Situated robot learning for multi-modal instruction and imitation of grasping

J. J. Steil[a], F. Röthling[a], R. Haschke[a], and H. Ritter[a]

[a] Bielefeld University, Neuroinformatics Group, Faculty of Technology,
P.O. Box 10 01 31, D-33501 Bielefeld, Germany

A key prerequisite to make user instruction of work tasks by interactive demonstration effective and convenient is situated multi-modal interaction aiming at an enhancement of robot learning beyond simple low-level skill acquisition. We report the status of the Bielefeld GRAVIS-robot system that combines visual attention and gestural instruction with an intelligent interface for speech recognition and linguistic interpretation to allow multi-modal task-oriented instructions. With respect to this platform, we discuss the essential role of learning for robust functioning of the robot and sketch the concept of an integrated architecture for situated learning on the system level. It has the long-term goal to demonstrate speech-supported imitation learning of robot actions. We describe the current state of its realization to enable imitation of human hand postures for flexible grasping and give quantitative results for grasping a broad range of everyday objects.

1. Introduction

How can we endow robots with enough cognitive capabilities to enable them to serve as multi-functional personal assistants that can easily and intuitively be instructed by the human user? A key role in the realization of this goal plays the ability of *situated learning*: Only, when we can instruct robots to execute desired work tasks by means of a combination of spoken dialog, gestures, and visual demonstration, robots will loose their predominant role as specialists for repeatable tasks and become effective to support humans in everyday life.

A basic element of *situated learning* is the capability to observe and successfully *imitate actions* and – as a prerequisite for that – to establish a common focus of attention with the human instructor. For multi-modal communication, additional perceptive capabilities in the fields of speech understanding, active vision, and in the interpretation of non-verbal cues like gestures or body posture are essential and have to be included and coordinated.

We report on progress in building an integrated robot system within the framework of the Special Collaborative Research Unit SFB 360 "Situated Artificial Communicators". In the course of this long-term program, many modules implementing partial skills were at first realized and evaluated as stand alone applications [4,7,18,20,34], but their integration is an additional research task and a key issue towards the realization of intelligent machines.

As the development of integrated learning architectures for real world tasks poses an enormous challenge, there can hardly be found any efforts to scale learning from the lower level of training single skills up to a multi-stage learning across the overall system. A primary reason is that most learning approaches rely on highly pre-structured

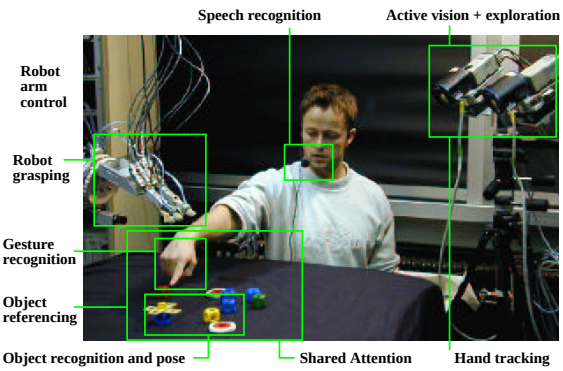


Figure 1. Interaction with the GRAVIS-system using speech and gesture.

information and search spaces. Prominent examples are supervised learning of target outputs, unsupervised learning of clusters, or learning of control tasks with a (usually small) number of predefined variables (pole balancing, trajectory learning). Here exist well understood approaches like gradient based learning, support vector machines, vector quantization, or Q-learning, which yield for certain tasks remarkable results, e.g. in speech-image integration [26], trajectory learning [22,19,44], in object recognition and determination of grasp postures [28], sensor fusion for grasp planning [1], or grasp optimization [30].

In real world learning a well defined prestructuring of the data with respect to the given task is an essential part of the learning itself: the system has to find lower-dimensional relevant manifolds in very high-dimensional data and to detect important regularities in the course of learning to use these to improve its capabilities. Furthermore, for a sophisticated robot with many motor degrees of freedom or for a cognitive system – as the one discussed here – finding a solution by exploration of new actions is not suitable because the search spaces involved are extremely high-dimensional and by far too complex.

Current practice aims at developing well-scalable, homogeneous and transparent architectures to create complex systems. Somewhat ironically, successful examples of this strategy tend to cluster in the small- or mid-size range, while truly large and complex systems seem to defy our wishes for "formatting away" their complexity by good bookkeeping alone. It seems not unlikely that it is one of the hallmarks of complex systems that they confront us with limited homogeneity, evolutionarily grown layers of overlapping functionality and bugs that may even amalgamate with features. Looking at biological systems with their enormous complexity, we see that these by no means resemble orthogonal clockworks; instead, they consist of a tangle of interwoven loops stabilized by numerous mechanisms of error-tolerance and self-repair. This suggests that a major challenge for moving to higher complexity is to successfully adopt similar approaches to come to grips with systems that we cannot analyze in their full detail.

In the present paper, we address these issues in the context of a longer-term research project aiming at the realization of a robot system that is instructable by speech and gestures. For the aforementioned reasons, we have pursued the development of this system in an evolutionary fashion, without the requirement that a global blueprint had to be available at each stage of its development. In Sec. 2, we report our experiences with this approach and give an overview of the current stage of the evolved system.

In Sec. 3, we focus our discussion on the issue of learning within such a system and argue for three major levels at which learning has to be organized: (*i*) an *ontogenetic level* which exploits learning methods in order to create initial system functions (such as object classifiers) from previously acquired training data in an off-line fashion, (*ii*) a *refinement level* at which on-line learning is used locally within a functional module, with the main effect of increasing the module's robustness or refining its performance, but with no or little need of explicit coordination with adaptive processes in other modules, and (*iii*) a *situated level* at which different learning methods are combined in a highly structured way in order to achieve short-term situated learning at the task level. While all three learning levels are important, undoubtedly it is the uppermost, *situated level* which currently poses the most exciting research challenge.

In Sec. 4, we propose an approach how to organize learning at this level. Our proposal is strongly motivated by the idea of imitation learning [2,3,6,8,23,24,32], which attempts to find a successful "action template" from the observation of a (human) instructor. This requires (i) to endow the robot system with sufficient perceptive capabilities to recognize and observe the action to imitate; (ii) to transform the observed action into an internal representation, which is well matched to the system's own operating characteristics (in particular, its different "sensory perspective" and "instrumentation" with actuators); (iii) to be able to physically execute a suitable action by itself. Focusing on the important task of *imitation grasping*, we describe in Sec. 5 an initial implementation of this scheme, using our current

system as a platform for the necessary and considerable, perceptual and motor anchoring of such an imitation learner in its environment. Sec. 6 then presents some results on imitation grasping of common everyday objects with the system implemented so far.

At all levels, the results of learning can – by its very nature – at best be partially predicted, eroding further the idea of the availability of a fixed system blueprint. In Sec. 7, we therefore argue for a *datamining perspective* for coping with systems of such kind. As a concrete example, we briefly describe a powerful multi-modal monitoring system (AVDisp) that has been developed in our lab very recently and we report some experiences from applying this approach to our robot system. Finally, Sec. 8 presents some conclusions.

2. System Design and Overview

Due to the long-term development of our system, the ideal perspective to define constraints and a unified framework beforehand to facilitate building a cognitive learning architecture had to be replaced by an "evolutionary approach" to integrate also modules that were developed in different research contexts and not necessarily designed in view of being utilized in the described system. This led to the development of a rather flexible architecture, based on a distributed architecture communication system (DACS [14]) developed earlier in the framework of the SFB 360. In this framework, very heterogeneous components can be accommodated as separate and parallel processes, distributed over several workstations and communicating mainly via message passing supported by DACS (some modules also use more sophisticated communication facilities of the DACS package). In this way, we have been able to integrate a large number of modules, which use different programming languages (C, C++, Tcl/Tk, Neo/NST), various visualization tools, and a variety of processing paradigms ranging from a neurally inspired attention system to statistical and declarative methods for inference and knowledge representation.

The current system can be coarsely subdivided into four major functional clusters depicted

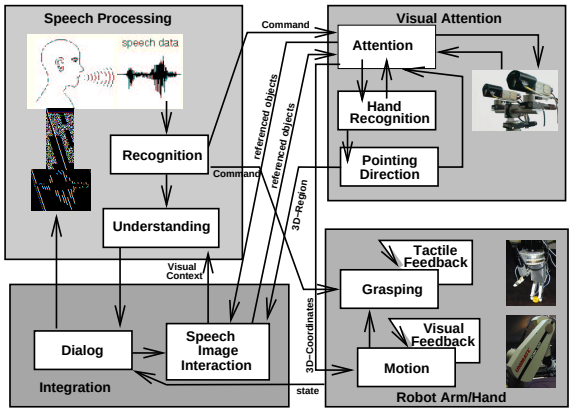


Figure 2. Schematic picture of the current system architecture.

schematically in Fig. 2. The *speech processing* (left) and the *visual attention* mechanism (right) provide linguistic and visual/gestural inputs converging in an *integration* module. This maintains a short-term memory of objects and their 3D-coordinates and passes control to the *arm/hand manipulator* if an object is unambiguously referenced by speech or gesture or their combination. We found that the coordination of all these functional modules can be very conveniently achieved by finite state machines implementing data driven state transitions. Additionally, speech commands like *"calibrate skin", "park robot arm", etc.* can trigger specific actions. The most important functionalities of these four building blocks are summarized below for convenience.

2.1. Visual Attention

The basic behavior of the active vision system is – driven by an attention system – to explore the scene autonomously and to search for salient points and objects, including hands and pointing fingers. The attention behavior is based on an active stereo vision camera head and works in full 3D-space. At the lower level it consists of a layered system of topographically organized neural maps for integrating different feature maps into a continually updated saliency map [20], similar to mechanisms proposed in [9,12,38]. Incorporating an additional fadeout map and results from a hand detection module, it forms a final atten-

tion map whose highest peak determines the next fixation. If a pointing hand is detected, a further module computes a 3D-pointing cone and restricts the attention to the corresponding region. A holistic, neural object recognition system [18] determines whether a known object has been seen and can be transferred into the short-term memory of the integration module.

2.2. Speech Processing and Understanding

To enable speech interaction and communication between the user and the artificial communicator, our system imports a module for speaker-independent speech understanding [13]. The recognition process is directly influenced by a partial parser which provides linguistic and domain-specific restrictions on word sequences derived from previous investigations on a word corpus. Therefore, partial syntactic structures instead of simple word sequences are generated, like e.g. object descriptions (*"the red cube"*) or spatial relations (*"in front of"*). These are combined by the subsequent speech understanding module to form linguistic interpretations. The instructor neither needs to know a special command syntax nor the exact terms or identifiers of the objects. Consequently, the speech understanding system has to face a high degree of *referential uncertainty* from vague meanings, speech recognition errors, and un-modeled language structures.

2.3. Modality Integration and Dialog

In integration of speech and vision, this referential uncertainty has to be resolved with respect to the visual object memory. Here the system uses a Bayesian network approach [40], where the different kinds of uncertainties are modeled by conditional probability tables that have been estimated from experimental data. The objects which are denoted in the utterance are those explaining the observed visual and verbal evidences in the Bayesian network with the maximum a-posteriori probability. Additional causal support for an intended object is defined by an optional target region of interest that is provided by the 3D-pointing evaluation. The intended object is then used by the dialog component for system response and manipulator instruction. The dialog

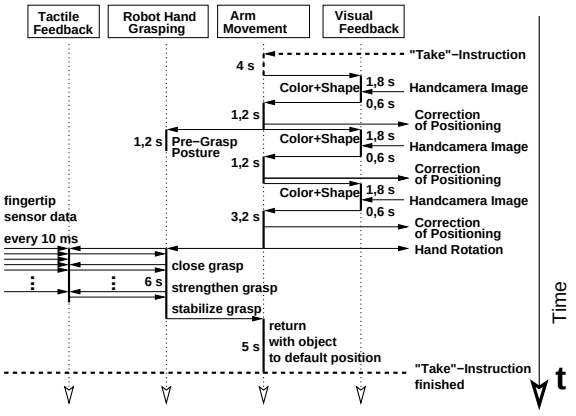


Figure 3. Evaluation of a predefined grasp sequence based on data collected by the system monitoring tool AVDisp.

system is based on an investigation of a corpus of human-human and simulated human-machine dialogs [7]. In particular, it asks for a pointing gesture to resolve ambiguities in the current spoken instruction with respect to the actual state of the memory. The overall goal of this module is to continue the dialog in every situation.

2.4. Robot Arm and Hand

Manipulation is carried out by a standard 6-DOF PUMA manipulator operated with the real-time RCCL-command library [36]. It is further equipped with a wrist camera to obtain local visual feedback during the grasping phase. The grasping is carried out by a 9-DOF dextrous robot hand developed at the Technical University of Munich [37]. It has three approximately human-sized fingers driven by an hydraulics system. The fingertips have custom built force sensors to provide force feedback for control and evaluation of the grasp. Recently we have added a palm and rearranged the fingers in a more human-like configuration in order to allow a larger variety of two- and three-finger grasps and to equip the hand with a tactile display of 8×8 force sensors on the palm.

The grasp sequence starts with an approach movement to the 3D-coordinates determined by the vision and integration modules. Based on visual feedback, it centers the manipulator above the object and finally grasps it, starting from an

initial hand pre-shape, slowly wrapping all fingers around the object and employing force-feedback from the fingertip and palm sensors to detect contacts. Notice, that this grasping strategy significantly differs from analytical approaches, which first compute optimal grasp points on the known object surface and finally employ inverse hand kinematics to achieve these contact points [17]. Fig. 3 shows a time-evaluation of the grasp sequence for a preprogrammed grasp. After a successful grasp, a similar chain of events allows the robot to put the object down at another gesturally selected location. To add flexibility to this fixed grasp behavior and to enable grasping of irregular shaped real world objects, we added an imitation capability in the choice of grasp prototypes based on human observation as detailed in Sections 4 and 5.

3. Structuring of Learning

Learning is a very multi-faceted phenomenon and its complexity is amply reflected in the numerous different proposals how to relate and implement its various aspects. Theoretical considerations motivate a "horizontal subdivision" of learning into the major types of unsupervised, supervised, and reinforcement learning (with still a substantial number of approaches distinguishable both at the conceptual and algorithmic level within each type). Recently there has been a stimulating discussion that such a subdivision may even be reflected in anatomically distinguishable subsystems of the brain [11]. While such a subdivision is highly attractive in many respects, we think that there is another very important but different dimension of learning: the *time scale* at which learning can take place.

At the slowest time scale (which we will call the *ontogenetic level*) learning methods are used in order to create initial system functions by off-line algorithms, usually operating on rather large data sets prepared in contexts that do not require an already trained system. Examples in our system are the unsupervised training of the sensory front end of a neural-network-based object classifier, the supervised training of recognition modules (e.g. for object identity [18] or for continuous hand posture [27] as explained in some more detail in Sec. 5). Learning at this ontogenetic level does not involve any behavior of the robot; instead, it permitted us the creation of important initial subsystem functionalities which would have been much harder to obtain by explicit programming alone. While this level can extend even into quite high levels of abstraction (e.g. the computation of probability tables for Bayes nets to integrate visual and speech recognition results at the symbolic level [40]), its contribution becomes "frozen" afterwards, since the employed techniques frequently assume the availability of the target modules in isolation, without the complex interactions resulting from being embedded in the complete and working system.

The second *local refinement level* comprises those learning processes whose adaptive changes occur on-line, during (and based on) the actual behavior of the robot and refine its initial capabilities. The increased complexity imposed in this way is compensated by requiring the adaptive changes to be (at least largely) local to each module, so that learning processes at the second level can become "encapsulated" in a single functional module, thereby allowing to achieve a good balance between benefits and implementability. Typical examples in our system are the adjustment of subsystem calibrations (e.g. on-line color-recalibration of the vision system), slow adaptation to changing environment conditions (e.g. "habituation" of the fingertip sensors, dynamic renormalization of the feature weights in the saccadic system with respect to the current feature statistics, continuous update of the skin color model, or forgetting factors in the object memory), and mechanisms to ensure that system variables remain well inside their operating ranges. This level can be considered as hosting most of the "long-term plasticity" of our system, and its algorithms can largely be based on ideas of statistical learning. We found that the isolated contribution of the adaptivity of single modules on system performance is often rather small; however, the good tuning of *many* parameters has a big impact on overall system reliability and performance.

Finally, the third *situated level* addresses the

challenge to make rapid (ultimately "one-shot") learning feasible. This cannot rely exclusively on slow and repeated adaptations; instead, this level has to rapidly coordinate adapted subsystem functionalities in very structured, situation-specific and cross-modular ways. Clearly, coming up with workable learning mechanisms at this situated level poses a significant research challenge. The notion of imitation learning has emerged as a very promising paradigm to cope with this challenge. There is a considerable debate in current literature, in which way an architecture for imitation learning could coordinate the required subsystem functionalities. Some researchers focus on using "neurally inspired" building blocks of visual processing [42] as their primitives, others approach the problem from the perspective of attention modeling [12], or perception-action systems [6], implementing either a kind of "imitation at the processing level", or focusing on imitation at the level of joint angles [19]. Whenever such systems execute actions in the real world, their "inner workings" are considerably affected by the physical constraints of the real robot system available. In particular, the advent of more human-like and human-sized robots has had a major impact on the development of techniques for motor-learning and skill transfer in this area.

In the context of our own system, we have started to pursue these ideas within a paradigm of *situated learning for imitation grasping*. In the following Sec. 4 we propose a learning architecture intended to address the issues at the situated level. With the system described in Sec. 2 as a basis, we have started to implement learning on this highest level of abstraction to enable our robot system to carry out a variety of grasps of everyday objects, visually observing and imitating a human instructor who indicates useful grasp postures with his or her own hand.

4. An Architecture for Situated Learning

As pointed out in the previous section, to enable learning at the short time scale of the situated level will depend in an essential way on the highly structured interplay of several functional loops, complementing each other in a tightly coupled fashion to cope with the joint constraints of high-dimensional search spaces and small number of training samples. In the following, we argue that a suitable interlocking of the three functional loops of *observation*, *internal simulation*, and *sensorimotor exploration* can lead to a scheme that appears sufficiently powerful to enable fast situated learning.

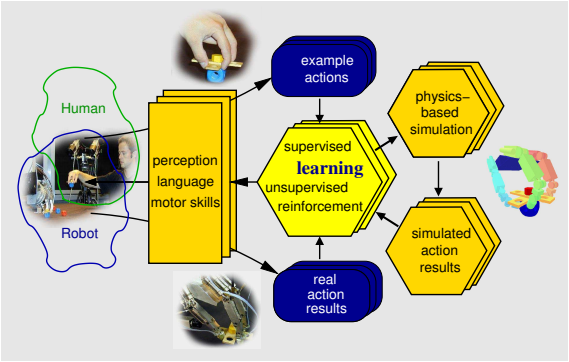


Figure 4. Multistage architecture for integration of different learning paradigms to enable situated learning on the system level. Shown are the interlocking three functional loops of *observation* of example actions, *internal simulation*, and *sensorimotor exploration*.

The objective of the **observation loop** is to gain a promising "action template" that permits to dramatically cut down the otherwise huge search spaces for the other two learning loops. This requires to watch the environment, respectively the instructor, in order to (i) extract relevant features, events, and chains of observed partial actions, (ii) translate these from the observed to an intrinsic perspective, and (iii) exploit them for forming an action template that is focused on promising regions of the a-priori very high-dimensional search space.

At a second stage the observed action template has to be improved further using an **internal simulation loop**, exploring possible actions in the vicinity of the observed template and selecting promising action candidates. At this stage it is important to perform learning from an "intrinsic" perspective in order to incorporate available model knowledge (e.g. about kinematic and sen-

sor constraints of the actually used hardware), which becomes merged with constraint information made available from the observation component. The computational basis of this component is a dynamics-based grasping simulation, allowing the application of reinforcement learning methods to improve the grasping strategy. Due to the availability of full information about joint angles, applied wrenches, or contact points in the simulation, we can make predictions (valid to the extent that the model is accurate) about grasp quality to generate a suitable reward signal even in the absence of corresponding tactile sensors in the actual TUM-hand.

We think that this *search space restricted reinforcement learning*, where the exploration of actions can be restricted to a neighborhood of the observed successful trajectory, is the adequate technique to generate promising action candidates. This combination of observation and reinforcement learning appears very flexible: the neighborhood can be chosen small where highly reliable observations are available, whereas more exploration may be needed where poor data are given. A typical candidate for the application of this approach in our scenario would be the initialization of a grasping sequence with respect to the approach direction and hand pre-shape based on visual observation of a human hand, which can be obtained by earlier developed hand and fingertip recognition methods (see Sec. 5 and [27,29]).

Finally, the third **sensorimotor loop** is responsible for actually carrying out those action candidates that have been identified as most promising by the other two components. Since the internal simulation model and its observation-based refinement can always only approximate the actual situation, actual execution of the action is faced with two alternative evaluation criteria: (i) a maximization of knowledge gain with the consequence of "risky" actions, for instance near decision boundaries; (ii) a maximization of robustness with the consequence of conservative actions in maximal distance to decision boundaries, leading only to minimal information gain.

Fig. 4 shows the interaction of the different loops. The *observation loop* acquires example actions from the human instructor, the *simulation loop* provides refinements of the observed strategies and adaptation to the constraints imposed by the robot system. This changes robot learning into an interactive situated learning process, which uses speech and the multi-modal perceptual channels for an effective optimization of the system's exploration.

5. Towards Imitation Grasping: Observation, Simulation, and Control of Hand Posture

While we have not yet a full implementation of the described architecture, we can report an initial implementation of some of its major features for the scenario of situated learning of grasping of common everyday objects, such as depicted in Fig. 8. In this scenario, the observation component is a vision module permitting observation and 3D-identification of a human hand posture indicating to the robot a sample of the to-be-used grasp type. The identified hand posture is transformed to the joint angle space of the robot hand and is used at the same time as an initial posture for the physics-based simulation of a corresponding grasp as shown in Fig. 5.

The **hand posture recognition** uses a system for visual recognition of arbitrary hand postures which was previously developed in our group. It works in real time, however, currently is restricted to a predefined viewing area. For a more detailed description of the underlying multi-stage hierarchy of neural networks which first detect the fingertips in the hand image and then reconstruct

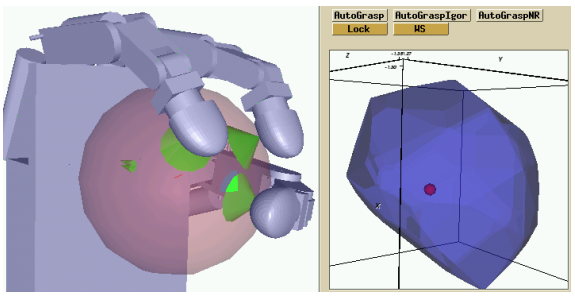


Figure 5. Grasp evaluation in physics-based simulation using contact friction cones (left) and the force closure polytope (lower right).

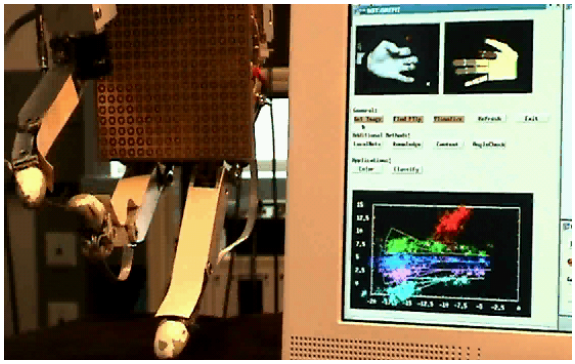


Figure 6. Observation and recognition of continuous human hand postures to obtain an initial pre-grasp posture for autonomous grasping of the anthropomorphic robot hand. In the upper right part the observed human hand is shown on the screen together with the reconstructed hand posture visualized by the rendered computer hand. Below is shown the operation of the PSOM network which obtains the inverse transform from fingertip positions found from observation in the 2D-images to the joint angles of the hand. To the left the resulting TUM-hand posture can be seen.

the joint angles from the fingertip positions see [27,29]. In the context of the present paper, it is a good example for a neurally inspired processing module, which has been refined in an evolutionary way over a number of years and employs learning mainly at the ontogenetic and, to a minor extent, also at the refinement level. Image locations of the fingertips are identified by a two-level hierarchy of several neural networks trained during an off-line phase. A further processing stage, employing a *Parameterized Self-organizing Map* (PSOM, [41]), transforms the obtained 2D-features (fingertip locations) into the joint angles of an articulated hand model approximating the geometry of a five-fingered human hand. Because the PSOM replaces the discrete lattice of the SOM with a continuous manifold and allows to retrieve the inverse of each learned mapping automatically, it can be trained with data of the analytically computable forward transform from the 3D-posture to the 2D-locations.

However, to **actuate the robot-hand**, the reconstructed joint angles cannot be used directly, because the robot hand shown in Fig. 6 has three fingers only and differs in palm and finger sizes, proportions of the phalanges, and the arrangement of the fingers at the palm. Additionally, the sensory equipment and control concepts differ, such that we first have to transform the observed human joint angles into an internal perspective to obtain a posture of the robot hand that is approximately equivalent with respect to the functional properties of the grasp. Geometrically this transformation maps the different joint angle workspaces and reflects the kinematic constraints imposed by the coupled joints of the robot hand. Further we lack direct joint angle measurements in the TUM-hand and therefore rely grasping on force feedback obtained from the custom built fingertip sensors and a recently added palm sensor (see Fig. 6). With the latter we can evaluate the shape of an object when carrying out a power grasp, while the fingertip sensors are primarily used for precision grasps.

From the perspective of imitation learning, these incompatibilities between a human and our robot hand are nothing but the manifestation of the gap between the characteristics of the observed system and the imitator's own sensorimotor equipment plus the different sensory views of the situation. In the present system, the observation component contributes to closing that gap by providing a good initial hand posture from which the robot grasping is started.

This initial posture can be used by the **internal simulation** to generate a grasping sequence. The simulation utilizes the real-time dynamics-based package Vortex [39], which allows accurate object motion and interaction based on Newtonian physics. We extended the package to provide static and dynamic friction for contacts, which is crucial for successful grasp simulation. To generate the finger trajectories, we use an algorithm that attempts to confine the object by incrementally flexing the fingers in a cage-like fashion, evaluating contact conditions on the way. Although contacts are simulated on the basis of point contacts and thus are necessarily coarse, they provide full force feedback, which is not available with our real world tactile fingertip and palm sensors. When all further finger movements have become

blocked by object contacts, the grasp is considered as complete and evaluated according to a quality function [5], which we evaluate by numerical solution of Linear Matrix Inequalities as recently suggested in [15]. Fig. 5 illustrates the friction cones of a successful grasp together with the resulting polytope of applicable forces. One of the next steps will be to use this information in order to conduct a search for an improved initial condition before the grasp sequence is actually carried out with the real hand.

6. Results for Imitation Based Grasping

Before implementation of the imitation grasping subsystem described above, our system had to use pre-programmed associations between known objects and suitable grasps that had been "hand-tuned" for a limited range of objects in rather labor-intensive experiments. From this work, we also knew that our artificial hand – despite its serious limitations – can grasp a large number of real world objects. However, due to the enormous range of possible shapes, generalizing pre-programmed grasps to new and general objects is a rather hard task. Therefore we tried two imitation strategies with our system: (i) a "naive" imitation strategy, in which the observed joint angle trajectories (after their transformation into the three-finger geometry) were directly applied to control the fingers of the TUM hand during the grasp, until complete closure around the object; (ii) a strategy in which the visually observed hand posture is matched to the initial conditions of a power grip, a precision grip, a three-finger and two-finger grip, respectively, in order to identify the grip type. Then, using the initial condition for that grip type, the closing of the fingers takes place autonomously using the same algorithm as employed in the simulation, evaluating tactile feedback from the fingertips to sense stable object contact (Fig. 7).

Success with the "naive" strategy (i) was very limited such that a quantitative analysis was not even worthwhile. This reflects the fact that a purely visual servoing is hardly appropriate for successful grasping, which has to take into account haptic feedback as well. It underlines that

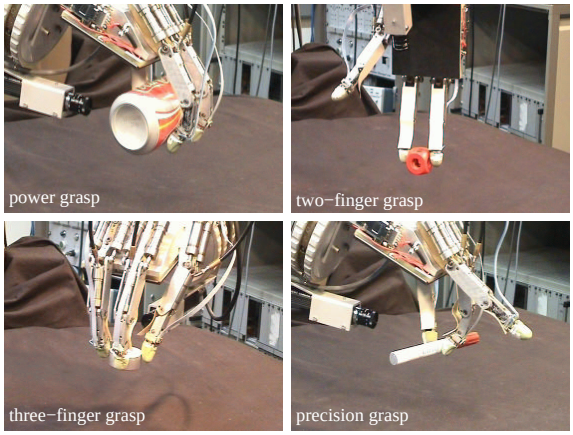
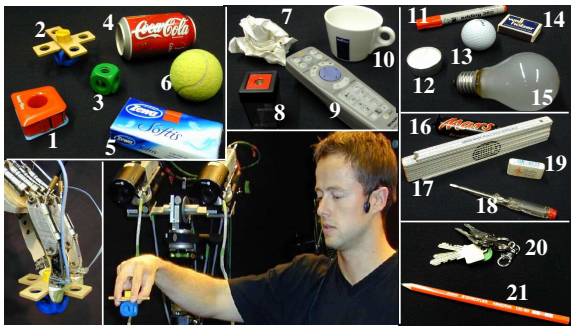


Figure 7. Prototypic grasps with two or three fingers.

in such cases a mixed strategy is required, using information from visual observation as a useful constraint for an action sequence guided to a significant extent under proprioceptive feedback. Indeed, and in line with this expectation, strategy (ii) yielded by far better results, permitting successful grasping of many previously unknown objects. Here, the human grasping gesture instructs the robot to select the most appropriate grasp type for the given object and its orientation on the desk. The Table in Fig. 8 shows the results of a quantitative experiment in which 21 objects were repeatedly grasped (10 trials for each object) under setting (ii) (numeric entries showing the number of successful trials for the particular grasp type/object pairing, while dashes indicate infeasible pairings).

The evaluation of the experiment reveals some interesting details. Six objects can successfully be grasped in all of 10 trials if the most suitable grasp type is used. An additional object, the wooden "propeller" (object no. 2 in Fig. 8), is of a particularly complex shape and cannot be grasped by any of the aforementioned grasp types. However, successful grasping of this object is possible with a specialized grasp derived from the three-finger grasp (see lower left corner image of Fig. 8). The last two objects (keys, no. 20, and pencil, no. 21) are very close to the limits of the hardware capabilities of our hand and can



no.	power grasp	prec. grasp	three finger	two finger	grasp stability
1	10	-	+	-	+
2	requires specialized grasp				
3	(+)	-	+	10	+
4	10	-	-	-	0
5	10	-	+	-	+
6	10	-	+	-	+
7	9	-	(+)	-	+
8	+	(+)	8	(+)	+
9	8	-	-	-	5
10	9	-	-	-	+
11	-	7	-	-	5
12	-	-	6	-	+
13	7	-	-	-	+
14	+	(+)	7	(+)	+
15	6	-	(+)	-	4
16	-	-	5	-	4
17	-	4	-	-	3
18	-	3	-	-	2
19	-	4	-	-	+
20	-	-	0	-	0
21	-	0	-	-	0

Figure 8. Imitation based grasping of everyday objects sorted with respect to the number of successful trials (10 to 0) out of 10 grasp attempts, using the most suitable strategy. The remaining standard grasps are indicated as "+": also possible, "(+)": possible but with less chances of success, "-": not possible. The propeller (no. 2) needs a specialized grasp. The final column gives the number of trials, which are robust against rotation of the hand after lifting up ("+": robust in all trials).

currently not be grasped at all, remaining as a challenge for further improvements.

Further experiments revealed that often small changes in the pre-grasp posture have a large impact on the grasping success. Therefore our next step will be to exploit the simulation stage for optimizing the initial condition before actually executing the grasp, as indicated in Sec. 5. Possible free parameters to be evaluated in such a simulation are the exact initial joint angles of the fingers, the exact relative position of the hand to the object, and the closing speeds of the fingers.

7. A Datamining Perspective on Robot Learning

Regarding learning as a central ingredient to facilitate the construction of complex systems shifts our view from a complex robot whose behavior unfolds according to well-chosen, explicitly designed control mechanisms to a view in which a robot much more resembles a kind of "datamining engine", foraging flexibly for information and regularities in the sensory images of its environment. This suggests to adopt a similar perspective as in the field of datamining, and to exploit algorithms from that area, which appear of considerable interest for advanced robots with their need to cope with uncertainty and situations too complex to be amenable to full analysis on the basis of "first principles".

Specifically, we think that recent progress in content-based image database indexing and retrieval [31,33] may have much to offer for learning robots. Future intelligent robots should be endowed with some kind of episodic long-term memory to accumulate visual and other sensory data. Using raw sensor images for this purpose has many attractive features, provided we can solve the task of efficiently indexing into and navigating within large collections of such data [21]. Unlike symbolic scene descriptions, raw sensor images are easy to acquire and collect. Moreover, they do not enforce a prior commitment onto a narrow range of possible future queries but remain "open" to inspection under aspects that may be unforeseeable at the time of their acquisition. This flexibility may be one of the reasons why also the memory system in our brain offers an apparently visually organized interface to our episodic memory.

In fact, recent progress in semantically organizing large image collections with machine learning techniques for unsupervised category formation [10] and automatic labeling with classifiers previously trained on a variety of visual object domains [43] (so that human keyword assignment becomes dispensable) can be seen as the first promising steps towards the self-organized structuring of a larger body of sensory experience for an artificial cognitive system. Obviously, any progress along these lines is of immediate significance for robotics also, a field which may motivate us to extend such approaches to additional modalities, e.g., collecting and organizing a database of haptic experiences for dextrous object manipulation. Such "life-long" learning may turn out to be the only viable solution to acquire the huge mass of world knowledge that is required for even moderately "intelligent" behavior.

A very important prerequisite in this respect is a system for data collection, diagnosis, and monitoring. Our system currently employs more than 30 distributed processes with many functional submodules such that tracing their behavior, the collection of sensor data, and the detection of errors becomes a non-trivial task. To better cope with this challenge, we have combined adaptive visualization techniques with the rather recent approach of sonification in order to convey an informative, yet intuitive multi-modal display [16] (AVDisp – Audio Visual Display, Fig. 9). This application collects computational results, useful status information, and data from all distributed processes, tags all these messages with time-stamps, and displays the state of the overall system and the interactions of its functional modules. It enables us to build up a database of all relevant information describing the system behavior and to carry out an analysis of the time-behavior of the complete system (like shown in Fig. 3). We found that this interface, originally inspired from the datamining perspective, was already very helpful in the final debugging and tuning phases of the complex system on the engineering level and now provides a valuable basis for implementing more sophisticated learning capabilities of our system.

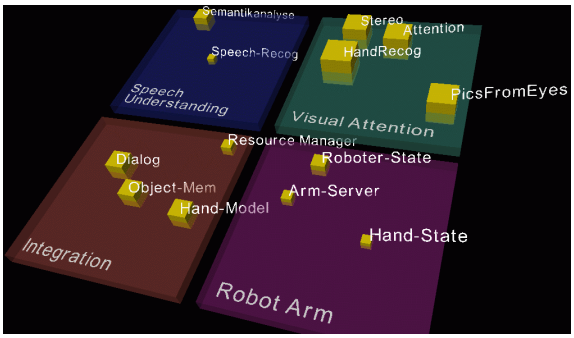


Figure 9. AVDisp: an Audio-Visual Display for system monitoring, data collection, and generation of user feedback (see [16]).

8. Conclusions and Outlook

Our initial assumption is that situated and multi-modal interaction is a key prerequisite for learning in artificial intelligent perception-action systems. Thus, we will proceed with the development of the current platform and use it as a basis for a systematic refinement of the described learning architecture. The longer-term goal is to demonstrate speech enabled imitation learning for instructing grasping tasks, because multi-fingered grasping combines many of the highly developed capabilities of the human cognitive system: the recognition of object shape, type, position and orientation in space; the respective and for the intended task appropriate choice of a grasp; the actual realization of the grasp under complex kinematic constraints; and the following immediate optimization of finger positions and contact force with respect to grasp stability and manipulability.

We believe that this research program is promising if a sufficiently developed technological basis is available. This basis seems crucial for higher level architectures and includes sophisticated hardware for data acquisition and action like an articulated dextrous hand as well as algorithms for robust implementation of the perceptual skills. In particular for the imitation of grasps, we expect in the nearer future progress from improvements in the field of multi-fingered hands, especially with respect to robustness and tactile sensing. Concerning intelligent control it

is important to have at our disposal a sufficiently high number of robust and adaptive partial skills, a prerequisite toward which many efforts have been made in the course of the Special Collaborative Research Unit SFB 360.

The key towards an integrated architecture is a design, which endows the system with a fruitful interrelation of different aspects of learning and their various techniques on the different levels to generate a flexible and incrementally improving combination of these partial skills and modules. Here we see the main focus of the described learning architecture, knowing that this goal may be reached only by long term efforts and in incremental steps. We are aware, that in view of the enormous complexity of the respective challenges, this research program also calls for a close collaboration of robotics with neighboring disciplines like neurobiology or cognitive science and we expect many insights and inspirations from these fields.

Acknowledgments:

Among many people who contributed to the robot system, we thank in particular G. Fink, J. Fritsch, G. Heidemann, T. Hermann, J. Jockusch, N. Jungclaus, F. Lömker, P. McGuire, R. Rae, G. Sagerer, S. Wrede, S. Wachsmuth, J. Walter. For further contributions of the SFB 360 "Situated Artificial Communicators" and the neuroinformatics and practical informatics groups at the Faculty of Technology of the Bielefeld University see the references.

REFERENCES

1. P. K. Allen, A. Miller, P. Y. Oh, and B. Leibowitz, "Integration of vision, force and tactile sensing for grasping,", *Int. J. Intelligent Machines*, vol. 4, no. 1, pp. 129–149, 1999.
2. P. Andry, P. Gaussier, S. Moga, J. P. Banquet, and J. Nadel, "Learning and communication via imitation: An autonomous robot perspective," *IEEE SMC*, vol. 31, pp. 431–442, 2001.
3. P. Bakker and Y. Kuniyoshi, "Robot see, robot do : An overview of robot imitation," in *Proc. AISB Workshop on Learning in Robots and Animals*, Brighton, pp. 3–11, 1996.
4. C. Bauckhage, G. A. Fink, J. Fritsch, F. Kummert, F. Lömker, G. Sagerer, and S. Wachsmuth, "An Integrated System for Cooperative Man-Machine Interaction," in *IEEE Int. Symp. on Comp. Int. in Robotics and Automation*, pp. 328–333, 2001.
5. A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," in *Proc. ICRA*, pp. 348–353, 2000.
6. A. Billard and M. J. Mataric, "A biologically inspired robotic model for learning by imitation," in *Proc. 4. Int. Conf. on Autonomous agents*, Barcelona, Spain, 2000.
7. H. Brandt-Pook, G. A. Fink, S. Wachsmuth, and G. Sagerer, "Integrated recognition and interpretation of speech for a construction task domain," in *Proc. Int. Conf. on Human-Computer Interaction*, vol. 1, pp. 550–554, 1999.
8. C. Breazeal and B. Scassellati, "Challenges in building robots that imitate people," in *Imitation in Animals and Artifacts*, K. Dautenhahn and C. Nehaniv, Eds. MIT Press, 2002.
9. C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc.IJCAI*, pp. 1146–1151, 1999.
10. Y. Chen, J. Z. Wang, and R. Krovetz, "An unsupervised learning approach to content-based image retrieval,", in *Proc. IEEE Int. Symp. Signal Processing and its Applications*, Paris, France, 2003.
11. K. Doya, "What are the computations of cerebellum, the basal ganglia, and the cerebral cortex ?," *Neural Networks*, vol. 12, pp. 961-974, 1999.
12. Joseph A. Driscoll, Richard Alan Peters II, and Kyle R. Cave, "A visual attention network for a humanoid robot," in *Proc. IROS*, Victoria, B.C., 1998.
13. G. A. Fink, "Developing HMM-based recognizers with ESMERALDA," in *LN in AI*, V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, Eds., Berlin, vol. 1692, pp. 229–234, 1999.
14. G.A. Fink, N. Jungclaus, H. Ritter, and

G. Sagerer, "A communication framework for heterogeneous distributed pattern analysis," in *Int. Conf. on Algorithms and Architectures for Parallel Processing*, Brisbane, pp. 881–890, 1995.

15. L. Han and J. C. Trinkle and Z. X. Li, "Grasp Analysis as Linear Matrix Inequality Problems," *IEEE. Trans. on Robotics and Automation*, vol. 16, no. 6, pp. 663–673, 2000.
16. T. Hermann, and C. Niehus, and H. Ritter, "Interactive visualization and sonification for monitoring complex processes," in *Proc. of the Int. Conf. on Auditory Display*, pages 247–250, Boston MA USA, 2003.
17. Ch. Borst, M. Fischer, and G. Hirzinger, "Calculating hand configurations for precision and pinch grasps," in *Proc. IROS*, pages 1553–1559, Lausanne, 2002.
18. G. Heidemann, D. Lücke, and H. Ritter, "A system for various visual classification tasks based on neural networks," in *Proc. ICPR*, Barcelona, A. Sanfeliu et al., Ed., pp. 9–12, 2000.
19. A. J. Ijspeert, J. Nakanishi, and S. Schaal, "Trajectory formation for imitation with nonlinear dynamical systems," in *Proc. IROS*, pp. 752–757, 2001.
20. N. Jungclaus, R. Rae, and H. Ritter, "An integrated system for advanced human-computer interaction," in *UCSB-Workshop on Signals and Images*, USA, pp. 93–97, 1998.
21. T. Kämpfe, T. Käster, M. Pfeiffer, H. Ritter, and G. Sagerer. "INDI – Intelligent Database Navigation by Interactive and Intuitive Content-Based Image Retrieval," in *IEEE 2002 International Conference on Image Processing*, volume III, pages 921–924, Rochester, USA, 2002.
22. R. Dillmann, M. Kaiser, and A. Ude, "Acquisition of elementary robot skills from human demonstration," in *In Int. Symp. on Intelligent Robotic Systems*, Pisa, Italy, pp. 185–192, 1995.
23. Y. Kuniyoshi, M. Inaba, and H. Inoue, "Learning by watching: extracting reusable task knowledge from visual observation of human performance," *IEEE. Trans. on Robotics and Automation*, vol. 10, no. 6, pp. 799–822, 1994.
24. M. J. Mataric, O. C. Jenkins, A. Fod, and V. Zordan, "Control and imitation in humanoids," in *AAAI Fall Symposium on Simulating Human Agents*, North Falmouth, MA, 2000.
25. P. McGuire, F. Fritsch, J. J. Steil, F. Röthling, G. A. Fink, S. Wachsmuth, G. Sagerer, and H. Ritter, "Multi-modal human-machine communication for instructing robot grasping tasks," in *Proc. IROS*, pp. 1082–1089, 2002.
26. P. McKevitt, Ed., *Integration of natural language and vision processing*, Kluver, Dordrecht, 1994.
27. C. Nölker and H. Ritter "Visual Recognition of Continuous Hand Postures," *IEEE Trans. NN*, vol. 13, no. 4, pp. 983–994, 2002.
28. J. Pauli, "Learning to recognize and grasp objects," *Autonomous Robots*, vol. 5, pp. 407–420, 1998.
29. H. Ritter, J. J. Steil, C. Nölker, F. Röthling, and P. McGuire, "Neural Architectures for Robot Intelligence," *Rev. Neurosci.*, vol. 14, no. 1-2, pp. 121-143, 2003.
30. B. Roessler, J. Zhang, and M. Hoechsmann, "Visual Guided Grasping and Generalization Using Self-Valuing Learning," in *Proc. IROS*, pp. 944–949, 2002.
31. Y. Rui, T. Huang, and S. Chang, "Image retrieval: current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, 10(4):39–62, 1999.
32. Stefan Schaal, "Is imitation learning the route to humanoid robots ?," *Trends in Cognitive Sciences*, vol. 3, no. 6, pp. 233–242, 1999.
33. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.
34. J. J. Steil, G. Heidemann, J. Jockusch, R.Rae, N. Jungclaus, and H. Ritter, "Guiding attention for grasping tasks by gestural instruction: The GRAVIS-robot architecture," in *Proc. IROS 2001*. IEEE, pp. 1570–1577, 2001.

35. Joseph O'Sullivan, "Towards a robot learning architecture," in *Learning Action Models* , Wei-Min Shen, Ed., 1993, pp. 47–51, AAAI Press, 1993.
36. J. Lloyd and M. Parker, "Real time control under Unix for RCCL," in *Robotics and Manufacturing ISRAM'90*, vol. 3, pp. 237–242, 1990.
37. R. Menzel, K. Woelfl, and F. Pfeiffer, "The development of a hydraulic hand," in *Proc. 2. Conf. Mechatronics and Robotics*, pp. 225–238, 1993.
38. S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proc IROS*, pp. 2332–2337, 2001.
39. CMLabs "Vortex – physics engine for real-time simulation,", www.cm-labs.com
40. S. Wachsmuth and G. Sagerer, "Bayesian Networks for Speech and Image Integration," in *Proc. of 18th National Conf. on Artificial Intelligence*, Edmonton, pp. 300–306, 2002.
41. J. Walter, C. Nölker, and H. Ritter, "The PSOM Algorithm and Applications," in *Proc. Symposion Neural Computation (Berlin)*, pp. 758–764, 2000.
42. H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant recognition," *Neural Computation*, vol. 15(7), pp. 1559-1588, 2003.
43. J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE PAMI*, 23:947–963, 2001.
44. M. Yeasin and S. Chaudhuri, "Toward Automatic robot programming: Learning Human Skill from Visual Data," *IEEE SMC*, vol. 30, no. 1, pp. 180–185, 2000.