions of miscommunication and re-adaptation alternately, and emerges new interaction manners depending on robots' and humans' body dynamics through the generalization capability of neuro-dynamical system. It is interest that the similar phenomenon is also observed in the study of the robot cognition based on dynamical system [6].

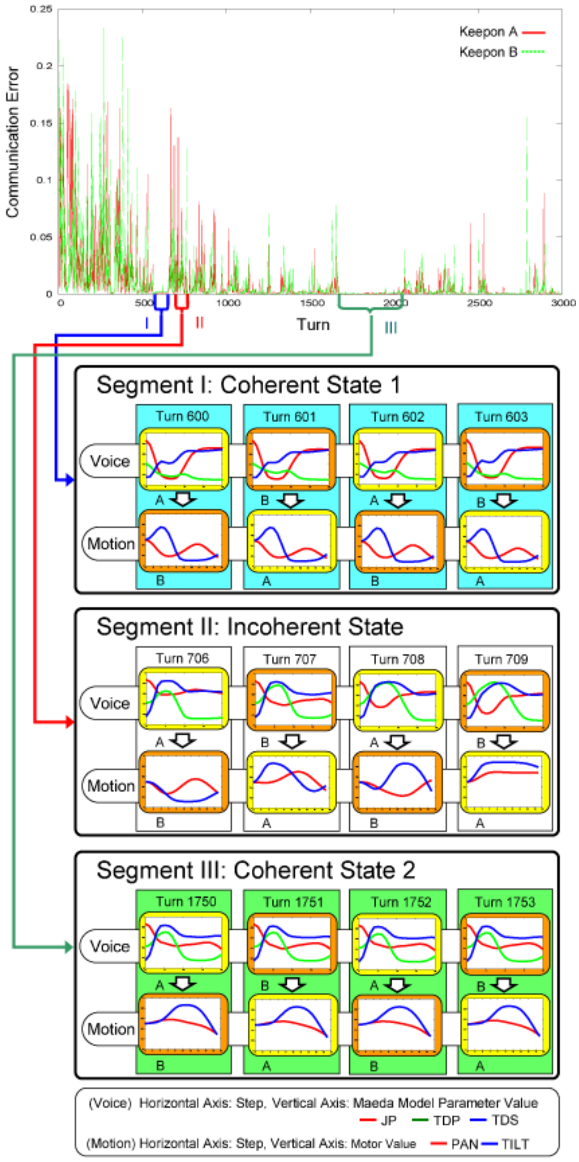


Fig. 6 Result of the Interaction Experiment

Note that these interactions do not include any random factors (noises) which cause the phase transition. The systems are purely dynamical systems without stochastic processes. Each agent (robot and/or human) tries to predict the other's reactions and behaviors in the mutual adaptation process. However, even after such a long interaction, the slight prediction error remains in each agent, because they interact in the physically continuous and complex world. Thus, the agent in the real world always has to adapt to the confliction between the internal dynamical model and the external world including the interaction partners. The small error could be a trigger which causes the drastic changes like phase transition.

This instability nature in dynamical interaction systems is essential for emergence of various type of interaction manner including primitive signs. Moreover, their behaviors seem to relate to the concept of "chaotic itinerancy [7]", though the interactions described in this paper cannot be regarded as pure dynamical systems in the sense that they update its development rule in the neural networks.

V. Conclusion

This paper proposes a research theme of "interaction emergence" for flexible human-like interaction systems to prevent user's boring.

We tackle this theme by taking a synthetic approach coupling multiple dynamics of the neural networks, the robot system, and the environment. Concretely, the concept of open-ended interaction is introduced. We also show the neuro-dynamical systems models which can adapt the dynamics of mutual adaptation between robots and human.

Future work includes mathematical investigation of RNN model and implementation of the model to large-scale systems. The ultimate goal is to obtain explanation model for emergence of interaction.

Acknowledgment

This research was supported by Grant-in-Aid for Creative Scientific Research 19GS0208, Scientific Research (B) 21300076, and RIKEN.

References

[1] T. Ogata, S. Sugano, and J. Tani: Open-end Human-Robot Interaction from the Dynamical Systems Perspective -Mutual Adaptation and Incremental Learning, Advanced Robotics, VSP and Robotics Society of Japan, Vol.19, No. 6, pp. 651-670, July, 2005.

[2] W. Hinoshita, T. Ogata, H. Kozima, H. Kanda, T. Takahashi, H. G. Okuno: Emergence of Evolutional Interaction with Voice and Motion between Two Robots using RNN, Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS-2009), pp.4196-4291, 2009.

[3] D. Rumelhart, G. Hinton, and R. Williams, *Learning internal representation by error propagation*. MIT Press, 1986.

[4] Y. Yamashita and J. Tani, "Emergence of Functional Hierarchy in a Multiple Timescales Recurrent Neural Network Model: A Humanoid Robot Experiment," PLoS Computational Biology, Vol. 4, No. 11, e1000220, (2008).

[5] H. Kozima, C. Nakagawa, and H. Yano, "Using robots for the study of human social development," AAAI Spring Symposium on Developmental Robotics, (2005).

[6] J. Tani, Model-based learning for mobile robot navigation from the dynamical systems perspec- tive, IEEE Trans. Syst. Man Cybernet. B (Special Issue on Robot Learning) 26, 421–436 (1996).

[7] I. Tsuda: "Chaotic itinerancy and Cantor coding can afford a mathematical basis of information processing in cortical dynamics", Invited, Riken, (2003).

PM2-2

The Cognitive Nature of Action - A Bi-Modal Approach towards the Natural Grasping of Known and Unknown Objects

Kai Essig, Jonathan Maycock, Helge Ritter, and Thomas Schack

***Abstract*—To gain a deeper understanding of human manual dexterity we investigate the grasping of known and unknown objects by recording hand kinematics (using a CyberGlove) and eye movements (with a mobile eye-tracking system) in a two-stage experiment: subjects first had to look at and grasp 8 known and 8 unknown objects, which they had neither seen nor handled before. In the second stage they had to interact with the objects. Mental representations of the objects were analyzed using a hierarchical sorting paradigm both, before and after contact with the objects. The results show that subjects first separated the objects according to their shape and size. However, when tested after having had haptic contact with the objects, they then tended to cluster the objects based on their functional aspects. This first aspect was also found in the grasping data: clusters in the principle component (PC) space can first be distinguished by object size and shape. For the unknown objects subjects showed less complex movements in the first part of the interaction phase – presumably where they tried to figure out what the unknown object is, followed by more complex postures in the second part of the interaction phase. We also found that unknown objects require more intensive cognitive processing, indicated by a slightly higher number of fixations with shorter saccade lengths and a wider attention distribution when interacting with them. Our results support the notion that there are close links between perception, conceptual factors and grasping movements when confronted with objects being seen for the first time. These insights are of importance for the field of cognitive robotics, where it is hoped that robots can select and adjust actions flexibly according to the given situation.**

I. Introduction

Manual action is a skilled behavior requiring intricate control of the musculoskeletal system of the human hand. Especially when we grasp, manipulate and interact with objects, movements of the hand have to be accurately adapted to the object's shape and the task we want to perform. During such skilled movements, the large number of degrees of freedom (DOF) in the human hand has to be controlled in a highly efficient way [1]. There are many examples of carefully hand-programmed robots which can carry out impressive manipulation tasks, such as using tools to assemble complex objects and pouring water into a glass [2]. However, fully autonomous grasping of previously unknown objects remains a challenging problem, because it is difficult to obtain an accurate 3D reconstruction of a novel object to find suitable grasping positions [3-5]. Some proposed solutions are based on geometric models of the objects, obtained by stereovision or a 3D laser sensor, or by learning the contact points between the fingers and the object surface and therefore imitating human grasping behavior [3],[6],[7]. In [8] an architecture for grasp synthesis inspired by the human neurophysiology of action-oriented vision was presented, proposing an adaptation of brain models to the peculiarities of robotic setups. Additionally, there are learning algorithms that use a data glove and neural network techniques to realize a mapping of human and artificial hand workspaces, to learn the different joint values for the robotic hand from the human grasping data [9],[10].

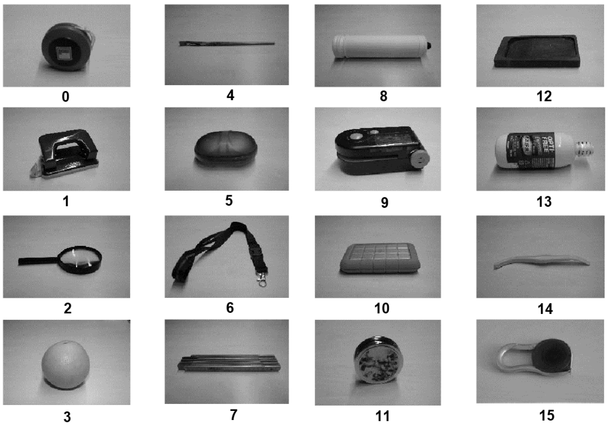


Fig. 1. The 8 known (left two columns) and 8 unknown (right two columns) objects. 0. yo-yo, 1. hole puncher, 2. magnifying glass, 3. orange, 4. brush, 5. soap, 6. lanyard, 7. folding rule, 8. balloon pump, 9. CD cleaner, 10. HDD, 11. (Chinese style) magnifying glass, 12. ink stone, 13. contact lenses solution, 14. orange peeler, 15. tape measure.

These approaches work quite well in controlled world environments. However, when it comes to action in the real world, precise models are an unrealistic idealization and uncertainty about many parameters prevails, such as geometric shape, object weight, surface friction, or mechanical stiffness. For many common daily tasks object models as they are currently used in robotics, with their focus on an explicit description of detailed material physics, are simply unfeasible to parametrize properly [2]. In

Manuscript received August 15, 2011. This work was supported by the Center of Excellence Cognitive Interaction Technology (CITEC) at Bielefeld University, Germany.

Kai Essig and Thomas Schack: Neurocognition and Action Research Group, Faculty of Psychology and Sport Science, and Center of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University, PB 100 131, 33501 Bielefeld, Germany (corresponding author: Kai Essig; phone: +49-521-106-6057; fax: +49-521-106-6432; e-mail: kai.essig@uni-bielefeld.de).

Jonathan Maycock and Helge Ritter: Neuroinformatics Group, Technical Faculty, and Center of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University.

contrast, human actions are heavily informed by large amounts of knowledge about the characteristics of the encountered objects and goals, and how to counteract the numerous disturbances and mishaps that usually occur during even moderately complex movements [2].

II. Cognitive Nature of Action

Results from movement research and neurophysiology led to the cognitive-perceptual hypothesis of central motor planning. It suggests that movement representations have the same spatio-temporal structure within the brain as the actual movement. Experimental studies [11] showed that representational frameworks were organized in a hierarchical tree-like structure and revealed a good match with the biomechanical demands of the task. After measuring kinematic parameters, we previously investigated the relationship between the structure of motor representation and the kinematic parameters of different movements [11-13]. These studies revealed significant correlations between kinematic parameters (time structure, angles according to the take-off-phase, tilt angle, angular velocities, etc.) of movement and the corresponding parts of mental representations. According to this perspective, the representation structure can access all the topological properties that support the movement. It has been shown that mental (action) representation plays a central role in the control and implementation of actions. As current studies show, human grasping movements are cognitively represented on the basis of movement concepts (e.g., *Basic Action Concepts*) and build on effect-orientated target codes (in relation to space rather than to body) [13],[14]. In [1] we compared grasps directed towards real and virtual spherical objects varying in size. We found that the grasping movement is influenced by object characteristics (i.e., object size) at an early stage of the movement. We also found a separation of smaller objects from larger ones in the analysis of mental representation of grasps and concluded that grasping movement is strongly influenced by conceptual factors. Altogether, our experimental results support the hypothesis that voluntary movements are executed in a person-task-environment constellation and are directly stored in memory through representations of their anticipated perceptual effects. A technique for investigating mental representation of movements in athletes, the *structural dimensional analysis-motoric (SDA-M)* has been used to study different fields of sports such as volleyball, gymnastics, sky diving, and dancing. It has also been applied to everyday scenarios, such as drinking from a cup, using tools, and grasping objects of different sizes [13], [14].

The importance of attention in guiding manual actions has seen renewed interest of late [15], [16]. In order to achieve good eye-hand coordination, hands and eyes must work together in smooth and efficient patterns. Johansson et al. [17] analyzed the coordination between gaze behavior, fingertip movements, and movements of a manipulated object. Obligatory gaze targets were those regions where contact occurred. They concluded that gaze supports hand movement planning by marking key positions to which the fingertips are subsequently directed.

We also investigate whether the Eye-Mind Hypothesis [18] is supported by the captured data. This states that the number and the distribution of fixations reflect the degree of cognitive processing required for the understanding of particular scene regions. Long fixation durations and short saccade lengths signify a fine and deep processing of a particular image region and is an indicator that the understanding of visual input is quite difficult. In contrast, long saccade lengths and short fixation durations indicate a fast and coarse scanning of a particular scene region, signaling that the information content of that particular image region is easy to process or less important for the current task.

In this paper we use a bi-modal approach to study the perceptual and grasping behavior of humans while picking up and interacting with known and unknown objects. We additionally investigate how the mental representation structures of the different objects in the long-term memory are build up and how this representation changes after interaction with the objects. Changes in the visual and grasping behavior of subjects are also studied. Here we present a first test of the SDA-M in a grasping task in which there are known and unknown objects. We expect that the subjects are first guided by their visual perception when clustering the objects. Objects, which they cluster together, may be approached by using similar grasping postures. After performing a typical interaction with the objects, subjects got an impression of how to handle them. We expect that this lead to a different hierarchical clustering of the objects. These changes in the mental representation structures may also be reflected in a different grasping behavior.

III. Methods

To test our bi-modal setup a simple three-stage experiment was designed in which five subjects (three males, age from 25-30 years) were shown 8 known and 8 unknown objects (see Fig. 1). All subjects had normal or corrected-to-normal vision and had no known impairments related to arm or hand movements. All gave written informed consent to be part of the study and the experiment was carried out according to the principles laid out in the 1964 Declaration of Helsinki. Subjects performed all three experiments in the same order, starting with Experiment 1, directly followed by Experiment 2 and Experiment 3. Finally, they repeated Experiment 1 again. The experiments were carried out in the Manual Intelligence Lab of Bielefeld University, making use of its sophisticated multimodal set-up for investigating manual interactions [19]. During data collection, the subjects sat in front of a table (with dimensions l=210cm, w=130cm, h=100cm). Subjects wore an Immersion CyberGlove II wireless data glove (Immersion

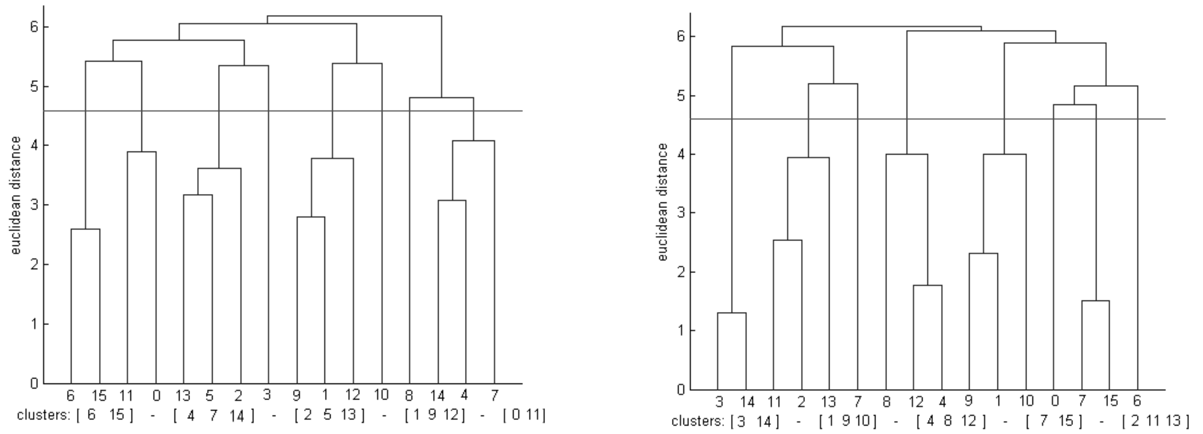


Fig. 2. Average cluster solution of all subjects before (left) and after (right) the grasping experiment. On the horizontal line the numbers of the single objects are represented as corresponding with the list shown in Figure 1. The Euclidean distance between the single objects is depicted on the y-axis. The solid horizontal line marks d_{crit}.

Corp., San Jose, CA; data acquisition rate: 100Hz; sensor resolution: <1°) on the right hand that allowed for the recording of whole hand kinematics (22 DOF) [9].

Two markers were placed on the table 40cm from the subject. The one on the left hand side indicated the position of the objects before being picked up. The marker on the right hand side marked the position where the subjects had to put the objects down. In order to get insights into human visual perception processes while handling known and unknown objects, we used an SMI iViewX (monocular) mobile eye tracking system [20] (sampling rate of 200 Hertz, with a gaze position accuracy < 0.5°) for our experiments.

A. *Experiment 1*

In a first experiment we analyzed how subjects clustered the 16 objects by means of the SDA-M. When performing Experiment 1 for the first time, subjects had to cluster the objects using the pictures (see Fig. 1), without having had haptic contact with them. The images (200 x 133 pixels) were presented in black and white in order to prevent clustering based on color information. This experiment took approximately 15 minutes. The subjects were seated in front of a laptop computer on whose screen they saw one of the 16 objects, which served as a reference object. The remaining 15 objects were presented underneath the anchoring unit aligned along one column. Each of these 15 objects were presented and subjects had to judge whether each one was "functionally related in day-by-day use" to the anchor object. If the subjects decided these objects were functionally related, they pressed the right arrow key and the current object shifted to the right side of the screen (positive list). Otherwise, subjects pressed the left arrow key and the object was shifted to the left side of the screen (negative list). This procedure repeated until all 15 objects were compared to the anchor object. Each object occupied the anchor position once. Using this splitting procedure, sixteen decision trees per subject were created. First, the binary decisions generated by each subject were used to calculate a correlation matrix, *R,* that is a reflection matrix of the mental representation structure of these objects in subjects' long-term memory. Low values (close to -1) mean concepts (or in our case, objects) are dissimilar or high values (close to 1) mean that concepts are similar. Next a distance matrix, *D,* is computed from the correlation matrix:

$D_{ik} = \sqrt{2N(m)} \cdot \sqrt{1 - R_{ik}}$, with N = number of concepts (for the 16 objects), *i* and *k* are the indices of the correlation matrix *R,* and *m* is usually equal to *N* (but is used as a correction factor if *R* is not square). The entries in the distance matrix reflect the Euclidean distances between individual concepts. Based on the distance matrix a greedy algorithm is used for a hierarchical cluster analysis with the distances based on the subjective distance judgments of all combinations of pairs of 16 objects obtained in the previous step. As a result we obtained the individual partitioning of the 16 objects, the so called dendograms (see Fig. 2). Cluster solutions were calculated for all individual subjects and for the whole group. Each cluster solution was established by determining an critical Euclidean distance d_{crit}, with all junctures lying below this value forming an individual concept cluster. The critical value d_{crit} depends on the number of concepts. The algorithm tries to calculate the optimal critical value $d_{crit,}$ but in cases of sub-optimal d_{crit} calculation, some manual tuning is possible. For further details on the splitting procedure see [13]. If two objects are often labeled as being "functionally related in day-by-day use", they have a small Euclidean distance, and this results in a low projection of the objects on the vertical line in the tree diagram (see objects 3 and 14 in Fig. 2 right). If two objects are not judged to be "functionally related in day-by-day use", and therefore not selected together during the splitting procedure, the Euclidean distance is big and the projection of the two objects is high in the tree diagram (see objects 0 and 11 in Fig. 2 left).

B. *Experiment 2*

In this experiment all subjects had to grasp 8 known and 8 unknown objects (see Fig. 1). First, the eye-tracker and CyberGlove were calibrated. Then, subjects had to pick up each object, hold then for 10s and finally put them down again. Shortly before the time came to an end, the experimenter gave the subjects a verbal signal. Then subjects moved their arm to the right marker and put the object down. Subjects then had to move their hand back to the starting position and wait for the next trial. Subjects were instructed to place the right hand at the starting position between the two markers at the edge of the table in front of them and wait for a “go” signal to grasp the objects. Each object was presented in a suitable random order.

C. *Experiment 3*

Here, the experimental procedure was nearly the same as for Experiment 2. However, this time subjects had to perform an interaction with each object for 20s. They had to actively explore the objects in order to discover what possible functions they had. For known objects this was clear, but for the unknown objects their functionality could only be determined by this interaction phase and sometimes it remained unclear. Because the task was to interact with the object, they could use both hands, but only the right hand was tracked. Again, their gaze movements were recorded with a mobile eye-tracking system. Finally, subjects had to return to Experiment 1 with the same stimuli.

IV. Results

A. *Experiment 1: Mental Representation of Objects*

In order to compare the cognitive representation structures before and after grasping the known and unknown objects, we computed the average dendrograms over all subjects from the results of the first time subjects partook in Experiment 1 (Fig. 2 left) and second time subject partook in Experiment 1 (Fig. 2 right). When comparing the results of the cluster analysis ($d_{crit} = 4.59$), we can see clear differences in the clustering of the objects between the dendrograms computed before interaction with the objects and after interaction. Initially, subjects had to judge if two objects were "functionally related in day-by-day use” just by comparing pictures of the objects, i.e., without seeing the objects in real or touching them. The resulting dendogram (see Fig. 2 left) reveals five clusters: (cluster 1: lanyard – tape measure, cluster 2: orange peeler – brush – folding rule, cluster 3: contact lenses solution - magnifying glass – soap, cluster 4: CD cleaner – ink stone – hole puncher, cluster 5: (Chinese style) magnifying glass – yoyo). Subjects found it difficult to find objects with similar functionalities, and thus the Euclidean distance of objects within clusters is quite high. Failing to find functional similarities, subjects seemed to cluster the objects according to similarities in shape and size. Only one cluster (cluster 3: contact lenses solution - magnifying glass – soap), was classified according to the functionality of the objects, i.e., cleaning functionality. Fig. 2 (right) shows the results of the SDA-M after subjects had haptic contact with the objects and had performed typical movements with them. The tree diagram again shows five clusters, but this time different objects were grouped together (cluster 1: orange – orange peeler, cluster 2: CD cleaner – HDD – hole puncher, cluster 3: balloon pump – ink stone – brush, cluster 4: folding rule – tape measure, cluster 5: (Chinese style) magnifying glass - magnifying glass - contact lenses solution). After using the objects, subjects grouped them according to their functionality and not according to similarities in their shape and size features. This is not only true for tools sharing similar function, like the folding ruler and the measuring tape, but also for objects that are used together in everyday tasks, such as the peeler and the orange or the brush and the ink stone. These results reveal that subjects were at first unfamiliar with the functionality of the unknown objects, but by interacting with them, their mental representations of these objects were updated and meaningful functional clusters could be formed. As can be seen from Fig. 2, the known objects are not clustered in the same way before and after the subjects had haptic contact with them. A possible reason is that the pictures were too small in order to recognize all objects correctly. After subjects interacted with the objects, they were more familiar with them (shown by the higher number of clusters according to functionality).

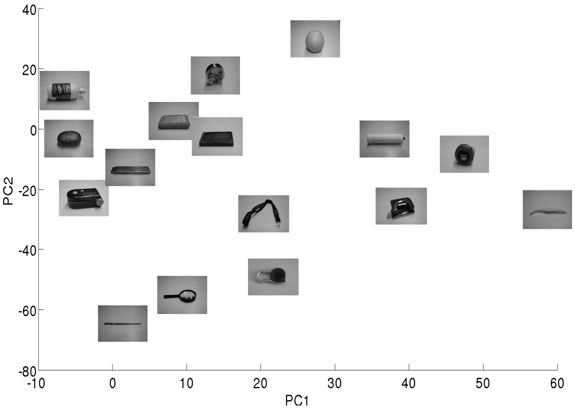


Fig. 3. All 16 objects in PC space (PCs 1 and 2) at the sampling time when subjects grasped them in Experiment 2. There is a clear separation between large and small objects by a diagonal running through the lanyard and balloon pump.

B. *Experiment 2: Grasping of the Objects*

We analyzed the gaze data with regard to the number of fixations, fixation durations, number of saccades and saccadic length in x- and y- directions. With regard to the fixations, we found only slightly higher number of fixations (19 versus 21) as well as longer fixation durations (386.4ms versus 400.9ms) for the unknown objects. There are also virtually no differences with regards to the saccade lengths between known and unknown objects (44.11 and 45.97 pixels in horizontal, and 60.19 and 59.21 pixels in the vertical direction for known and unknown objects, respectively).

A separation of smaller objects from larger ones can also be seen in the principal component space computed from the hand posture data recorded in Experiment 2. Subjects use similar movement synergies for similar sized objects [1], [21]. Fig. 3 shows that there is a clear separation of the objects according to object size: smaller objects are located in the lower part of the space and larger objects in the upper part. No differences in terms of functionality could be observed in the data. In order to find differences in the functionality, subjects need to have a suitable cognitive representation for the objects according to their “day-by-day” use.

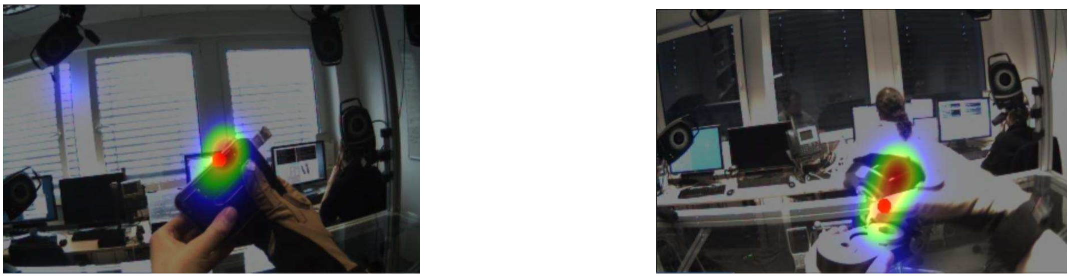


Fig. 4: Attention maps showing main areas of fixation activity when subject 1 is interacting with a known object (hole puncher, left) and an unknown object (CD cleaner, right). Colors are used to visualize the extent to which image areas are fixated, where red means that the area is highly fixated, black means that the area was not fixated at all.

C. *Experiment 3: Interacting with the Objects*

The analysis of the recorded gaze videos of the interaction phase revealed that for the case of unknown objects, subjects needed nearly the first half of the available experimental time to analyze the objects in hand in order to find out how to interact with them. Therefore, we analyzed the gaze and grasping data separately for two time intervals: The first one was from 2s to 10s (1st_half), reflecting the interval when subjects mostly held the object in their hands for further analysis. The second interval lasted from 10s to 18s (2nd_half). This was the interval in which they tended to interact with the objects.

With regards to the fixation durations, we found a higher number of fixations (17.8 versus 18.5) as well as longer fixation durations (398.1ms versus 411.3ms) for the unknown objects. The number of fixations is only slightly higher and the duration is approximately 5% longer for the unknown objects. However, when interacting with the objects, there are clear differences with regard to the saccade lengths between known and unknown objects (49.61 and 62.36 pixels in horizontal, and 55.9 and 78.44 pixels in the vertical direction for known and unknown objects, respectively). The overall results are in line with the Eye-Mind Hypothesis: the number and distributions of fixations reflect the degree of cognitive processing required to understand the scene. Long fixation durations and short saccade lengths signify a fine and deep processing of a particular object. For the case of unknown objects, subjects required more cognitive processing in order to identify what the object was and then react with the object in a meaningful way. Subjects were not only able to immediately identify known objects, but also able to interacting with them. An attention map [22] was calculated to highlight which areas were looked at and how intensely they were studied [see Fig. 4]. In an attention map each pixel is assigned a value between 0 and 1 depending on how long the subjects focused on a particular pixel (values closer to 1 indicate higher attention by the subject). Color-spots indicate which areas were attended and how intensively the subject looked at them. The attention maps of subject 1 for the known object (hole puncher, see Fig. 4 left) shows a smaller attention area – thus is because the subject knows the object and can immediately perform a typical movement with it. For the case of an unknown object (CD cleaner, see Fig. 4 right), the attention area is larger, especially for the highly fixated area (red spot). This indicates that the subject had to analyze the object in detail in order to figure out what to do with it.

We also computed the variance of the first two PCs for each object over all subjects separately for each of the two interaction intervals (Fig. 5). Lower values indicate more complex movements. The figure shows that there are only small differences in the variances for the known objects between the two time intervals (58.16 and 57.45, respectively). Subjects know the objects and immediately start with the typical movements. In case of the unknown objects, there are slightly larger differences (60.61 and 56.02, respectively). This indicates that subjects made less complex movements initially as they first explored the object, but that in the second half of the interaction phase they performed more complex movements. We stress that these results are not conclusive, but are certainly a good starting point for further research.

V. Conclusions and Outlook

Using a bi-modal approach we investigated motor synergies of hand postures as subjects interacted with 8 known and 8 unknown objects. Our results clearly reveal differences in the mental representation structures for the objects in the long-term memory before and after the experiment. First, subjects were not familiar with all objects and clustered them according to shape and size features. Later, they got familiar with them – resulting in a higher number of object clusters according to functionality.

The analysis of the gaze data is, however only weakly pronounced, in line with the Eye-Mind Hypotheses, revealing a higher cognitive load for the analysis of the unknown objects (especially with respect to the saccadic

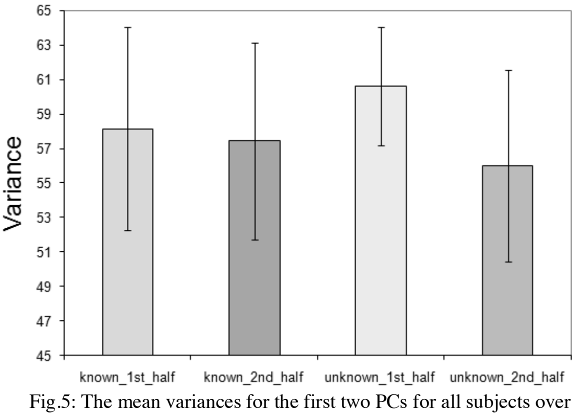


Fig.5: The mean variances for the first two PCs for all subjects over all objects in the interaction phase.

length in the interaction phase). The small differences in the eye tracking parameters may be due to the fact that there is a high variance in objects size and shape, which may equalize existing differences between grouped objects in the SDA-M. For example, when comparing the overall saccadic length in horizontal and vertical directions for the ink stone and the hole puncher, the data shows smaller values for the saccade length for the unknown object. Additionally, subjects use visual data to get more information from the object when grasping it. In the interaction phase, the distributed attention shows that subjects explore unknown objects in more detail in order to figure out how to manipulate them. Interaction with the object in complex real world environments is clearly shaping the cognitive representation of the single objects closer to functional similarity – resulting in a different grasping behavior. We will repeat the experiment with younger and older people, in order to investigate the variations in humans' perceptual and grasping behavior over the developmental process.

Our results support the notion that the grasping movements are strongly influenced by conceptual factors, as the separation of small objects from larger ones does not only show up in the mental representation structures, but also in the PC space. The posture analysis in PC space was very difficult because of the variety of object shapes and sizes and the unconstrained way the subjects could act with the objects. Therefore, it was not possible to find meaningful clusters of grasping postures for known and unknown objects. We build on the theory that the development of cognition is linked closely with the capability of acting on one's environment and causing changes to it [2]. These insights into the human learning and perceptual processes when encountering unknown situations are of importance for the field of cognitive neuroscience robotics in order to build artificial cognitive systems that can interact with a human in an intuitive way. By measuring the cognitive structure of motor action/object relations and attentional processes, robotic systems can get insights into the level of expertise and visual interests of interaction partners. Such cognitive structures can be implemented using self-organizing maps, allowing a seamless integration into neural network based architectures for robot control [23]. The future goal is to combine learning of high-level cognitive memory structures with low level, automatic learning of grasping movements using methods like goal babbling [24]. In term of practical applications, cognitive neuroscience robots could possibly select and adjust their actions flexibly in a given situation, e.g., while assisting an older or handicapped person.

ACKNOWLEDGMENT

We wish to thank Patrick Artner for his help in the lab and all of our subjects for their patience.

REFERENCES

[1] J. Maycock, B. Bläsing, T. Bockemühl, H. Ritter, and T. Schack, "Motor synergies and object representation in virtual and real grasping," in *Conf. Rec. 1st Int. Conf. on Applied Bionics and Biomechanics (ICABB)*, Venice, Italy, 2010.

[2] T. Schack and H. Ritter, " The cognitive nature of action – functional links between cognitive psychology, movement science, and robotics," *Progress in Brain Research*, vol. 174, pp. 231-251, 2009.

[3] A. Saxena, J. Driemeyer, J. Kearns, C. Osondu, and A.J. Ng, "Learning to grasp novel objects using vision," *Int. Journal of Robotics Research*, vol. 27, pp. 157-173, 2008.

[4] S. El-Khoury, A. Shababani, and V. Perdereau, "Learning the natural grasping component of an unknown object," *in Conf. Rec. IEEE/RSJ Intern. Conf. on Intell. Robots and Systems (IROS), 2007*, pp. 2957-2962.

[5] M. Richtsfeld and M. Vincze, "Grasping of Unknown Objects from a Table Top," in *Workshop Vision in Action: Efficient strategies for cognitive agents in complex environments*, Marseille, France, 2008.

[6] C. Borst, M. Fischer, and G. Hirzinger, "A fast and robust grasp planner for arbitrary 3D objects", *in Proc. IEEE International Conference on Robotics and Automation*, 1999, pp. 1890-1896.

[7] C. Michel, C. Rimond, V. Perdereau, and M. Drouin„ "A robotic grasping planner based on the natural grasping axis," in *Proc. of the Intern. Conf. on Intell. Manipulation and Grasping*, 2004.

[8] G. Recatalá, E. Chinellato, Á.P. Del Pobil, Y. Mezouar, and P. Martinet, "Biologically-inspired 3D grasp synthesis based on visual exploration," *Journal Autonomous Robots*, vol. 25, pp. 59-70, 2008.

[9] J. Steffen, C. Elbrechter, R. Haschke, and H. Ritter, " Bio-inspired motion strategies for bimanual manipulation task, " *Humanoids*, 2010, pp. 625-630.

[10] S. Ekvall and D. Kragic, "Interactive grasp learning based on human demonstration," in *IEEE/RSJ International Conference on Robotics and Automation*, New Orleans, USA, 2004, pp. 3519-3524.

[11] T. Schack, "The relationship between motor representation and biomechanical parameters in complex movements. Towards an integrative perspective of movement science," *Eur J Sport Sci.*, vol 3, pp. 1-13, 2003.

[12] J. Maycock, K. Essig, R. Haschke, T. Schack, and H. Ritter, "Towards an Understanding of Grasping using a Multi-sensing Approach," ICRA workshop on Autonomous Grasping, 2011.

[13] T. Schack, *A method for measuring mental representation. Handbook of Measurements in Sports*. New York: Human Kinetics, 2011.

[14] T. Schack, "The cognitive architecture of complex movements," *Int. Jour. of Sport and Exercise Psychology*, vol. 2, pp. 403-438, 2004.

[15] B. Rasolzadeh, M. Bjorkman, K. Huebner, and D. Kragic, "An active vision system for detecting, fixating and manipulating objects in real world," *Int. Jour. of Robotics Research*, vol. 29, pp. 133–154, 2010.

[16] D. Jonikaitis and H. Deubel, "Independent allocation of attention to eye and hand targets in coordinated eye-hand movements," Psychological Science, vol. 22, pp. 339-347, 2011.

[17] R. S. Johansson, G. Westling, A. Backström, and J. R. Flanagan,"Eye-hand coordination in object manipulation," *The Journal of Neuroscience*, vol. 21, pp. 6917–6932, 2001.

[18] M. A. Just and P. A. Carpenter, *The Psychology of Reading and Language*. Newton: Allyn and Bacon, 1987.

[19] J. Maycock, D. Dornbusch, C. Elbrechter, R. Haschke, T. Schack, and H. Ritter, "Approaching manual intelligence," *in KI - Künstliche Intelligenz, Issue Cognition for Technical Systems*, pp.1-8, 2010.

[20] Sensomotoric Instruments (SMI) [Online]: http://www.smivision.com.

[21] M. Santello, M. Flanders, and J.F. Soechting, "Postural hand synergies for tool use," *J. Neurosci.*, vol. 18, pp. 10105-10115, 1998.

[22] D.S. Wooding., "Eye movements of large populations: II. deriving regions of interest, coverage, and similarity using fixation maps," *Behavior Research Methods, Instruments & Computers*, vol 3, pp. 518-528, 2002.

[23] M. Rolf, J. J. Steil, and M. Gienger, "Learning Flexible Full Body Kinematics for Humanoid Tool Use", Int. Symp. Learning and Adaptive Behavior in Robotic Systems , pp. 171 - 176 , 2010.

[24] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse Kinematics", IEEE Trans. Autonomous Mental Development, vol. 2, pp. 216 - 229, 2010.

PM2-3

The Future of Cognitive Neuroscience Robotics

Minoru Asada

Dept. of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Japan

The cognitive neuroscience robotics has been taking interdisciplinary approaches to understanding of human cognitive processes aiming at designing future products in our daily life. A wide range of research issues have been attacked from very fundamental, scientific issues such self/other cognition related to MNS to more engineering ones such as human-machine interface to control robots inside and/or outside the buildings. These issues are not separated nor independent, but closely related to each other. Future issues may include deeper understanding of human cognitive functions, new applications of BMI studies, and more friendly and natural man-machine interfaces including humanoids that can assist people not only physically but also mentally in our daily life. In my talk, we discuss these issues.