

# **Towards Cognitive Robotics**

**Christian Goerick**

**2008**

**Preprint:**

This is an accepted article published in *Creating Brain-like Intelligence*. The final authenticated version is available online at: [https://doi.org/\[DOI not available\]](https://doi.org/[DOI not available])

# Towards Cognitive Robotics

Christian Goerick

Honda Research Institute Europe GmbH  
Carl-Legien-Strasse 30  
63073 Offenbach / Main  
Germany  
`christian.goerick@honda-ri.de`

**Abstract.** In this paper we review our research aiming at creating a cognitive humanoid. We describe our understanding of the core elements of a processing architecture for such kind of an artifact. After these conceptual considerations we present our research results on the form of the series of elements and systems that have been researched and created.

## 1 Introduction

Research about intelligent systems interacting in the real world is gaining momentum due to the recent advances in computing technology and the availability of research platforms like humanoid robots. Some of the most important research issues are architectural concepts for the overall behavior organization of the artifacts. The spectrum spans from mechanisms for action selection in a direct fashion [1] towards research with the target of creating cognitive architectures [2]. One long-term goal of the research presented in this paper is aiming at incrementally creating an autonomously behaving system that learns and develops in interaction with a human user as well as based on internal needs and motivations. The other long-term goal is to understand how the human brain works, the only truly intelligent system as of today. Both goals are coupled in a way that is called analysis by synthesis. We would like to create brain-like intelligent systems, hence we have to understand how the brain works. The artifacts we are creating should show what we have understood from the brain so far, and should help formulating the next questions aiming at further understanding the brain.

The vehicle for the research considered here are humanoid robots. Their anatomy and embodiment is considered a necessary condition in order to create intelligence in an anthropocentric environment. In this paper we report on our current research efforts towards cognitive robotics: the researched elements and the endeavors aiming at a brain-like control architecture for humanoid robots.

## 2 Towards an Architecture

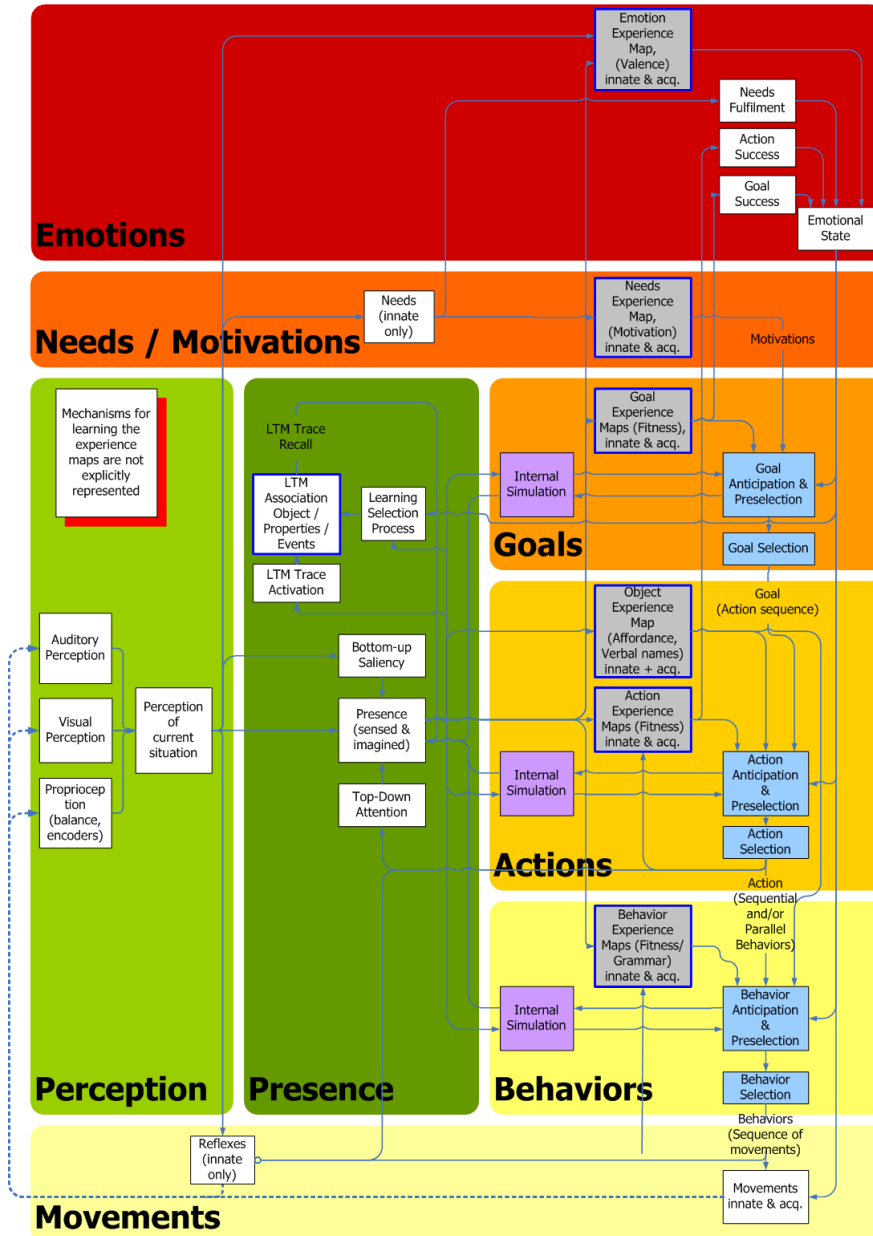
As stated above, the long term goal of our research is establishing a cognitive architecture for controlling humanoid robots. We are convinced that cognitive

or intelligent performances of artifacts can only be achieved within an architecture orchestrating the individual elements in a phenomenologically coherent way. What those elements exactly are and how they are to be arranged is subject to current research. Nevertheless, it is understood that there is a minimal subset of those elements that have to be addressed. The analysis of biological systems teaches us about those principled elements and their possible role within animals' brains [3]. We consider the following elements to belong to this subset:

- Sensory perception: Comprising extero- (vision, audition, tactile) and proprioception (measured internal states like posture).
- Presence or working memory: A short term representation of behaviorally meaningful percepts or internally generated entities as the basis for external and internal actions. The content of this presence is modulated by top down attention processes.
- Plastic (i.e. learnable) long term memory for the storage and retrieval of consolidated persistent entities like object, words, scenes and mental concepts.
- Elements for predictions and internal simulation for creating expectations about the world for selecting the relevant information from the sensory streams and choosing internally the most effective action from a set of possible alternatives without testing all alternatives externally.
- Basic motor control and coordination means for efficiently controlling systems with a large number of degrees of freedom.
- A basic behavior repertoire building on the motor control level for a more abstract and robust representation of actions.
- A more abstract behavior organization comprising the traditionally separated issues of communication and action, leading to behaviorally routed definitions of the semantics of language.
- A representation of goals and processes working on those representations for organizing a meaningful system behavior on mesoscopic time scales above purely reactive sensory actions and below strategically driven decision processes. Those elements will also influence the perception in a top down fashion for focusing on goal relevant entities in the external world.
- A set of internal drives and motivations for establishing internal forces towards a continuous self-development of the system and controlling the balance between explorative and exploitative actions.
- A value system (“emotions”) providing basic guidelines for the limits of autonomous behavior by unconditioned preferences concerning elements and states of the environment.

This subset is far from being complete, but we consider those elements as the most pressing research issues that will provide major progress in the field of cognitive robotics.

Figure 1 shows a sketch relating the above mentioned functional entities to each other. Those functional elements can be again related to the corresponding areas of the brain. The depicted resulting architecture is called PISA, the **P**actical **I**ntelligence **S**ystems **A**rchitecture. It represents our current best understanding of an abstract long term research goal. It is to be understood more



**Fig. 1.** PISA: The Practical Intelligence Systems Architecture, showing the major functional elements and relating them to each other. It represents joint work with Frank Joublin and Herbert Janßen dating back to 2004. Please refer to the text for a more detailed description.

in the fashion of a strategic means rather than a concrete goal that is to be realized according to the master plan derived from the drawing. It is very valuable for the strategic organization of research and incremental systems architectures, because all research activities can easily be put into relation and the crucial communication between researchers evolves naturally.

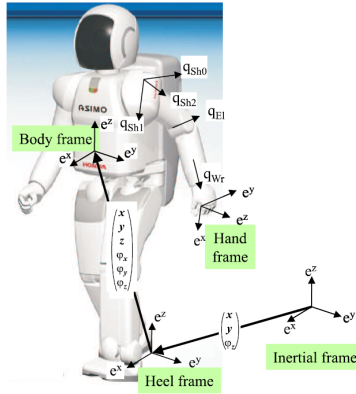
A major issue for most of the elements stated above is learning and adaptation, always in interaction with the real world. We consider this as an important issue, because otherwise we may not ask the research questions stated above in the right way. Nevertheless, we take the freedom to ignore technological issues like on-board computing resources for now, as long as we consider the property of scaling of the researched methods from the beginning.

In the following we will focus on selected issues and created systems.

### 3 Task and Body Oriented Motion Control

The first focus is on the movement generation. In PISA it is located in the bottom-most section called “movements”. Before we can consider researching how to learn movements, actions and behaviors, we have to understand how to do the basic control of the body similar to the spinal cord and the brain stem in biology. Technically, controlling the physical motions of a biped humanoid robot is not a trivial task. The number of degrees of freedom that have to be coordinated is high, the balance of the system has to be maintained under all circumstances and it is currently possible for such kind of robots to mechanically destroy themselves by commanding position of the limbs that lead to self collisions. On the other hand, humans and animals effortlessly control their end-effectors for solving tasks without continuously reflecting on the level of joints about their current motions. Additionally, they have some kind of body image representing the anatomy and boundaries of the own body that helps acting in complex tasks without continuously having contacts with the own body or external objects. From the constructive point of view, it seems desirable in a cognitive architecture to be able to cognitively control only the task relevant parameters and leave the “tedious” details to underlying levels of control. This should include the avoidance of self collisions, which is in technical systems much more disastrous than in natural ones.

As a results of current research we have established a stable layer for the motion control with a so called motion interface that fulfills the requirements stated above and allows cognitive control processes to perform complex control tasks with the humanoid ASIMO with a minimum effort [4]. In contrast to classical joint level control, the robot is controlled by a task level description. The tasks can e.g. be defined by four separate targets for the two hands, the head gaze direction and a position of the centre of mass projected on the ground, respectively. The corresponding coordinate systems are depicted in Figure 2. The coupling between the tasks and the mapping to the actually controlled joints is performed by a whole body motion controller. This controller implements a redundant control scheme considering all degrees of freedom of the robot simul-



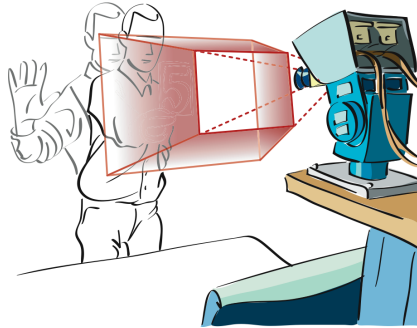
**Fig. 2.** Kinematics model of ASIMO used for the whole body motion control.

taneously. The commanded tasks usually do not determine all degrees of freedom uniquely. For the remaining degrees of freedom it is possible to state potential functions that model the task unspecific preferences of those. This could be closeness to rest positions and the avoidance of extreme joint positions close to the physical limits.

What do we gain by pursuing such kind of task description and whole body control in a cognitive architecture? First of all we have a description of tasks in a more natural way than in the joint space. For example, the right hand is commanded to a position in 3D space with a certain attitude. The necessary joint trajectories are computed automatically online on-board the robot. Therefore, higher level processes don't have to care about the details of the robot motion. Additionally, since the whole body including walking is employed for reaching the commanded target, the motion range is extended incrementally. Imagine a 3D position for the right hand is commanded that is not reachable by arm motions alone. The whole body motion controller first induces a leaning motion of the upper body in order to reach the target, and if this does not suffice ASIMO starts walking for finally reaching the commanded target. Again, higher level cognitive tasks still command only the 3D target position of the right hand. During those movements the appearance of the robot motion is naturally relaxed, because the redundant degrees of freedom are "softly" adapted to the requirements given by the hard task constraints. This fact can be envisioned by assuming springs between the segments of the robot, the task command would correspond to a force pulling the respective hand and the rest of the robot's body adapting to the influence of this force. The walking is conceptually more advanced but can also be treated within the same framework. Such kind of natural appearance can help solving acceptance problems with robots. Summarizing, the task space description and the whole body control approach gives more freedom to the motion control level and disburdens the higher levels of control. Further extensions in the same spirit are commanding task intervals instead of crisp tasks positions

and including self collision avoidance on the level of motion control [5]. Current research is concerned with strategy selection based on internal simulation, allowing ASIMO to autonomously choose the hand to grasp the commanded target with. The proposed approach performs interactively with visually specified targets [6], which is in contrast to state of the art as described in [7, 8]

## 4 Visually and Behaviorally Oriented Learning



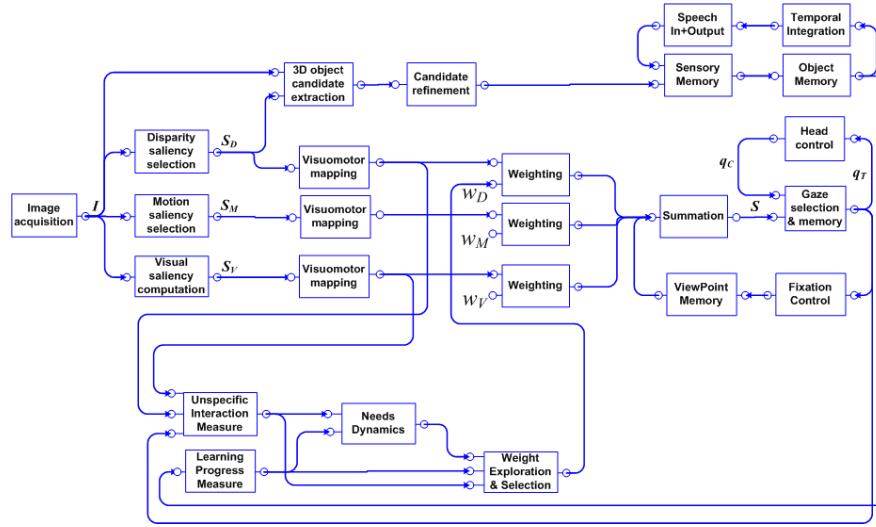
**Fig. 3.** Interaction with the stereo camera head. Waving attracts attention, showing objects within the peripersonal space fixes attention to the presented object. The fixated object can be learned interactively and recognized immediately.

In the introduction we have stated that the long term goal of this work is the creation of a humanoid robot that is equipped with mechanisms for learning and development.

We move the focus away from the movement generation towards vision and the generation and exploration of visually oriented behaviors, including mechanisms for learning and development.

The concrete goal here is to present an interactively behaving vision system for the humanoid that comprises already both kinds of mechanisms: autonomous developmental mechanisms governing the behavior generation and selection, and interactive learning mechanisms allowing for teaching the system new objects to be recognized online. Regarded separately, both mechanisms already represent a valuable step towards autonomous adaptive systems. But the emphasis is more on the principled combination of both. In contrast to statistical learning we are here less concerned about the representation of the variability of the input space but rather in learning behaviorally relevant external and internal entities.

For studying those issues we have created a biologically motivated interactive vision system with adaptive basic behaviors being able to learn and recognize freely presented objects in interaction. The learning is governed by mechanisms building on an internal needs dynamics based on unspecific and specific rewards governing and exploring the parameterization of the basic behaviors [9].



**Fig. 4.** Schematics of the active vision system. See text for detailed description.

On a schematic level, the system can be described as follows. Please refer to Figure 4 for a graphical representation. Based on the images from a stereo camera pair a set of features is computed. Those features comprise the general visual “interestingness” (saliency)  $S_v$  of image locations, the most prominent region in the image based on visual motion  $S_m$ , and the most prominent region in the image based on closeness  $S_d$  to the system. The information is represented as activation maps over the image location, with high values of the maps corresponding to interesting locations. Those maps are weighted with weights  $w_v$ ,  $w_m$ , and  $w_d$  respectively, and added. Based on this combined saliency  $S$  and some memory about previous gaze directions the new gaze direction is determined by means of an integrative peak selection with hysteresis. This simple system exhibits some interesting behaviors. It is a homogeneous control loop that is constantly executed without any structural changes. The behavioral spectrum of the system is spanned by the weighting parameters  $w_v$ ,  $w_m$ , and  $w_d$  of the maps. What kind of behaviors can such kind of a system show? Priming the saliency computation for a certain color like red and having a weight  $w_v$  greater than zero will yield a system that gazes at locations in its environment with red color. If there is no red color in the current view it will randomly look around and shifts its field of view until it finds a spot with red color. For an external observer it looks like the system is “searching for red color”, even though the system was never directly programmed to show a search behavior. Next, imagine having an interacting human in the scene approaching the system and trying to raise the attention of the system (see Figure 3). Assigning a value to  $w_m$  greater than  $w_v$  will cause the system to look at a spot where the waving is located in the image, i.e. with such kind of parameter setting the system will look at regions containing visual



motion rather than general visual saliency. Nevertheless, since the two channels are superimposed, they support each other for stabilizing the gaze selection. If the user continues to approach the system, the distance based map  $S_d$  will have an activation corresponding to the closest part of the user's body to the system. If this contribution receives the highest weight  $w_d$ , the system will continuously focus on, e.g. , the hand of the user. This behavior corresponds to the biological concept of a peripersonal space around the robot. The psychological concept of the peripersonal space is defined as the space wherein individuals manipulate objects, whereas extrapersonal space, which extends beyond the peripersonal space, is defined as the portion of space relevant for locomotion and orienting [10]. Here, the peripersonal space establishes a very important concept for any further meaningful interaction: Sharing the "attention" between the system and the user. An interacting user can show something to the system and the system will focus on the shown entity. Based on this capability we have addressed several new scientific concepts. The first one is the online learning of complex objects freely presented to the system. The object within the peripersonal space is segmented from the image and processed by a biologically motivated visual object recognition and learning system [11]. Two different memory stages and speech interaction serve to continuously learn and label objects in real-time, allowing for online correction of errors during learning. Secondly, we have introduced an internal homeostatic control system representing internal drives allowing for learning the weights of the different maps. This corresponds to learning the visual interaction behaviors instead of working with hard coded weights. The third one is the combination of the previous ones. The system learns to interact with the human user in such a way that its internal needs are in the temporal average equally satisfied including the curiosity for learning new objects. This represents a new quality in interactively learning systems.

The implementation is currently limited to controlling the gaze direction of a head, but the concepts are sufficiently general in order to allow for interactively learning behaviors including manipulators. This is currently being investigated. Research aiming in a similar direction can be found in e.g. [12].

We would like to point out that the system described above is internally not organized in terms and structures of externally visible behaviors. There is no "tracking" or "interaction" module within the system, even if those terms can be attributed to externally observable behaviors. In our opinion it is crucial to make this distinction, because a system organized in terms of externally observable entities will by definition be confined to the predefined set of behaviors put into the system, and no self-driven cognitive self-development will be possible. Or phrased differently, the mechanisms generating certain functions should clearly be distinguished from the semantics that can be associated with them. In the next section this point will be further elaborated.

The research elements presented here are in PISA located in the areas for visual perception, behavior generation and needs.

## 5 ALIS: Autonomous Learning and Interacting System

The concrete system considered in this section is called ALIS, an acronym for “Autonomous Learning and Interacting System”. It is our current design of an incremental hierarchical control system for the humanoid robot ASIMO comprising several sensing and control elements. Those elements are visual saliency computation and gaze selection, auditory source localization for providing information on the most prominent auditory signals, a visual proto-object based fixation and short term memory of the current visual field of view, the online learning of visual appearances of such proto-objects and an interaction-oriented control of the humanoid body. The whole system interacts in real-time with users. It builds on and extends incrementally the research presented in the previous two sections, i.e. the motion generation and control as well as the visual behaviors. The corresponding original publication can be found in [13]. The focus is not on single functional elements of the system but rather on its overall organization and key properties of the architecture. We will describe the architecture by means of a conceptual framework that we developed. The clear focus of this framework is to have a general but not arbitrary means for describing incremental architectures, focusing on the hierarchical organization and on the relations and communication between hierarchically arranged units when they are being created layer by layer. We are convinced that researching more complex intelligent systems without such a kind of framework is infeasible.

To our knowledge, ALIS represents the first system integrated with a full size biped humanoid robot that interacts freely with a human user including walking and non-preprogrammed whole body motions, in addition to learning and recognizing visually defined object appearances and generating corresponding behaviors.

Our architectural concepts point in a similar direction as presented in [14], where a subpart of a mammalian brain has exemplarily been modeled as a hierarchical architecture. We share the view that such kinds of hierarchical organizations are promising for modeling biological brains. We go beyond the arguments presented there by considering explicitly the internal representation and the dependencies in the sensory and behavioral spaces. This is the main difference to classical subsumption-like architectures as summarized in [15]. The approach we pursue is incremental w.r.t. the overall architecture, which goes beyond an incremental local addition of new capabilities within already existing layers. This is the main difference to the state of art in comprehensive humanoid control architectures including learning as presented in [16], [17, 18] and [19]. A similar reasoning applies to the comparison to classical three-layer architectures [20]. The hierarchies we are considering are not fixed to the common categories of deliberation, sequencing and control.

In the next subsection (5.1), we will formulate the framework and discuss the biological motivation. Subsequently, we will present the realized system in more detail. In subsection 5.4 we will report on experiments performed in interaction with the system.

## 5.1 SYSTEMATICA

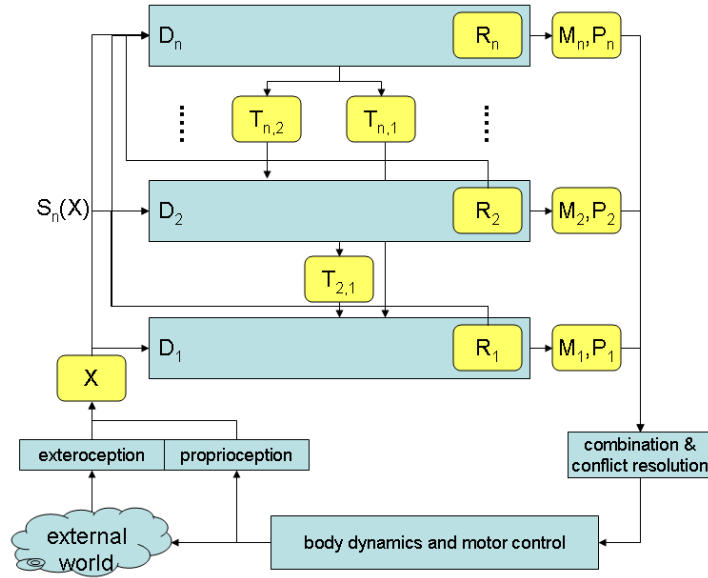


Fig. 5. Schematics of SYSTEMATICA.

We call the framework “SYSTEMATICA”. It was devised for describing incremental hierarchical control architectures in a homogeneous and abstract way. Here, we will introduce the notation that is necessary for making the points of the concrete system instance presented in this contribution. One future target of our research are comparative studies of different kinds of hierarchical control architectures by means of the presented framework.

Each identifiable processing unit or loop  $n$  is described by the following features (see Figure 5 for reference):

- it may process independently from all the other units;
- it has an internal process or dynamics  $D_n$ ;
- its full input space  $X$  is spanned by exteroception and proprioception;
- it can create some system-wide publicly accessible representations  $R_n$  used by itself and other units within the system. The indices may be extended in order to denote the units that are reading from the representation, e.g.  $R_{n;m,o,\dots}$  means that representation  $R_n$  is read by units  $m$  and  $o$ ;
- it may use a subspace  $S_n(X)$  of the complete input space  $X$  as well as the representations  $R_1, \dots, R_{n-1}$ ;

- it can be modulated by top-down information  $T_{m,n}$  for  $m > n$ ;
- it can send top-down information / modulation  $T_{n,l}$  for  $n > l$ ;
- it may emit autonomously some behaviors on the behavior space  $B_n$  by issuing motor commands  $M_n$  with weight / priority  $P_n$  and / or by providing top-down modulation  $T_{n,l}$ ;
- the value of the priority  $P_n$  is not necessarily coupled to level  $n$ , see for example underlying stabilizing processes like balance control etc.;
- a unit  $n$  can choose to work solely based on the input space  $X$  without other representations  $R_{m \neq n}$ ;
- the coupling between the units is such that the behavioral space covered by the system is  $\bigoplus_n B_n$ , denoting the vector product or direct sum of the individual behavior spaces;
- the behaviors  $B_n$  may have different semantics  $Z_j$  depending on the current situation or context  $C_i$ , i.e. the behaviors  $B_n$  represent skills or actions from the system's point of view rather than observer dependent quantities;
- the motor commands of different units may be compatible or incompatible. In the case of concurrently commanded incompatible motor commands a conflict resolution decides based on the priorities;
- all entities describing a unit may be time dependent.

The index  $n$  represents the index of creation in an incremental system. Therefore, units with a lower index  $n$  cannot observe the representations  $R_m$  of units with a higher index  $m$ . The combination and conflict resolution is not to be understood as the primary instance for such cases but rather as the last resort. Conflicts and combinations must be treated as major issues between and inside of the units of the architecture, e.g. according to the biological principles of inhibition and disinhibition. The sensory space  $S_n(X)$  can be split into several aspects for clearer reference. The aspects that are concerned with the location of the corresponding entity in the world are termed  $S_n^L(X)$ , and the features are termed  $S_n^F(X)$ . Correspondingly, the behavior space  $B_n$  can be split into parts concerned with the potential location of the actions (termed  $B_n^L$ ), and the qualitative skills or motions (termed  $B_n^S$ ).

We use the term behavior in the meaning of an externally observable state change of the system. This comprises actions and motion as well as speech and communication. The behavior space  $B_n^S$  is spanned by the effective degrees of freedom or order parameters of the dynamical system  $D_n$  of the unit. In a wider sense, it is spanned by the parameters that are governing changes in the stereotypical actions controlled by the respective unit.

The presented framework allows to characterize the architecture of such systems with respect to the following issues: Find a system's decomposition or a procedure to decompose or construct units  $n$  consisting of  $S_n(x)$ ,  $D_n$ ,  $B_n$ ,  $R_n$ ,  $M_n$ ,  $P_n$ ,  $T_{m,n}$  such that

- an incremental and learning system can be built;
- the system is always able to act, even if the level of performance may vary;
- lower level units  $n$  provide representations and decompositions that
  - are suited to show a certain behavior at level  $n$ ,

- are suited to serve as auxiliary decompositions for higher levels  $m > n$ , i.e. make the situation treatable for others, provide an "internal platform" so that higher levels can learn to treat the situation.

In our understanding, a necessary condition for achieving the abovementioned system properties is a hierarchical arrangement of sensory and behavioral subspaces, the representations and top-down information. Another crucial aspect is the separation of behaviors from the semantics of the behaviors in a certain context. We will discuss this aspect in more detail in subsection 5.3.

Due to space limitations we forbear from a further in-depth mathematical definition and treatment of the presented terms. The concrete system presented in subsection 5.3 should elucidate the underlying concepts in a graspable fashion.

## 5.2 Biological Embedding of SYSTEMATICA

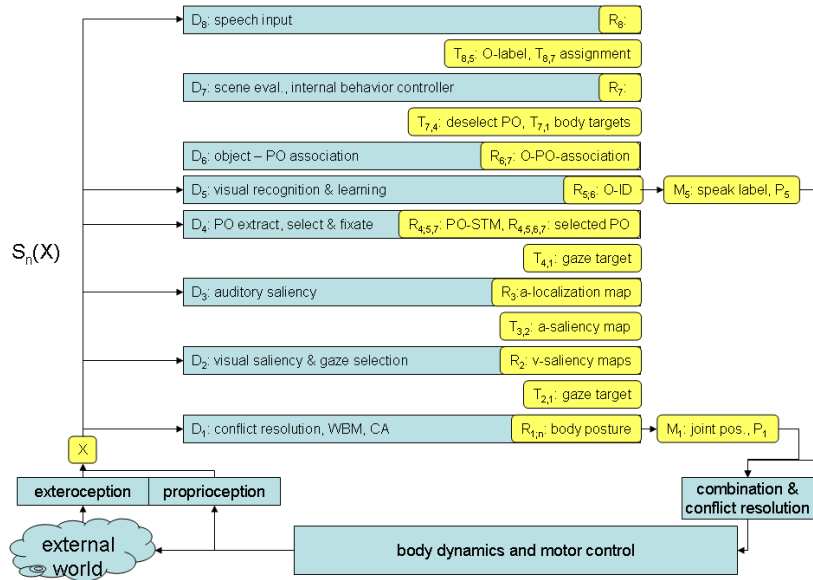
If the goal is to research brain-like intelligent systems, the creation of a fixed hierarchy with units stacked on top of each other is not sufficient: the interplay of the units is the crucial issue. In the classical subsumption paradigm the interplay within a hierarchy is modeled as inhibition of sensory signals and motor commands. We argue that a deeper communication between the units is biologically more plausible and beneficial, because it is more efficient in terms of (re-)using already established representations and processes. The biological motivation of a sensory space  $X$  that is in principle accessible for all levels of the hierarchy has already been discussed in [14]. The individual subspaces  $S_n(X)$  may of course differ. The same applies to the direct access from higher levels of the hierarchy to the motors and actuators, with additional evidence given in [21]. This may not correspond to the predominant signal flows, but is in some cases necessary for the acquisition of completely new motions. The difference between lower and higher levels is mainly that lower levels act on a coarser level of the sensory signals and do not allow for a fine control of actuators. A very fine analysis of sensory signals and a corresponding fine control of e.g. finger motions is subject to cortical and not sub-cortical regions of the brain [22]. What is mainly not addressed in the technical literature is the synergistic interplay of the different levels of the hierarchy. The main issues are the following:

- underlying control processes in the brain perform a basic stabilization and allow higher areas to modulate those stabilizations according to some goals. This is e.g. the case for the balance and the upright standing of the human body that is maintained by the brain stem (mid brain, hind brain and medulla oblongata) [23]. The higher areas in the brain rely on those functional loops.
- Specific structures in the brain maintain representations  $R_n$  for their own purposes, but those representations are also observed and used by areas created later in evolution. This is e.g. the case for the superior colliculus. The target for the next gaze direction is observed by the cortex [24]. A similar reasoning applies to the area AIP, where the coarse information about graspable objects is maintained, which is observed by the Premotor Cortex and used for configuring and target setting of the motor cortex [25].

c) Lower level structures can autonomously perform certain actions but can be modulated from higher level structures by top-down information ( $T_{n,m}$ ). An example is here again the superior colliculus. In reptiles it directly controls sensory based behaviors as the highest level of control. In humans, it can control the gaze direction based on visual and auditory signals if “permitted” by the cortex. If the cortex is damaged, the superior colliculus can take over control again.

The presented SYSTEMATICA serves to organize such a kind of incremental design in a way that the resulting complexity and cross dependencies are still treatable. Compared to so-called cognitively oriented architectures, the approach presented here is de-central with respect to processes and representations involved. The incremental direction is here to be understood in a developmental sense with a number of levels, less as incrementally adding more functionality at already existing levels in the system.

### 5.3 ALIS: Architecture and Elements



**Fig. 6.** Schematics of ALIS formulated in the framework SYSTEMATICA. For explanation please refer to subsection 5.3.

Based on SYSTEMATICA we will now describe the architecture and elements of ALIS. ALIS represents an incrementally integrated system including visual and auditory saliency, proto-object based vision and interactive learning, object dependent autonomous behavior generation, whole body motion and self collision

avoidance on the humanoid robot ASIMO. The elements of the overall architecture are arranged in hierarchical units that produce the overall observable behavior, see Figure 6. The corresponding areas from PISA are auditory and visual perception, presence, movements and behaviors.

The first unit with dynamics  $D_1$  is the whole body motion control of the robot, including a basic conflict resolution for different target commands and a self collision avoidance of the robot’s body as described in section 3. It receives the current robot posture as sensory data. The top-down information  $T_{n,1}$  providable to the unit is in the form of targets for the right and left hand respectively, the head and the walking. Any other unit can provide such kind of targets. Without top-down information, the robot is standing in a rest position with a predefined posture at a predefined position. The posture and the position are controlled, i.e. if the top-down information is switched off, the robot walks back to the predefined home position while compensating for external disturbances. The behavior subspace  $B_1$  comprises target reaching motions including the whole body while avoiding self collisions. The subspace  $B_1^S$  is spanned by variables controlling the choice of the respective actuator group: mainly the gaze, the hands and the body’s position and orientation in 3D. The subspace  $B_1^L$  comprises the area that is covered by walking and that can be reached by both hands. Many different kinds of semantics  $Z_j$  can be attributed to those motions like “pointing”, “pushing”, “poking” and “approaching” etc. The representation  $R_1$  used and provided is a copy of the overall posture of the robot. Unit 1 provides motor commands  $M_1$  to the different joints of the robot and establishes the body control level many other units can incrementally build upon. It unloads much of the tedious control from higher level units.

The second unit with  $D_2$  comprises a visual saliency computation based on contrast measures for different cues and gaze selection as partially described in section 4. Based on the incoming image, visually salient locations in the current field of view are computed and fixated by providing gaze target positions as top-down information  $T_{2,1}$  to unit 1. The spatial component  $S_2^L(X)$  of the sensory space comprises the field of view covered by the cameras.

The representation  $R_2$  comprises the saliency maps, their modulations and the corresponding weights. As top-down information  $T_{n,2}$ , the modulations and the corresponding weights can be set. Depending on this information, different kinds of semantics  $Z_j$  like “visual search”, “visual explore” and “fixate” can be attributed to the behavior space  $B_2$  emitted by this unit. The subspace  $B_2^S$  is spanned by the weights of the different cues, the time constant of the fixation and the time constant for inhibition of return as described in [26]. The unit performs an autonomous gaze control that can be modulated by top-down information. It builds on unit 1 in order to employ the whole body for achieving the commanded gaze direction.

The unit with  $D_3$  computes an auditory localization or saliency map  $R_3$ . It is provided as top-down information  $T_{3,2}$  for unit 2, where the auditory component is higher weighted than the visual. The behavior space  $B_3$  comprises the fixation of prominent auditory stimuli, which could semantically be inter-

preted as “fixating a person that is calling the robot”. The space is spanned by the weight balancing the auditory versus the visual saliency maps. The sensory space  $S_3^F(X)$  is spanned by binaural time series, the spatial component  $S_3^L(X)$  is the area all around the robot. The corresponding auditory processing is described in [27]. Unit 3 builds on and employs the gaze selection mechanism of unit 2. The combination of both units 2 and 3 corresponds to an autonomous gaze selection based on visually and auditory salient stimuli.

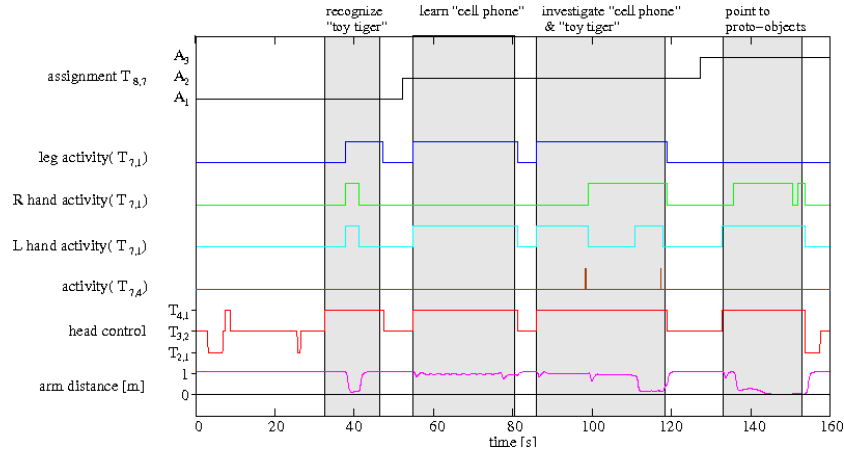
Unit 4 extracts proto-objects from the current visual scene and performs a temporal stabilization of those in a short term memory (PO-STM). The computation of the proto-objects is purely based on depth and peripersonal space (see below), i.e.  $S_4^L(X)$  is a range limited subpart of  $S_2^L(X)$ . The PO-STM and the information which proto-object is currently selected and fixated form the representation  $R_4$ . The top-down information  $T_{4,1}$  provided to unit 1 are gaze targets with a higher priority than the visual gaze selection, yielding as behaviors  $B_4$  the fixation of proto-objects in the current view. The unit accepts top-down information  $T_{n,4}$  for deselecting the currently fixated proto-object or for directly selecting a specific proto-object. The concept of the proto-object as we employ it for behavior generation is explained in more detail in [6]. The main difference between the approach described there and this one is the extraction of the proto-objects from the scene. Here we are extracting three dimensional descriptions of approximately convex three dimensional blobs within a certain distance range from the robot. As introduced in section 4, we call this range the peripersonal space.

The combination of the units 1-4 autonomously realizes the framework for the interaction with the robot. Seen from the robots point of view, the “far-field” interaction is governed by the visual and auditory saliency computation and gaze selection computations. The close-to-the-body or peripersonal interaction is governed by the proto-object fixation. Those processes run continuously without an explicit task and take over control depending on the location of the interaction w.r.t. the robot’s body.

Unit 5 is based on the incrementally established interaction framework. It performs a visual recognition or interactive learning of the currently fixated proto-object without own control of the robot. The sensory input space  $S_5^L(X)$  is the same as  $S_4^L(X)$ , the feature space  $S_5^F(X)$  is the full color image and the corresponding depth map. The unit relies on the representation  $R_4$  for extracting the corresponding sub-part of the information from  $S_5(X)$ . The three-dimensional information of the currently fixated proto-object is used to extract the corresponding segment from the high resolution color image space. The segments are being classified w.r.t. the object identity O-ID. For newly learned objects, the target identity has to be provided as top-down information  $T_{n,5}$ . The representation  $R_5$  is the object identity O-ID of the currently fixated proto-object. The motor commands  $M_5$  emitted by the unit are speech labels corresponding to the object identity. The unit described here corresponds mainly to our work described in [9, 28] and section 4. The object identity O-ID is the first instance of fixed semantics, since we use user-specified labels like “blue cup” or “toy car”.



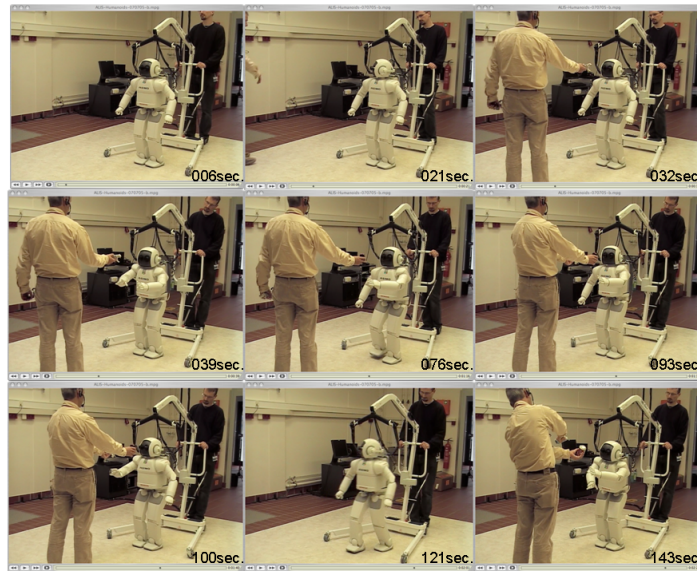
From the incremental architecture point of view, we now have a system that additionally classifies or learns the objects it is currently fixating.



**Fig. 7.** Measurements from the interactive experiment. In the time range from sec. 0 until sec. 32, ALIS is mainly driven by saliency based interaction with the world. From sec. 32 until sec. 47 the human is presenting a known object, from sec. 54 until sec. 81 the system is learning an unknown object. From sec. 86 until sec. 118 two objects are presented by the human, sequentially attended and recognized. From sec. 135 on two objects are being presented by the human and continuously pointed at by the robot. Please refer to subsection 5.4 for further explanations.

Unit 6 performs an association of the representations  $R_4$  and  $R_5$ , i.e. it maintains an association  $R_6$  between the PO-STM and the O-IDs based on the identifier of the currently selected PO. This representation can provide the identity of all classified proto-objects in the current view. Except for the representations it has no other inputs or outputs. From the incremental point of view we have now an additional memory of all classified proto-objects in the current view.

Unit 7 with  $D_7$  builds on the sensory processing and control capabilities of many of the underlying units. It governs the control of the robot's body except for the gaze direction. This is achieved by deriving targets from the proto-object representation  $R_4$  and sending them as top-down information  $T_{7,1}$  for the right and the left hand as well as for walking to unit 1. Additional top-down information  $T_{7,4}$  can be sent to the proto-object fixating unit 4 for requesting the selection of another proto-object. Details of the internal dynamics  $D_7$  can be found in [29]. Here, it is based on the evaluation of the current scene as represented by  $R_4$  (proto-object short term memory) and  $R_6$  (association object identifier and proto-object identifier) and the top-down information  $T_{n,7}$  concerning the current assignment. An assignment is an identifier for a global mode of the internal dynamics of unit 7. The first realized assignment (A1) is pointing once with the most appropriate hand or both hands to the fixated and



**Fig. 8.** Image series from the interactive experiment. From top left row-wise to bottom right. Rest position (sec. 6), saliency based interaction (sec. 21), proto-object fixation (sec. 32), fixation and both handed pointing after recognition (sec. 39), learning of a new object (sec. 76), fixation and pointing to first object of two (sec. 93), fixation and pointing to second object of two (sec. 100), return to rest position (sec. 121), pointing to two proto-objects (sec. 143). For further explanation see subsection 5.4.

classified proto-object. The second assignment (A2) differs from the first one in the respect that pointing is continuous and immediate to the fixated and not yet classified proto-object. Whether the pointing is done using a single hand or both arms depends on the currently arbitrarily defined category of the classified object: both-handed pointing for toys, single handed pointing for non-toys. The definition is currently associated with the labels of the objects. During both assignments, the distance to the currently fixed proto-object can autonomously be adjusted by walking. Additionally, the autonomous selection of a new proto-object is requested ( $T_{7,4}$ ) from the proto-fixation if the currently fixated one has been classified successfully two times. This allows for a first autonomous scene exploration. The third assignment (A3) is pointing with each hand at a proto-object irrespective of the classification result and without walking. The behavioral space spanned by this unit is a subspace or a sub-manifold of  $B_1$ . The semantics of the behaviors are currently fixed by design, like “both handed pointing to toys” etc. From the incremental design point of view unit 7 is a thin layer controlling different kinds of interaction semantics for the body based on the sensory processing and control capabilities provided by the underlying system.

The last unit 8 works on another audio stream  $S_8(X)$  and processes speech input. The results are currently provided as object labels for the recognizer ( $T_{8,5}$ ) and as assignments for unit 7 ( $T_{8,7}$ ). It serves for establishing verbal interaction with the user in the current setting.

In summary, the presented system consists of several independently defined units that build on each other in an incremental way for yielding the combined performance. Due to the incremental nature of the architecture, the units can be implemented, tested and integrated one after the other, which is an important means for dealing with the increasing complexity of the targeted system.

The described system, except some parts of unit 1, is implemented in our framework for distributed real-time applications [30] and runs with 10Hz for the command generation in interaction. The implementation consists of 288 processing components. The workload is distributed across 10 standard CPUs in 6 computers without any further optimization.

## 5.4 Experiments

Users can freely interact with the running ALIS. The behavior of the system is governed mainly by the interaction. Figure 7 shows the measurements of a recorded experiment. The bottom most graph shows the measured minimal distance between the arms, because the self collision of the arms constitutes in this experiment the highest risk. The next higher graph depicts which of the possible top-down feedback  $T_{4,1}$  (proto-fixation),  $T_{3,2}$  (auditory saliency) or  $T_{2,1}$  (visual saliency) is controlling the gaze direction. The graph with the label “activity  $T_{7,4}$ ” shows the occurrence of the request for fixating a new proto-object by the proto-fixation unit 4. The graphs with the labels “L hand activity ( $T_{7,1}$ )”, “R hand activity ( $T_{7,1}$ )” and “leg activity ( $T_{7,1}$ )” depict the active control of the

respective effector group by unit 7. The topmost graph with the label  $T_{8,7}$  shows the currently valid assignment, namely A1, A2 and A3 in a sequence.

The following time course is shown in Figure 7. From the beginning until second 32, ASIMO is mainly interacting with its environment by gazing at far distance visual and auditory stimuli. Beginning with second 32, the user presents an object in the peripersonal space, which is immediately fixated by means of the control of unit 4. At second 37, the object is successfully recognized as a “toy-tiger” and pointed at once with both hands since it belongs to the category “toys”. After pointing, the object is still fixated and the distance is adjusted by walking until second 47. After termination of the close interaction by the human, ASIMO returns autonomously to the rest position. At second 52 the assignment is switched to A2, and starting with second 54 ASIMO fixates and continuously points to the presented proto-object. It is unknown and learned in interaction as “cell phone” until second 81 when ASIMO returns back to the home position. At second 86 the previously trained “cell phone” is presented together with the “toy-tiger”. The cell phone is fixated and pointed at, and successfully recognized at second 91. At second 98 it is successfully recognized for the second time and the fixation of a new proto-object is requested from unit 7 to unit 4 by the activity of  $T_{7,4}$ . At second 105 the toy-tiger is first misclassified, but subsequently recognized at second 111 and second 117. At second 127 the assignment is changed to A3, and at second 135 ASIMO starts pointing at two objects with both hands. The user tries to force a self collision crossing the arms with the fixated proto-objects until the arms touch each other. This is depicted in the arm distance plot, which comes close to the limit of a self-collision but never reaches it. The self collision is prevented by the continuously running self collision avoidance of unit 1. After the termination of the close interaction, ASIMO returns to the rest posture. Figure 8 shows some snapshots from the running experiment.

The sequence of the interaction is just an example, the resulting behavior as well as all motions of the robot are computed online and depend on the interaction of the user with the robot.

## 5.5 Discussion

After the presentation of the conceptual framework (SYSTEMATICA), the instance (ALIS) and the experiments we would like to point out some of the key features.

- Units run autonomously and without explicit synchronization mechanisms in parallel. The undirected publication of the representations  $R_n$  and the directed top-down information  $T_{n,m}$  establish a data driven way of synchronization depending on activity.
- The top-down information flow is not restricted to the communication between two adjacent layers but can project from any higher to any lower level.
- Unit 1 provides the basis for higher level units to control the robot’s hands, head and steps positions including the avoidance of self collisions. It “unloads” a lot of detailed knowledge about the robots kinematics of the higher

level units. This kind of unloading allows for an easier incremental design or development of the system.

- The space  $S_3^L(X)$  covered by the audio saliency is the largest one: it includes the space  $S_2^L(X)$  covered by the visual saliency, which again includes the sensory space  $S_4^L(X)$  of the current implementation of the peripersonal space. The arrangements of these spaces and the corresponding behavior space serve as the basis for getting and staying in interaction with the system.
- The lower level units are to a large extent free of specific semantics. Higher-level units like 5 and 7 temporarily define the semantics for the lower units.
- The same physical entity can be represented / perceived by different sensory spaces. The proto-object extraction of unit 4 is based on grey value stereo image pairs on a low resolution for extracting the three-dimensional information. The visual recognition of unit 5 is based on a high resolution color image segment. The segment is extracted from this color image based on the information from the currently fixated proto-object. The segment is extracted at the time of the classification, not at the time of the extraction of the proto-object. Based on this arrangement, the classifier can easily be combined with the proto-object fixation loop. The feature part of the sensory space of unit 4 is more coarsely resolved than the feature space of unit 5.
- The location part of the behavior space of one unit may dynamically extend the location part of the sensory space of another unit. This is, for example, the case for the peripersonal space  $S_4^L(X)$  that is dynamically extended by adjusting the distance by unit 7.

The presented system has already a certain complexity and shows some important features, but the question of scalability has to be addressed. ALIS is already working in the real world in real-time interaction, which covers the aspect of scaling / bringing a concept to the real world. Asking about the scalability to more complex and prospective behaviors is a crucial point. We are confident to be on the right track because of the following reasons: Each of the hierarchical layers individually already performs some meaningful behavior, and some of them additionally serve as building blocks for more complex systems. This is facilitated via the coupling of the units by the publicly observable representations and directed top-down information, for us a key issue in successful scaling. A more loose argument for now but subject to current research is the following: Biology seems to have taken a similar route in evolving the brains of animals towards the brains of humans by phylogenetically adding structures on top of existing structures, and maybe mildly changing the existing structures. The communication between the “older” and the “newer” structures can be seen as providing existing representations and sending top-down information from the “newer” structures to the “older” ones. Does the presented approach scale in the direction of learning and development? We consider the visual object learning as a successful start in this direction. Nevertheless, the step towards learning is currently done only on the perceptive side. The learning on the behavior generation side is not explicitly addressed here, but in section 4 and [31] we showed our

approach towards using general developmental principles for the adaptation of reactive behaviors. Transferring this work into the presented architecture would formally require the addition of another unit and some changes in existing ones. This argument is of course made irrespective of the many open scientific questions involved in actually doing this step because the system considered in [31] is considerably simpler than the one discussed here. Nonetheless, it makes us confident about the scalability of the proposed architecture.

Summarizing this section, we have presented the conceptual framework SYSTEMATICA for describing and designing incremental hierarchical behavior generation systems. A framework like this is crucial for researching more complex intelligent systems. On the one hand, it provides the concepts handling the growing complexity, on the other hand it establishes a necessary common language for the collaboration of several researchers. Within this framework we have created the system ALIS, integrated with ASIMO. ALIS allows for the first time the free interaction of a human with a full size biped humanoid including non-preprogrammed whole body motions, interactive behavior generation, visual recognition and learning.

## 6 Summary

In this paper we have presented research concerned with elements and systems aiming at embodied brain-like intelligence and cognitive robotics. We started with presenting our guiding model PISA, followed by sections about movement generation and control as well as visually oriented behavior generation and learning. The last section contains a view on SYSTEMATICA and ALIS, showing our research in the area of large scale intelligent systems. The work presented here should show that we are researching and creating in an incremental and holistic fashion, leading to a better understanding of natural and artificial brain-like systems.

## 7 Acknowledgments

The authors would like to thank Michael Gienger, Herbert Janßen, Hisashi Sugiura, Inna Mikhailova, Bram Bolder, Mark Dunn, Heiko Wersing, Stephan Kirstein, Julian Eggert, Antonello Ceravola, Frank Joublin and Edgar Körner for their contributions, support and advice.

## References

1. Pirjanian, P.: Behavior coordination mechanisms – state-of-the-art. TR IRIS-99-375. Technical report, Institute of Robotics and Intelligent Systems, School of Engineering, University of Southern California (1999)
2. Vernon, D., Metta, G., Sandini, G.: A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation* **11**(2) (2007) 151–180

3. Kandel, E.R., Schwartz, J.H., Jessell, T.M.: Principles of Neural Science, Fourth Edition. McGraw-Hill (2000)
4. Gienger, M., Janßen, H., Goerick, C.: Task-oriented whole body motion for humanoid robots. In: IEEE/RSJ Int. Conf. on Humanoid Robots. (2005)
5. Sugiura, H., Gienger, M., Janßen, H., Goerick, C.: Real-time self collision avoidance for humanoids by means of nullspace criteria and task intervals. In: IEEE-RAS Int. Conf. on Humanoid Robots, IEEE Press (2006)
6. Bolder, B., Dunn, M., Gienger, M., Janßen, H., Sugiura, H., Goerick, C.: Visually guided whole body interaction. In: IEEE Int. Conf. on Robotics and Automation. (2007)
7. Nishiwaki, K., Kuga, M., Kagami, S., Inaba, N., Inoue, H.: Whole-body cooperative balanced motion generation for reaching. In: IEEE/RSJ Int. Conf. on Humanoid Robots. (2004)
8. Sian, N., Yokoi, K., Kajita, S., Tanie, K.: A framework for remote execution of whole body motions for humanoid robots. In: IEEE/RSJ Int. Conf. on Humanoid Robots. (2004)
9. Goerick, C., Mikhailova, I., Wersing, H., Kirstein, S.: Biologically motivated visual behaviors for humanoids: Learning to interact and learning in interaction. In: IEEE/RSJ Int. Conf. on Humanoid Robots. (2006)
10. Couyoumdjian, A., Nocera, F.D., Ferlazzo, F.: Functional representation of 3d space in endogenous attention shifts. The quarterly Journal of Experimental Psychology **56a**(1) (2003) 155–183
11. Wersing, H., Kirstein, S., Götting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J., Ritter, H., Körner, E.: A biologically motivated system for unconstrained online learning of visual objects. In: Proc. Int. Conf. Art. Neur. Netw. ICANN. (2006)
12. Ude, A., Cheng, G.: Object recognition on humanoids with foveated vision. In: IEEE/RSJ Int. Conf. on Humanoid Robots. (2004)
13. Goerick, C., Bolder, B., Janßen, H., Gienger, M., Sugiura, H., Dunn, M., Mikhailova, I., Rodemann, T., Wersing, H., Kirstein, S.: Towards incremental hierarchical behavior generation for humanoids. In: Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2007), Pittsburgh, USA. (2007)
14. Prescott, T.J., Redgrave, P., Gurney, K.: Layered control architectures in robots and vertebrates. Adaptive Behavior **7** (1999) 99–127
15. Pfeiffer, R., Scheier, C.: Understanding Intelligence. MIT Press, Cambridge, Massachusetts, USA (1999)
16. Brock, O., Fagg, A., Grupen, R., Platt, R., Rosenstein, M., Sweeney, J.: A framework for learning and control in intelligent humanoid robots. International Journal of Humanoid Robotics **2**(3) (2005)
17. Arkin, R.C., Fujita, M., Takagi, T., Hasegawa, R.: An ethological and emotional basis for human-robot interaction. Robotics and Autonomous Systems (3-4) (2003) 191–201
18. Chernova, S., Arkin, R.C.: From deliberative to routine behaviors: A cognitively inspired action-selection mechanism for routine behavior capture. Adaptive Behavior **15**(2) (2007) 199–216
19. Asfour, T., Regenstein, K., Azad, P., Schröder, J., Bierbaum, A., Vahrenkamp, N., Dillmann, R.: ARMAR-III: An integrated humanoid platform for sensory-motor control. In: Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2006), Genoa, Italy. (2006)

20. Gat, E.: On three-layer architectures. In: Artificial Intelligence and Mobile Robots. MIT/AAAI Press (1997)
21. Chouinard, P.A., Paus, T.: The primary motor and premotor areas of the human cerebral cortex. *The Neuroscientist* **12**(2) (2006) 143–152
22. Swanson, L.W.: Brain Architecture: Understanding the Basic Plan. Oxford University Press Inc, USA (2002)
23. Purves, D., Augustine, G.J., Fitzpatrick, D., , Hall, C., Lamantia, A.S., McNamara, J.O., Williams, S.M., eds.: Neuroscience. Sinauer Associates (2004)
24. Sommer, M.A., Wurtz, R.H.: What the brain stem tells the frontal cortex. I. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *Journal of Neurophysiology* **91** (2004) 1381–1402
25. Battaglia-Mayer, A., Caminiti, R., Lacquaniti, F., Zago, M.: Multiple levels of representation of reaching in the parieto-frontal network. *Cerebral Cortex* **13** (2003) 1009–1022
26. Goerick, C., Wersing, H., Mikhailova, I., Dunn, M.: Peripersonal space and object recognition for humanoids. In: Proceedings of the IEEE/RSJ International Conference on Humanoid Robots (Humanoids 2005), Tsukuba, Japan. (2005)
27. Rodemann, T., Heckmann, M., Schölling, B., Joublin, F., Goerick, C.: Real-time sound localization with a binaural head-system using a biologically-inspired cue-triple mapping. In: Proceedings of the International Conference on Intelligent Robots & Systems (IROS), IEEE (2006)
28. Wersing, H., Kirstein, S., Goetting, M., Brandl, H., Dunn, M., Mikhailova, I., Goerick, C., Steil, J.J., Ritter, H., Körner, E.: Online learning of objects in a biologically motivated visual architecture. *International Journal of Neural Systems* **17**(4) (2007) 219–230
29. Bergener, T., Bruckhoff, C., Dahm, P., Janßen, H., Joublin, F., Menzner, R., Steinhage, A., von Seelen, W.: Complex behavior by means of dynamical systems for an anthropomorphic robot. *Neural Networks* (7) (1999) 1087–1099
30. Ceravola, A., Stein, M., Goerick, C.: Researching and developing a real-time infrastructure for intelligent systems - evolution of an integrated approach. *Robotics and Autonomous Systems* **56**(1) (2008) 14–28
31. Mikhailova, I., von Seelen, W., Goerick, C.: Usage of general developmental principles for adaptation of reactive behavior. In: Proceedings of the 6th International Workshop on Epigenetic Robotics, Paris, France. (2006)