

Anticipation and Initiative in Human-Humanoid Interaction

Peter Ford Dominey¹, Giorgio Metta², Francesco Nori², Lorenzo Natale²

¹CNRS & INSERM U846, France, peter.dominey@inserm.fr

²Italian Institute of Technology, Italy, Giorgio.metta@iit.it; francesco.nori@iit.it; lorenzo.natale@iit.it

Abstract— One of the long-term goals for humanoid robotics is to have these robots working side-by side with humans, helping the humans in a variety of open ended tasks, which can change in real-time. In such contexts a crucial component of the robot behavior will be to adapt as rapidly as possible to regularities that can be learned from the human. This will allow the robot to anticipate predictable events, in order to render the interaction more fluid. This will be particularly pertinent in the context of tasks that will be repeated several times, or that contain sub-tasks that will be repeated within the global task. Through exposure to repetition the robot should automatically extract and exploit the underlying regularities. Here we present results from human-robot cooperation experiments in the context of a cooperative assembly task. The architecture is characterized by the maintenance and use of an “interaction history” – a literal record of all past interactions that have taken place. During on-line interaction, the system continuously searches the interaction history for sequences whose onset matches the actions that are currently being invoked. Recognition of such matches allows the robot to take different levels of anticipatory activity. As predicted sequences are successively validated by the user, the level of anticipation and learning increases. Level 1 anticipation allows the system to predict what the user will say, and thus eliminate the need for verification when the prediction holds. At Level 2 allows the system to take initiative to propose the predicted next event. At Level 3, the robot is highly confident and takes initiative to perform the predicted action. We demonstrate how these progressive levels render the cooperative interaction more fluid and more rapid. Implications for further refinement in the quality of human-robot cooperation are discussed.

I. INTRODUCTION

If humanoid robots are to engage with humans in useful, timely and cooperative activities, they must be able to learn from their experience, with humans. Ideally this learning should take place in a way that is natural and comfortable for the user. The results of such learning should be that during the course of an interaction, as the robot continuously acquires knowledge of the structure of the interaction, it can apply that knowledge in order to anticipate the behavior of the user. This anticipation can be expressed both in terms of the actions performed by the robot, as well as by its style of verbal communication.

Anticipation is the hallmark of cognition: von Hofsten for example says that: *Actions are directed to the future and must predict what is going to happen next* [22].

Spoken language has been extensively developed and applicable to human-robot interaction [1-4, 6-11]. Nicolescu and Mataric [12] employed spoken language to allow the user to clarify what the robot learned by demonstration. In order to explore how language can be used more directly, Lauria et al. [13] asked naïve subjects to provide verbal instructions to a robot in a visual navigation task. Their analysis of the resulting speech corpora, yielded a set of verbal action chunks that could map onto robot control primitives. They demonstrated the effectiveness of such instructions translated into these primitive procedures for actual robot navigation [14]. This indicates the importance of implementing the mapping between language and behavioural primitives for natural language instruction or programming [see 15]. Learning by imitation and/or demonstration likewise provide methods for humans to transmit desired behaviour to robots [16-17]. Thus, different methods have been used to allow users to transfer task knowledge to the robot, including traditional keyboard programming and the use of motion tracking for learning by demonstration.

II. OUR APPROACH

We have previously developed cooperation systems that allow a hands-free condition in which the user can actively perform one role in a cooperative task, while instructing the robot at the same time, such that the robot acquired new behaviours via its interaction with the human [11]. The current research thus takes place in the continuity of our studies of human-robot cooperation in the context of a cooperative construction task. The table construction task has repetitive subtasks (attaching the 4 legs) which provide an opportunity for learning and performance improvement within the overall task. In previous research, the user was required to explicitly instruct the robot about when to initiate and terminate behaviour learning, that is, the user was required to keep track of the segmentation of the overall task into subtasks.

The goal, and novelty of the current work, is to extend

the learning and anticipation capabilities of the robot within this interactive context by allowing the robot to automatically analyse ongoing behaviour with respect to its internal representation of its past. This will allow the robot to anticipate what the user will say (thus improving the spoken language interface), and to take a progressively more proactive role in the interaction. Most importantly, this frees the user from requirement to explicitly segment tasks into subtasks for teaching the robot, as this is now performed automatically.

In order to achieve this anticipatory and adaptive capability we will exploit the notion of the interaction history, as the temporally extended personal sensory motor history of the robot in its interaction with the world and the human [5]. By comparing ongoing interactions with previous experience in the interaction history, the system can begin to anticipate. Depending on the level of stability of these encoded interactions, the system can commit to different levels of anticipatory and initiative-taking behaviour. The stability of an interaction will increase as a function of its reuse over time. Here we describe the distinct levels of anticipation and initiative taking that will be implemented and tested. Level 1 anticipation allows the system to predict what the user will say, and thus eliminate the need for verification when the prediction holds. At Level 2 allows the system to take initiative to propose the predicted next event. At Level 3, the robot is highly confident and takes initiative to perform the predicted action.

A. Level 1: Dialog Anticipation

While the speech recognition provided by the RAD system (described below) is quite reliable, we systematically employ a subdialog in which, after the user makes a statement the system asks “Did you say ...?”, in order to correct for recognition errors. Most often, the recognition works correctly, and so this verification step is unnecessary, and most of all, it is tiresome for the user.

When the system has recognised that the current sequences of actions matches with a sequences that has been previously executed, and stored in the Interaction History, then it can anticipate what will be said next by the user. If this item matches with what is actually recognised as the user’s next statement, then the system can dispense with the need to explicitly validate. This can significantly improve the smoothness of the flow of interaction.

B. Level 2: Action Proposition

Once a sequence has been validated at level 1 (i.e. the system correctly predicts what the user will say), then that sequence is elevated to level 2. At this level, again, when the system detects that a level 2 sequence is being executed,

it will take initiative and propose to the user the next element in the predicted sequence. The user can then accept or decline the offer. In the context of a repetitive task, this is actually quite helpful as the user can rely on the record of her own previous history with the robot in order to guide ongoing action.

C. Level 3: Action Initiation

Once the sequence has been validated at Level 2, (i.e. the user has accepted the succession of actions proposed by the system), then it attains Level 3. At this level, when the sequence is detected, the robot takes full imitative and begins to execute the subsequent actions in the Level 3 sequence. At this level, the user is truly aided by the “apprentice” who has successively gained confidence, and can now proceed with its part of the interaction without the need to confer by language.

III. SYSTEM DESCRIPTION

The system is modular, with sensory motor control of the robot being accessed by higher level control (spoken language, anticipatory learning mechanism) via a well defined interface as specified below. The general approach we take is that the robot should come pre-equipped with a set of capabilities, including grasping particular objects, moving to useful postures that allow it to pass objects to the user, take objects from the user, hold objects while the user works on them etc. A relatively restricted set of such capabilities are predefined. They can be considered to correspond to words in a lexicon, and the system allows the user to compose them into useful behaviors, like language allowing the composition of words into sentences. We have demonstrated the practicality of this approach [11].

A. The iCub Humanoid

The current research is performed with the iCub, a humanoid robot developed as part of the RobotCub project [21]. It has been designed to reproduce the size of a three and a half years old child (approximately 1m tall). Its kinematic structure has a total of 53 degrees of freedom primarily located in the upper torso. The robot hands are extremely dexterous and allow manipulation of objects thanks to their 18 degrees of freedom in total. The iCub is strong enough to crawl on all fours and sit to free the hands for manipulating objects.

The iCub actuation relies on electric motors. The major joints are actuated by brushless DC motors coupled with frameless Harmonic Drive gears. This guarantees torques up to 40Nm at the shoulders, spine and hips. The head and hands are actuated by smaller brushed-DC motors.

The robot head is equipped with cameras, microphones, gyroscopes & linear accelerometers. Moreover the entire

body is equipped with force/torque sensors, position and temperature sensors. A fully sensorized skin and fingertips is under development.

The electronics of the iCub has been developed specifically to fit the limited space available. Each controller card runs a local position or velocity control loop on a special purpose DSP at 1KHz. Several cards are connected to a main relay CPU via a set of four CAN bus lines. These lines end into a multi-purpose I/O card which communicates to the relay CPU (a Pentium) which is also located inside the robot. More demanding computation can happen outside the robot. In a typical configuration sensory processing (e.g. vision) is performed on a cluster of PCs connected via Gbit Ethernet to the iCub (see Figure 1).

Additional electronics has been designed to sample and digitize the iCub sensors. Also in this case, everything converges on the main relay CPU by means of various additional connections (e.g. serial, firewire, etc.).

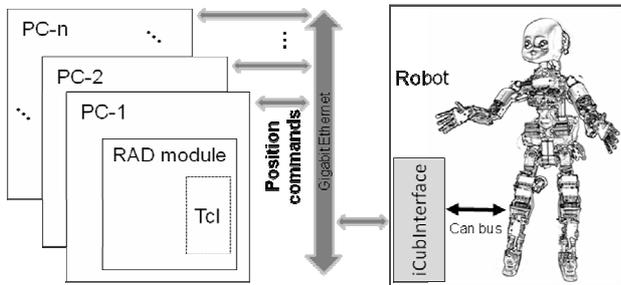


Figure 1. Hardware and software architecture for the presented experiment.

B. Yarp Architecture

The iCub software architecture uses YARP, an open source library written to support software development and integration in robotics [20]. The core of YARP is an inter-process communication layer which allows processes on different machines to exchange data across an Ethernet network (Figure 1). Communication in YARP is transport independent; details about the underlying network and protocol are hidden to the user. Similarly, YARP offers device driver wrappers, which help separating user-level code from vendor-dependent code related to sensors and actuators. Overall this contributes to achieve loose coupling between algorithms and hardware, and, in turn, favors modularity. In short, communication in YARP takes place through connections, called ports. Ports are named entities which move data from one process to another (or several others). iCub capabilities are implemented as a set of modules, interconnected through YARP ports. Each module is an executable which implements a given functionality, and creates a set of ports to receive and send data. Some modules provide access to the hardware. For example the iCubInterface module exports a set of ports to give access to the motors and broadcast the encoder feedback from all joints. Other modules in the architecture

control the robot by sending messages to these ports. Commands can be specified as joint space position or velocity.

C. Spoken Language Control of the Robot

Dialog management and spoken language processing (voice recognition, and synthesis) is provided by the CSLU Rapid Application Development (RAD) Toolkit (<http://cslu.cse.ogi.edu/toolkit/>). RAD provides a state-based dialog system capability, in which the passage from one state to another occurs as a function of recognition of spoken words or phrases; or evaluation of Boolean expressions.

Via the TCL language we can open YARP ports to the joints of the iCub. The robot is thus controlled via spoken language, via interaction with the different joints through the YARP port interface.

The behavioral result of a spoken action command that is issued either directly, or as part of a learned plan is the execution of the corresponding action on the robot. Based on the preliminary analysis of the table-building scenario described above, a set of primitive actions was identified for the iCub. Each of these actions, specified in Table 1, corresponds to a particular posture or posture sequence that is specified as the angles for a subset of the 53 DOFs. These actions have been implemented as vectors of joint angles for controlling the head, torso, and left and right arms.

Table 1. iCub Specific Action Commands

Motor Command	Resulting Actions
Reach	Position left hand next to closest table leg
Grasp	Close left hand
Lift	Raise left hand
Pass	Turn trunk and left shoulder towards user
Open	Open left hand
Hold	Bimanually coordinated holding
Release	Place both hands in upward safe position
Wait	Suspend until OK signal

In addition to the allowing the user to issue iCub specific motion commands, the Spoken Language Programming system implements the different levels of anticipation described in Section II. The RAD programming environment provides a GUI with functional modules that implement speech recognition and synthesis, and flow of control based on recognition results and related logic. Part of the system we developed for the current research is illustrated in Figure 2. Here we describe the control flow that implements the multilevel anticipation capability.

Nodes in the graph are referred to in italics.

The invariant situation is that the last action commanded has been added to the ongoing Interaction History, and the previous two actions are continuously compared in a sliding window with the Interaction History at *find_match*. If a match is found, the next action in the Interaction History is a candidate for anticipation. At *anticp_status*, if the sequence is recognized for the first time, then at *anticpate_next_command* the next command is identified as a target for speech recognition, and the sequence anticipation level is incremented for the current sequence element. If the sequence is recognized for the second time, at *propose_initiative*, the system proposes this action to the user. If it has been successfully recognized and validated more than twice, the next element is taken for execution at *take_initiative*.

At the main node, *Select*, the user chooses actions from Table 1. *check_anticipation*: Once the action is selected, the system determines what level of anticipation can be applied. If the current action is not part of a sequence of at least two elements recognized in the Interaction History, then there is no anticipation, and the system asks the user to confirm her command. If the command is part of a sequence that has been recognized for the first time – then the system will skip the verbal confirmation if command matches prediction. As already stated, at the 2nd recognition – propose to anticipate, and for 3rd + recognition: take initiative directly. In our previous work, only a single procedure sequence could be learned, or a limited number of behaviors to be learned.

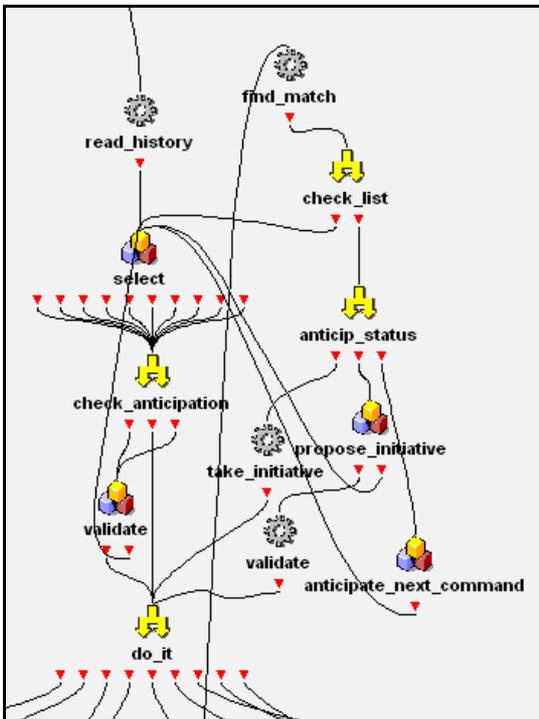


Figure 2. Control Flow in the Spoken Language Interaction.



Figure 3. Progress in the Table Assembly Task.

IV. EXPERIMENTAL RESULTS

In order to evaluate the implemented system, we performed an experiment that involved human-robot cooperation via SLP. In this experiment, the human user and the robot cooperate to construct a small table, as illustrated in Figure 3.

A. Assembling the table

The table assembly task is interesting as it involves cooperation (the robot must pass elements to the user, and hold things while the user works), and it has a repetitive structure that allows learning. For each leg, the user will ask the robot to reach to and grasp the leg, lift it, pass it to the user and open the hand, and then finally hold the table while the user attaches the leg. In the following interaction, note how the corresponding interaction becomes streamlined and fluid as the robot acquires knowledge of the task.

Table 2 provides the literal transcription of the human-robot dialog, Robot language is indicated as underlined, and human language in plain text.

B. Performance Analysis of SLP Effects

Table 2 provides evidence that as the levels of anticipation increase, so does the efficiency of the interaction. In order to quantify the effects of this form of anticipatory learning, we measured command processing time as the duration from the spoken onset of the user's command to the robot's completion of that action. We then grouped these actions based on the four assembly phases corresponding to the four legs of the table.

Because of the repetitive structure of the task, the four legs correspond to four levels of anticipation – starting at a naïve state with the first leg at Level 0. Figure 4 presents the mean values and their ranges. We see a reduction in processing time over the course of the task. A repeated measures ANOVA confirmed a significant effect of Level of anticipation on completion time, $F(3,15)=4.65$; $p<.0172$.

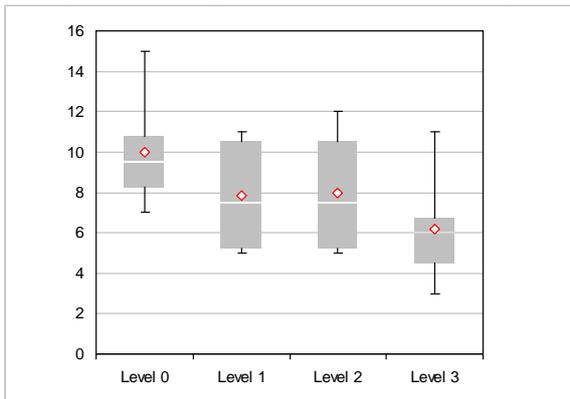


Figure 4. Action execution times (seconds) for robot actions in the assembly of the table, based on level of anticipation. These four levels correspond to the repetitive structure of the task for the 4 legs, respectively. Whiskers indicate max and min times, Boxes 25th and 75th percentile, point – mean, white line – median.

Human	Robot response	Human	Robot action
Reach.	<u>Did you say reach?</u>	Yes	<u>Reaching</u>
Grasp.	<u>Did you say grasp?</u>	Yes.	<u>Grasping.</u>
Lift.	<u>Did you say lift?</u>	Yes.	<u>Lifting</u>
Pass.	<u>Did you say pass?</u>	Yes.	<u>Passing</u>
Open.	<u>Did you say open?</u>	Yes.	<u>Opening hand.</u>
Hold the table.	<u>Did you say hold?</u>	Yes.	<u>Holding</u>
Wait.	<u>Did you say wait?</u>	Yes.	<u>Waiting for your signal; OK.</u>
Release.	<u>Did you say release?</u>	Yes.	<u>Releasing</u>
Wait.	<u>Did you say wait?</u>	Yes.	<u>Waiting for your signal; OK.</u>
Reach.	<u>Did you say reach?</u>	Yes	<u>Reaching</u>
Grasp.	<u>Did you say grasp?</u>	Yes.	<u>Grasping.</u>
Lift.			<u>Lifting.</u>
Pass.			<u>Passing</u>
Open.			<u>Opening hand</u>
Hold the table.			<u>Holding</u>
Wait.			<u>Waiting for your signal. OK</u>
Release.			<u>Releasing.</u>
Reach.	<u>Did you say reach?</u>	Yes	<u>Reaching.</u>
Grasp.	<u>Did you say grasp</u>	Yes	<u>Grasping.</u>
	<u>Shall I do lift?</u>	yes	<u>Lifting.</u>
	Shall I do pass?	Yes	Passing
	Shall I do open?	yes	opening hand
	Shall I do hold?	Yes	Holding
	Shall I do wait?	Yes	<u>Waiting for your signal. OK</u>
	Shall I do release?	Yes	Releasing
	Shall I do wait?	No	
Reach.	<u>Did you say reach?</u>	Yes	<u>Reaching</u>
Grasp.			<u>Grasping.</u>
	I know what to do.		<u>Lifting.</u>
	I know what to do.		Passing,
	I know what to do.		Opening hand,
	I know what to do.		Holding
	I know what to do.		Waiting for your signal OK
	I know what to do.		Releasing

Table 2. Transcription of the dialog from a experiment in the Table building scenario. Each line represents one interaction transaction. Note the three levels of anticipation in the Robot Response: (1) Elimination of the speech validation for correct predictions, (2) action proposition, (3) action initiative.

V. DISCUSSION

Progress in human-robot cooperation is being made in part via well-documented methods for action learning that include demonstration and imitation [16-17]. Language has been used in this context for correcting and clarifying what is being learned by demonstration [13-14]. One of the fundamental requirements is to establish the grounded meaning at the base of the communication, that is the link between human language, and robot action and perception [1-4]. This has recently been explored and developed in the domain of language based navigation [13-14]. Roy and colleagues further establish these links via an amodal Grounded Situation Model that integrates perception, action and language in a common framework for language based human-robot cooperation. They [15] have also demonstrated how visual perception can be used to generate an internal model of the physical scene that can then be used to allow the robot to act in the scene and describe the scene based on this internal representation. In this context, verbs correspond to sensory-motor networks or perceptually guided actions that the robot carries out (*lift, pickup, touch*). The internal model allows the robot to appropriately understand and use spatial relation words (e.g. left of, above) based on the human's perspective or its own. In that implementation, the system had a fixed repertoire of actions, and it did not allow any on-line modification of the behavioral repertoire of the robot, which is part of the goal of the work we describe here.

We have made progress with a system that can learn grammatical constructions which make the mapping between predicate-argument representation of action as perceived by a robot vision system, and natural language sentences that describe that action, generalizing to new action scenes [6-11]. Iwahashi and colleagues [1] have similarly demonstrated a robust capability to use <sentence, meaning> pair corpora in order to extract the grammatical structure of sentences that describe visual scenes. Once such mappings between sentences and meaning have been established they can be used in dialog contexts for human-robot interaction. Asoh et al [3] have implemented Jijo-2, a mobile robot that performs office tasks (notification of presence or absence of individuals, guiding within the office, updating schedules etc.) based on spoken language interaction. In related work on the robot assistant BIRON, particular attention has been paid to robustness in the speech understanding component by exploiting the appropriate situated semantic contexts [4].

In the context of the relation between meaning and language, we can consider that single events can be described at the level of single sentences. In contrast, an interlaced sequence of human and robot actions in a cooperation context is more appropriately described at the level of a dialog. Depending on the nature of the task, the structure of the dialog will vary. Lemon and colleagues [2] have developed a development environment that allows non-expert developers to produce complete spoken

dialogue systems based only on a standardized description of the process flow of control of their activity, such as verify customer identity, or pay a bill. It will be interesting to apply this technique in the development of more extended tasks for human-robot interaction.

The current research focused on the use of the interaction history for allowing the system to automatically identify useful behaviours. The user no longer needs to manage this learning explicitly, but instead can rely on the robot to extract pertinent regularities from their interactions. This line of research should be of significant impact in the future progress in human-robot cooperation.

ACKNOWLEDGMENT

This work is supported by the FP7 IST – PROJECT No.215805 – CHRIS, and by the French ANR-07-ROBO-0004-04 Amorces.

REFERENCES

- [1] Iwahashi N (2003) Language acquisition through a human-robot interface by combining speech, visual, and behavioral information, *Information Sciences*, 156 (1-2) 109-121
- [2] Lemon O, Liu X (2006) DUDE: a Dialogue and Understanding Development Environment, mapping Business Process Models to Information State Update dialog system, *Proceedings of EACL, 2006*
- [3] Asoh H, Motomura Y, Asano F, Hara I, Hayamizu S, Itou K, Kurita T, Matsui T, Vlassis, Bunschoten R, Kröse B (2001) Jijo-2: An Office Robot That Communicates and Learns, *IEEE Intelligent Systems*, Sept/Oct 2001, 46-55.
- [4] Hüwel S, Wrede B (2006) Spontaneous Speech Understanding for Robust Multi-Modal Human-Robot Communication *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 391–398.
- [5] Mirza NA, Nehaniv CL, Dautenhahn K, td Boekhorst R (2007) Grounded Sensorimotor Interaction Histories in an Information Theoretic Metric Space for Robot Ontogeny, *Adaptive Behavior* (15) 167-187
- [6] Dominey PF (2007) Spoken Language and Vision for Adaptive Human-Robot Cooperation, in (Matthias Hackel, Ed). *Humanoid Robotics*, ARS International, Vienna.
- [7] Dominey PF, Boucher (2005) Learning To Talk About Events From Narrated Video in the Construction Grammar Framework, *Artificial Intelligence*, 167 (2005) 31–61
- [8] Dominey, P.F., 2003. Learning grammatical constructions from narrated video events for human-robot interaction. *Proceedings IEEE Humanoid Robotics Conference*, Karlsruhe, Germany
- [9] Dominey, P. F., Boucher, J. D., & Inui, T. (2004). Building an adaptive spoken language interface for perceptually grounded human-robot interaction. In *Proceedings of the IEEE-RAS/RSJ international conference on humanoid robots*.
- [10] Dominey PF, Alvarez M, Gao B, Jeambrun M, Weitzenfeld A, Medrano A (2005) Robot Command, Interrogation and Teaching via Social Interaction, *Proc. IEEE Conf. On Humanoid Robotics 2005*
- [11] Dominey PF, Mallet A, Yoshida E (2007) Progress in Programming the HRP-2 Humanoid Using spoken Language, *Proceedings of ICRA 2007, Rome*.
- [12] Nicolescu M.N., Mataric M.J. : Learning and Interacting in Human-Robot Domains, *IEEE Trans. Sys. Man Cybernetics B*, 31(5) 419-430.

- [13] Lauria S, Buggmann G, Kyriacou T, Klein E (2002) Mobile robot programming using natural language. *Robotics and Autonomous Systems* 38(3-4) 171-181
- [14] Kyriacou T, Bugmann G, Lauria S (2005) Vision-based urban navigation procedures for verbally instructed robots. *Robotics and Autonomous Systems*, (51) 69-80
- [15] Mavridis N, Roy D (2006). Grounded Situation Models for Robots: Where Words and Percepts Meet. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*
- [16] Zöllner R., Asfour T., Dillman R.: Programming by Demonstration: Dual-Arm Manipulation Tasks for Humanoid Robots. *Proc IEEE/RSJ Intern. Conf on Intelligent Robots and systems (IROS 2004)*.
- [17] Calinon S, Guenter F, Billard A (2006) On Learning the Statistical Representation of a Task and Generalizing it to Various Contexts. *Proc IEEE/ICRA 2006*.
- [18] K.Kaneko, F.Kanehiro, S.Kajita, H.Hirukawa, T.Kawasaki, M.Hirata, K.Akachi, and T.Isozumi, "Humanoid robot hrp-2," in *Proceedings of the 2004 IEEE International Conference on Robotics & Automation*, vol. 2, 2004, pp. 1083–1090.
- [19] F. Kanehiro, N. Miyata, S. Kajita, K. Fujiwara, H. Hirukawa, Y. Nakamura, K. Yamane, I. Kohara, Y. Kawamura, and Y. Sankai, "Virtual Humanoid Robot Platform to Develop Controllers of Real Humanoid Robots without Porting," *Proc. Int. Conference on Intelligent Robots and Systems*, pp. 1093-1099, 2001.
- [20] P. Fitzpatrick, G. Metta, and L. Natale, "Towards Long-Lived Robot Genes", In *Journal of Robotics and Autonomous Systems Special Issue on Humanoid Technologies*. Vol. 56 (2008) 1–3.
- [21] N.G. Tsagarakis, G.Metta, G.Sandini, D.Vernon, R.Beira, F.Becchi, L.Righetti, J.S.Victor, A.J. Ijspeert, M.C.Carrozza and D.G.Caldwell. iCub – The Design and Realization of an Open Humanoid Platform for Cognitive and Neuroscience Research. In *Advanced Robotics special issue on "Robotic platforms for Research in Neuroscience"*. Vol 21, No. 10, June 2007.
- [22] von Hofsten, C. (2004). An action perspective on motor development. *Trends in cognitive sciences*, 8(6), 266-272