ADAPT

IST–2001-37173

Artificial Development Approach to Presence Technologies

Deliverable Item 1.9
Final report and management report

Delivery Date: due date: September 30th, 2005
Classification: Public
Responsible Person: Giorgio Metta – DIST
Partners Contributed: ALL

Short Description:
Contributors:

DIST: Giorgio Metta, Giulio Sandini, Riccardo Manzotti, Francesco Orabona, Carlos Beltran, Fabio Berton
CNRS: Jacqueline Nadel, Arlette Streri
UNIZH: Harri Valpola, Martin Krafft, Gabriel Gomez, Rolf Pfeifer

This document describes the final achievements of the project and contains a brief description of the deliverables, the structure of the experiments, the conclusions and future perspectives. Certain parts of this manuscript are derived and/or the same already prepared for other deliverables.

Contract started: October 1st, 2002
Contract duration: 36 months

This was prepared after the final review meeting on September 22nd, 2005 in London.

Project funded by the European Community under the "Information Society Technologies" Programme (1998-2002)

Summary

1 Executive summary

Adapt deals with a very basic question about the sense of presence: that is, how do we represent our world and, in particular, how do we represent our world of objects and people? There are two basic facts about this: we can ask first what a representation* is, encompassing in the answer quite a wide range of different disciplines, and second, how this representation can be used to reproduce the sense of presence in a human being. We chose (see the Technical Annex) to study mostly the first question and, as such, we are not going to work on the construction of any virtual reality device. Conversely, we are investigating on one side how representations are built by the brain during ontogenesis and, in parallel, how a model of this process can be reproduced in a robotic artifact. The reasons being that the study of development can provide precious hints that the study of adults can not, while, following the so-called synthetic methodology†, we aim at producing a working model of a similar process allowing a robot to acquire representations through the interaction with the environment.

1.1 Consortium

The consortium consists of three partners. The following table shows their main role within the project.

Partner	Role in the project
DIST - LIRA-Lab University of Genova, Italy	Coordinator: development of a humanoid robotic platform, theory of representation and intentionality, integration, manipulation
AI-Lab, Dept. of Information Technology University of Zurich, Switzerland	AI: Contribution to the definition of the developmental paradigm, study of the role of morphology in manipulation, development of multisensory features
UMR7593, CNRS, University Pierre & Marie Curie, Paris, France	Developmental psychology: definition and implementation of the behavioral experiments, comprehensive study on the development of certain representations in the brain

1.2 Division of work and workpackages

Adapt plan sees four technical and one management workpackage. The following table contains a summary of the organization of the work and the proposed experimental path. This table tries to present a clear-cut view of the different lines of investigation although the actual

* Representation: not to be confused with the classical view of classical AI (representation and symbol manipulation as in Newell and Simon approach).

† Synthetic methodology: it has been proposed that building robotic artifacts might be a useful endeavor to understand the extent and conditions of validity of models of the physiology (the functioning) of biological agents.

implementation is not so well defined and the structure of the deliverables, in fact, reflects this intermixing of different disciplines.

WP1	Management
WP2	Theory of consciousness
WP3	Study of embodiment and morphology in the development of representations
WP4	Developmental psychology experiments
WP5	Developmental robotics experiments

1.3 Aims

Adapt contributed to the formulation of a coherent theory of representation, and to the preparation of a set of experiments, both in developmental psychology and robotics, consistent with the theory. The long term goal of this activity is the complete validation of the theory and the understanding of the brain mechanisms responsible for the creation of representations.
The first activity of Adapt was the formulation of a theoretical framework based on psychology and philosophy of mind apt to describe both phenomenology/first person experience and objective/third person views of the world. The main characteristic of the theory is the unitary view of reality where the basic constituents are processes (named *onphenes* later on). This approach implicitly rejects the classical "dualistic" approach (a la Descartes for instance). The rationale is that to understand what constitutes a sense of 'being there', or specifically Presence, it is important to understand what gives rise to this sense of being there. It is argued by some authors that the sense of being-there depends on the capability of having semantic representations. The sense of being there is thus possible only in a *conscious subject* (called simply *subject* in the following). The sense of being there is the unified collection of phenomenal experiences that would be experienced by a subject in that particular place. In order to have these phenomenal experiences it is necessary to have the right sensory and motor capabilities plus something else, namely, the capability of having a phenomenal experience out of the right sensory motor contingencies. This something else can be equated with intentionality[‡] or phenomenal experience (conscious experience).

Unfortunately, the problem of consciousness or conscious experience is traditionally conceived as the impossible task of justifying the emergence of an inner world of experiences, qualia and/or mental representations out of a substratum of physical things believed to be autonomously existing. To solve the impasse, we argued that an alternative approach is possible but it requires a conceptual reconstruction of consciousness and existence, the two being different perspectives on the same underlying process. On this basis, we have presented a view of direct (conscious) perception that supposes the unity between the activity of the brain and the events of the external world. The outlined process is here referred to as the *onphene* and the theory itself "the Theory of the Enlarged Mind (TEM)". We will use an example later to introduce the new perspective, but eventually, the same approach can be used to explain other kinds of consciousness: illusions, memory, dreams, and phosphenes. The view presented

[‡] In the sense of "aboutness".

here shares some elements with neo realism and can be considered as a form of radical externalism.

In order to make our rationale as clear as possible, we emphasize a principle that we applied consistently throughout the formulation of the theory: *something, in order to exist, has to produce an effect*. It is a principle that has been advocated by most of Galileo's epistemological descendants. If something does not produce an effect, then there is no need for a difference between its existence and its absence: in either case the consequences of the existence or non-existence of this something would be the same. Here, we are not going to enter into the debate of the existence of logical entities and abstract concepts. We will keep our feet on the ground of physical reality. Ether, phlogiston, and epicycles were dismissed as being incapable of making a real difference. We will use this principle to show that the separation between the subject and the object is unfounded.

The *rainbow* is perhaps an example in which the separation between the observed object/event and the observing object/event is not evident. When the sun is sufficiently low on the horizon and projects its rays at an appropriate angle against a cloud with a large enough volume of drops of water suspended in the atmosphere, an observer can see the arch with all the colors of the spectrum. All drops of water reflect the sunlight in the same manner, yet only those which have a particular geometrical relation to the observer, due to his position, and due to the orientation of the rays of light, are seen as part of the rainbow. The position of the observed rainbow thus depends on the position of the observer.
An important caveat is required here: by observer of a phenomenon, we do not refer to a human being, or a conscious subject, or an agent with a mind. We refer to a physical system that is capable of "recognizing" an occurrence of that phenomenon. By recognizing we refer to the capability of selectively producing an outcome of some kind in response to the presence of that phenomenon. For instance, an observer of a rainbow is a system which can produce an outcome whenever it is in front of a rainbow. According to this definition, which has no pretension of being used outside of the scope of this manuscript, a digital or film camera is not an observer since it records all the visual information without being able to recognize explicitly anything. On the contrary, a human being, most animals, artificial pattern recognition systems are observers.
Let us consider a simplified version of a rainbow as that shown in Figure 1: a one-dimensional column of drops of water that float in the air. A stream of parallel white rays of light collides with them. As a result, each drop reflects a divergent stream of colored rays of light. Is there a unity? A whole? No. The rainbow as a unity, as a whole, is not there yet. Nevertheless, as soon as an observer selects a given combination of drops, then a rainbow takes place. If no observer were there, would the rays produce an effect as a whole? No, they would not, because they would continue their travel in space without interacting and eventually they would spread everywhere. Their opportunity to produce a joint effect would be lost. As William James wrote "In the parallelogram of forces, the 'forces' themselves do not combine into the diagonal resultant; a *body* is needed on which they may impinge, to exhibit their resultant effect" (James, 1890/1950), p. 159. Therefore their cause (the supposed rainbow) would not have produced any effect and as a cause would not have existed. It was only a theoretical existence. We assumed that there must have been a rainbow, but there wasn't. On the contrary, if an observer were there, the converging rays of light would have hit his/her photoreceptors and a

fast but complex chain of physical processes would have continued from the retina to the cortical areas up to a point where the recognition of the rainbow as a whole (a colored arch) would have taken place. Thanks to the existence of the physical structure of the observer, the drops of water of the rainbow have been able to produce a joint effect. As it is shown in Figure 1a, until the whole process is concluded, there is no actual rainbow as a whole. Something could happen at the very last moment interfering with the completion of the process. There are two possible outcomes (Figure 1b and Figure 1c). In the former, a perceiver is missing and the rays of light would loose their chance to produce an effect as a whole. In the latter, an observer allows the rainbow to take place as a process and as a whole. The cause seems to exist only thanks to the occurrence of its effect (Figure 1d): the cause of the cause is the effect and the effect of the effect is the cause. This is paradoxical. Yet the paradox disappears once we conceive the unity of the underlying process (Figure 1e). If we apply the principle, mentioned at the beginning of this paragraph, we must conclude that the rainbow remains a possibility, an abstraction, until the rays of light interact with something with the proper capability.

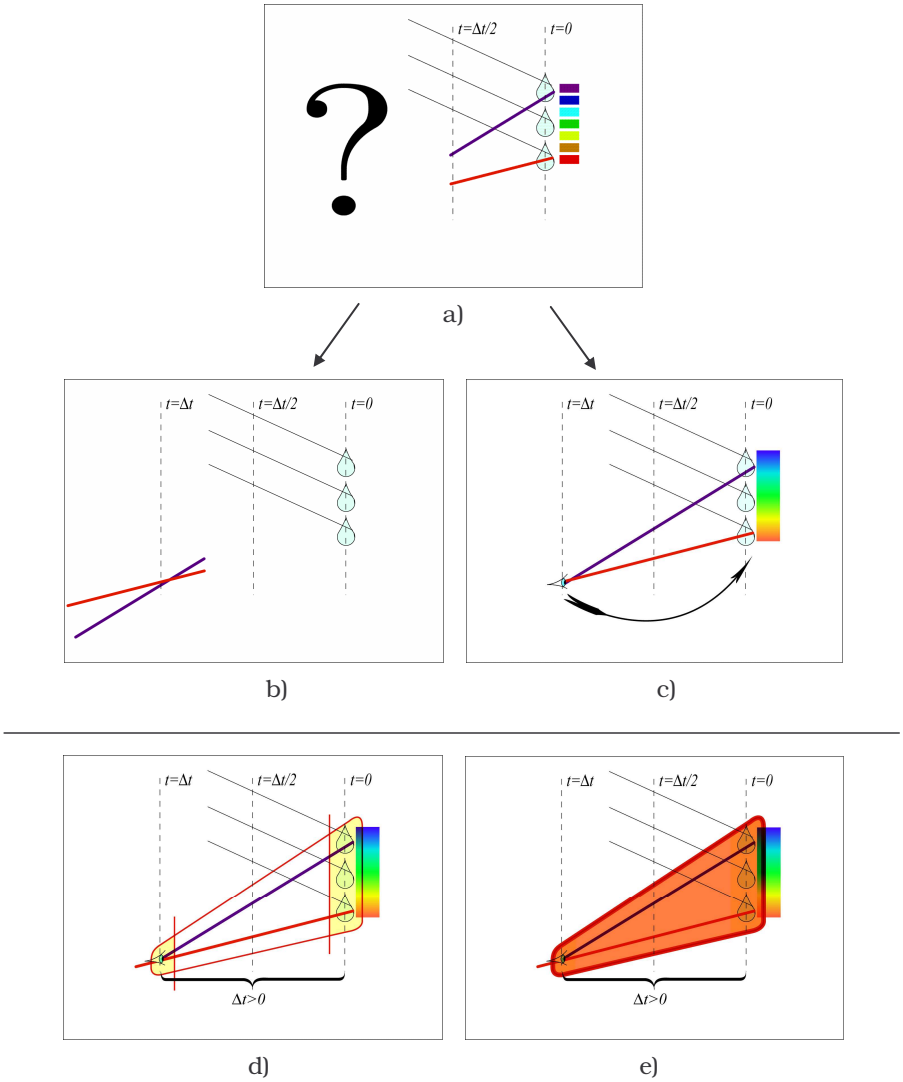


Figure 1: A rainbow (see text).

Thus the answer to the question: "Does the rainbow exist independently of the act of observation?" is obviously "No". Even from a logical point of view, to define the position of a rainbow, an expert physicist would need to know the precise point of view of the observer. Thus the rainbow is not a thing: it is *a process*, in which there is an entanglement between a physical complex and an observer. The drops of water do not constitute a distinctive whole (the rainbow) unless and until they produce an effect. The point is that the effect cannot be split from the cause, nor can the cause and the effect be split from their relation. The effect is responsible for the existence of the cause. Further, the existence of the rainbow depends not only on the presence of the physical conditions given above and the observer, but on a causal continuity between the two. This continuity consists of rays of light at the right location actually hitting the retina of the observer and setting up a continual discharge in the brain, as long as the physical relationships are maintained. Once these physical relationships are broken, the rainbow as – a process and as a whole – ceases to exist.
In the cloud there are almost infinite possible rainbows. Yet only a very limited number of them are actually able to produce an effect as a whole: those that are interacting with the proper kind of physical systems (normally human visual systems).
The concept of a "possible rainbow" is misleading because it entails the existence of something, while it would be much more precise something like "some of the conditions necessary to the occurrence of a rainbow". The physical conditions of the drops are only half of the story: the other half is in the observer's eyes and brains. The whole story is the occurrence of the process as a whole (which we call a rainbow).

Let us recapitulate the meaning of the previous example and see how to derive the core of a theory of mind. The traditional standpoint conceives reality as made of relatively autonomous objects or relatively autonomous events. This entails that the subject and the object, being both instantiated by an autonomous set of objects or events, are irremediably separate in time and in space. Therefore the problem of representation, the problem of mental causation and the problem of the ontology of mental events (secondary properties) arise. On the other hand, we – as human beings – do perceive the world not as an image of the world but as the world itself. Realism basically reminds us that our mental states are about the world. Externalism tries to get out of the boundary of the brain. Finally, a process-based ontology could be the tool to sustain both views and to overcome the subject/object dualism. The rainbow and other possible examples try to convey this insight: the world is not made of relatively autonomous events; the world is made of intrinsically related processes. Therefore, the subject and the object are not separate and there is no problem of "re-presentation", since the experience and the occurrence of the world are identical. In the previous example the cause does not exist isolated from its effect. They are both taking place as two different ways of describing a process which cannot be split. Whenever we consider perceptual events (like the perception of a rainbow or a face), we have to admit that the perceived object does not exist in isolation from its perception.

The traditional problems of consciousness are going to vanish once the *onphene* perspective is adopted. The world in which each subject is living is no longer a private bubble of phenomenal experiences concocted by the brain. Each subject is living in and experiencing the real world: the two being different descriptions of the same process. Each subject lives in that part of the world made by those processes with which s/he is identical with. The subject is *those*

processes. In our own experience, consciousness, existence and becoming cannot be split. As agents, we are part of a physical flow of processes that are possible thanks to our physical structure. These processes have the right properties of our own experiences as well as the right properties of the external world. The need for postulating a noumenal world of primary properties (and their bearer, the object) and a symmetrical world of secondary properties (and their bearer, the subject) arose from the undemonstrated Galilean hypothesis of an a-temporal domain of autonomous entities. By using a processual view such need vanishes. Adopting a processual point of view a different framework begins to unveil. Consciousness and existence can be explained as two perspectives on the same processes.

The world of the subject's experience is identical with the real world. It is then possible to simply discard many classical problems related to consciousness. In particular it is possible to discard the television view of the mind (see (Dretske, 1995)). This is not a complete novelty. Other authors criticized in a similar way the idea that what we experience is an image, possibly internally generated, of the external world. For instance James Gibson wrote that (James J Gibson, 1952): "The visual field, I think, is simply the pictorial mode of visual perception, and it depends in the last analysis not on conditions of stimulation but on conditions of attitude. The visual field is the product of the chronic habit of the civilized men of seeing the world as a picture … So far from being the basis, it is a kind of alternative to ordinary perception." In a strikingly similar way, the art historian Jonathan Crary wrote that "The idea of subjective vision – the notion that our perceptual and sensory experience depends less on the nature of an external stimulus than on the composition and functioning of our sensory apparatus – was one of the conditions for a severing of perceptual experience from a necessary relation to an exterior world." (Crary, 1992), p.12. Of course, this historical process of reshaping the observer had its theoretical origin with the work of Galileo. With respect to direct perception this approach has the advantage of solving all of the three classical problems outlined in the previous paragraphs: the hard problem (i.e. the ontology of mental events), epiphenomenalism (i.e. the problem of mental causation) and the problem of representation.

The so-called *hard problem* is solved since there is a candidate for the nature of phenomenal experience: the physical processes engaged between the brain and the external environment. There is no more dualism. The price to pay is to discard the assumption of the separation between the subject and the object as well as the autonomy of the existence of objects. Onphenes are neither objective nor subjective. They are private and public at the same time.

Epiphenomenalism is solved since phenomenal states are no longer separate from the physical world. Every phenomenal state is identical with a physical process that, as all physical processes, has causal powers and exerts its effects on the environment.

The *problem of representation* is solved since there is no more need to re-produce an internal image of the external world. Phenomenal experiences are identical – they coincide – with the aspects of reality they should represent. More precisely, they do not represent reality: they are the reality. The subject does not perceive *an image* of an object: a process takes place which is constitutive both of subject and objects; a process which can be described or as a subjective experience or as an objective event.

The next step in the formulation of the theory is then to introduce the concept of the Enlarged Mind. If every phenomenal experience is identical with a process (an onphene), the sphere of an individual consciousness is identical with a collection of onphenes. If a subject is conscious of a rainbow plus a face plus some speech to which s/he is listening to, then it means that at

least three separate processes are taking place. In reality there are almost countless processes going on in the environment. However, only a subset of them becomes entangled in that flow, which is the conscious experience of a given subject. The mental life of a subject is no longer constrained inside the cranium, compelled to the creation of a theatrical replica of the external world: the mental life is literally enlarged to the processes constituting everything that the mind is conscious of. There is not a mental life and a physical life: there is only one life where the mind is identical with everything the subject is conscious of; everything being a process and not an a-temporal static object. Furthermore, the existence of what the mind is conscious of is possible because of the occurrence of those processes that are identical with the mind itself. This is only apparently paradoxical. For instance, the fact that the subject is conscious of the rainbow as an arch of colors does not entail that the subject is responsible for the existence of the sun and the drops of water. Yet, without the subject's brain, the drops of water would have remained each by its own. No rainbow as a whole would have occurred. Their unity, as a colored arch, is the result of the process occurring. The rainbow, as a unity, does exist thanks to the same process which is identical with the observer.

We consider the physical process that begins in the external world and ends in the brain as a unity since it provides a unique framework for the description of physical reality and mental reality. If the hypothesis proves to be correct, then it is no longer necessary to look for a neural implementation of conscious activity. A conscious mind is the set of processes that have as causes the object of experience and as effects the recognizable events of the cognitive activity. Such causal processes named onphenes (achievable thanks to the particular structure of the brain, to the agent body and to the surrounding environment) constitute the external objects and the internal content of the mind, the two being different ways of describing the same thing. The rainbow is an excellent example of an onphene, in which observation, the observer and the observed entity cannot be split. All occur jointly. They are the same occurrence so this is coherent with the fact that they must constitute a unity. But the example of the rainbow, though a very compelling one, is not unique in leading to this conclusion. We propose that all perceived objects exist insofar as they "take place". The relevance of this argument lies in the fact that the brain is not self-sufficient with respect to mental events. We envisage the brain as the end part of a larger network of physical processes.

In short, clearly, this is all but a complete description of the Theory of the Enlarged Mind, many other aspects should be mentioned and several "classical" aspects (e.g. representation) rephrased within our framework in order to make serious claims about the accordance between theory and empirical evidence. Two deliverables describe these details: D2.1 and D2.2. D2.2 is a somewhat more complete and up to date version of the theory including also a larger set of references. And finally D2.3 contains a small vocabulary of terms with their interpretation within our theory. Nicely enough these terms can be now used unambiguously to describe both AI and psychological concepts.

To recapitulate, the core of the theory proposes that:

- presence is due to a series of phenomenal mental events that are contentful and integrated;
- the intentional and phenomenal status of mental events is due to their identity with physical processes that include the external target of these events;

- these processes have a role in shaping both the environment and the subjective experience; for this reason they have been named onphenes;
- a collection of these processes (or onphenes) constitutes a moment of presence
- the unity between separate onphenes is due to the progressive entanglement of causal processes in order to achieve a goal;
- the final unity of separate onphenes (which possess intentionality in the philosophical sense of aboutness) is eventually achieved by their cooperation to reach a given goal thus obtaining intentionality in the psychological sense.

In conclusion, although we believe the theory of the enlarged mind gives a fair account of a possible solution to the mind/body problem and proposes a solid philosophical framework for both psychology and AI, we are also aware that the presented theory once adopted has vast consequences which are virtually impossible to explore within a single project. Nevertheless, we believe it is worthwhile to attempt a complete description of the theory and the framework derived thereof. The work performed in Adapt and our future work, both experimental and theoretical, will provide the necessary elements for its further development.

1.3.1 Link with experimental work

Inside the various versions of externalism, a very closely related concept to that of the *onphene* is that of James J. Gibson's affordance. Although the concept of affordance is not completely unambiguous (Jones, 2003), Gibson defined an affordance of something as "a specific combination of the properties of its substance and its surfaces taken with reference to an animal" (J.J. Gibson, 1977), p.77. As in the case of the *onphene*, the affordance is neither entirely located in the object nor in the subject. Furthermore, an affordance depends on both terms it relates: on one side "the properties of its substance and surfaces" and on the other side "an animal", which is the observer. Gibson wrote that "[…] an affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is both a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer" (J.J. Gibson, 1979), p. 129. The *onphene* is very similar to an affordance in the sense that both bypass the subject/object distinction.
The concept of affordance is useful since experiment can be designed to study the properties of, for instance, the affordances of objects (see later in this document), in robotics (Arsenio, Fitzpatrick, Kemp, & Metta, 2003), and their neural correlates can be studied in animal (Sakata, Taira, Murata, & Mine, 1995) and human experiments (Fadiga et al., 1999).

1.4 Organization of the experimental work

The next step in Adapt was to derive a set of experiments. They represent the first attempts of validation of the theory on one hand, and the possibility of extending its application to different domains on the other. Clearly, investigating every possible facet was out of question; it is worth noting that these are issues and problems that plagued the last 50 years of artificial intelligence and philosophy of mind. As we mentioned in the introduction, we concentrated on the study of the interaction with "objects" in a rather generalized sense: that is, including both the interaction with objects proper and the interaction with people. These two types of interaction are linked with the theory of *affordances* and *multimodal integration*.

The theory of affordances proposed by J.J. Gibson as we mentioned is sort of the link between the theory of the *onphene* and the definition of a line of empirical research. Multimodal integration is the other side of the same coin. As the affordance is not entirely sensorial, in fact, it is defined in terms of both the external environment and the subject's state, including its motoric skills, which means multimodal integration in practice. In this sense, there is a nice connection with neurophysiology. In the last ten years, evidence has been accumulating showing how the neural sensory responses are deeply intertwined with the motoric ones (Gallese et al. 1996). This carries two messages: i) the embodiment of the agent is fundamental, and ii) the embodiment carries specific motor components that are then embedded into the sensory processing.

The last extremely important aspect to the study of the ontogenesis of representations is that of motivations. Motives are fundamental both for sensorimotor learning and for cognitive development (von Hofsten, 2004). Within Adapt we did not have enough resources to touch also this aspect although it is discussed in deliverable 2.2 in some length. One of our publications shows a possible motivation driven architecture and clearly this will be part of our future research.

The following diagram shows the conceptual link between the various domains in Adapt.

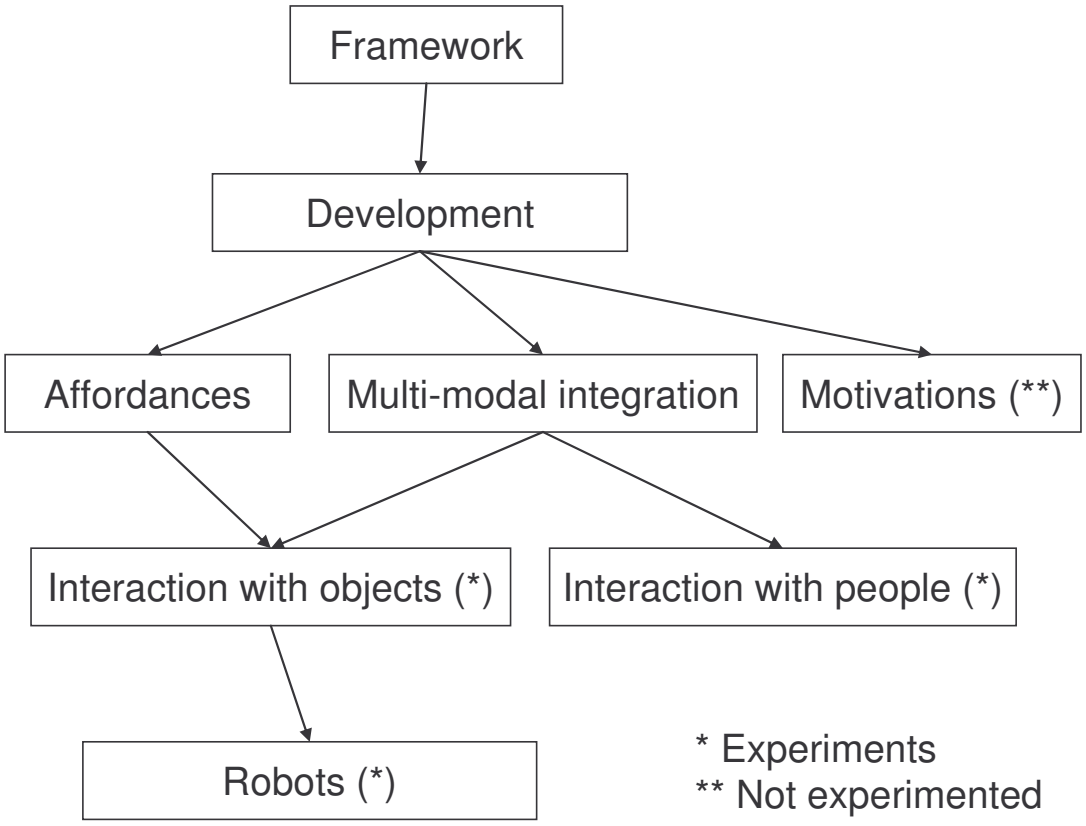


Figure 2: the experimental plan.

The two types of interaction were investigated by means of behavioral experiments with infants ranging from newborns to 12 months of age. These experiments were directed at the core topic of the project: that is, representation. In a first set of experiments the amount of multimodal integration at birth was investigated. Newborns were studied following the classical habituation paradigm. This first set consisted of five different experiments elucidating how different cues (texture vs. shape) are perceived by different sensory channels (touch vs. vision). The second set of experiments was directed at analyzing multimodal integration between audition and vision in 2 and 6 month olds. Also in this case four experiments were prepared investigating different aspects of the detection of delays and contingent speech vs. video. One last experiment started to look at multisensory integration and imitation in the context of object manipulation. In this case 6 and 12 month olds were recorded. At the moment of writing only the 6 month olds data are available. Globally, this experiments show that i) multimodal integration is present from birth (i.e. perhaps due to an amodal feature directly perceivable) and although patchy it forms the core of the representation of objects, and ii) time delay and contingency is also perceived starting at 2-6 month of age, meaning that the temporal integration of features develop quickly to serve, for example, social interaction. These "representation" of other agents presents aspects of multimodal integration but also of turn-taking (contingent behavior) thus containing a more sophisticated seed of a general representation of the environment. The last experiment is devoted to the observation of affordances in six month olds. The results show a high variability. Subjects were classified in three different groups: 1) not able to grasp, 2) grasp without adaptation to object properties, and 3) grasp with adaptation to object properties. This replicates earlier findings showing that by 6 months of age, infants are at a crucial cornerstone in developing and differentiating grasping abilities and affordance perception. The plan here is to analyze further the difference between grasping following a demonstrator in two situations: 1) copying an affordant behavior, and 2) copying a non-affordant behavior. In the latter, a higher degree of abstraction of affordances is required, while the former is thought to be simpler for the infants.

Finally, the project proposed the implementation of a "developing" robotic system. Clearly, the hardware is fixed; it is the software that evolves as it learns from examples. A rough developmental schedule is hypothesized and implemented into the humanoid robot. The robot's hand has been developed and realized within Adapt. The robot has been used for three sets of experiments. The final demonstration sees the robot searching, reaching and grasping an object. For elongated objects the robot also extracts some shape information (the seed of a grasping affordance) and uses it to regulate the orientation of the hand with respect to the object. In the second experiment we show how visuo-auditory processing is possible leading to a multisensory segmentation. In the last experiment, we have used the robot to collect data that are then analyzed off-line through unsupervised learning methods in search for multimodal features. To close the loop, we should have integrated the unsupervised learning with the control system, this as we mention in the management report has not been completed yet. It is worth noting that a good part of the control system of the robot shows learning and adaptation as presented in D4.4.

We have also analyzed grasping actions from an information theory point of view in search for the markers of the robot morphology. This means that the robot's shape and control strategy together determine specific information structures that could be potentially exploited by the

robot learning system. At the moment the technique is mostly used as an analysis tool, the final goal is clearly that of developing a synthesis instrument. See D3.3 for details.

2 Main achievements

Adapt main achievements are scientific including a fair amount of "real-world" robotic and psychological experiments. Generally speaking, Adapt contributed to the formation of an interdisciplinary group of scientists with mutual interest in bringing new approaches and methods to the "field". In particular, we have worked in close contact with developmental psychologists, roboticists, and AI people. Indeed, this can be the only possible route to cross-fertilization between disparate fields: that is, by actually working together in a project. Often, the opposite happens, and researchers are content with reading papers from different disciplines rather than acquiring first hand experience through collaboration. We believe the latter approach to be superior and potentially lead to better results.

The following list summarizes the major results of the project:

- Realization of two robotic hands with different mechanical solutions incorporating passive compliance and underactuation.
- A theory of mind contributing to a new ontological framework for the understanding of the brain (called the Theory of Enlarged Mind).
- A data analysis technique called Denoising Source Separation (semisupervised learning technique).
- An anthropomorphic robotic arm for research on morphology and materials (still under development).
- Implementation of a learning architecture to control reaching and grasping in a humanoid robot.
- Implementation of algorithms for detecting synchronous visual/auditory object features.
- A set of experiments in developmental psychology elucidating some of the aspects of the ontogenesis of representations.

3 Methods

Adapt required the realization of several new experimental apparatus. In particular we developed brand new:

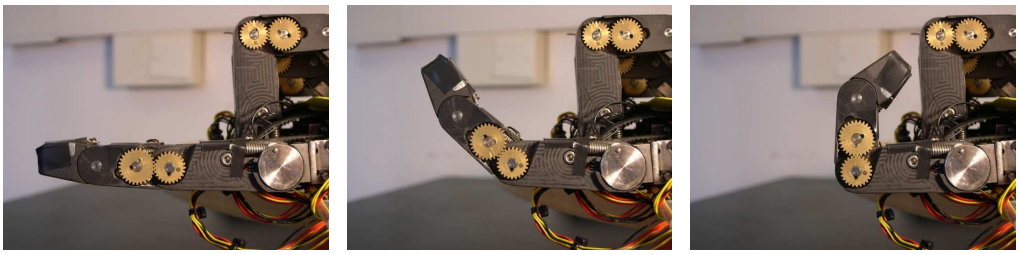
- A six degree of freedom robotic hand with passive compliance.
- A tendon driven robotic hand.
- A new version of the teleprompter device.

3.1 UGDIST robot hand

Each finger has 3 phalanges; the thumb can also rotate toward the palm. Overall the number of degrees of freedom is hence 16. Since for reasons of size and space it is practically impossible to actuate the 16 joints independently, only six motors were mounted in the palm. Two motors control the rotation and the flexion of the thumb. The first and the second phalanx of the index finger can be controlled independently. Middle, ring and little finger are linked mechanically

forming a single virtual finger controlled by the two remaining motors. No motor is connected to the fingertips; they are mechanically coupled to the preceding phalanges in order to naturally bend as explained in Figure 3. The mechanical coupling between gears and links is realized with springs. This has the following advantages:

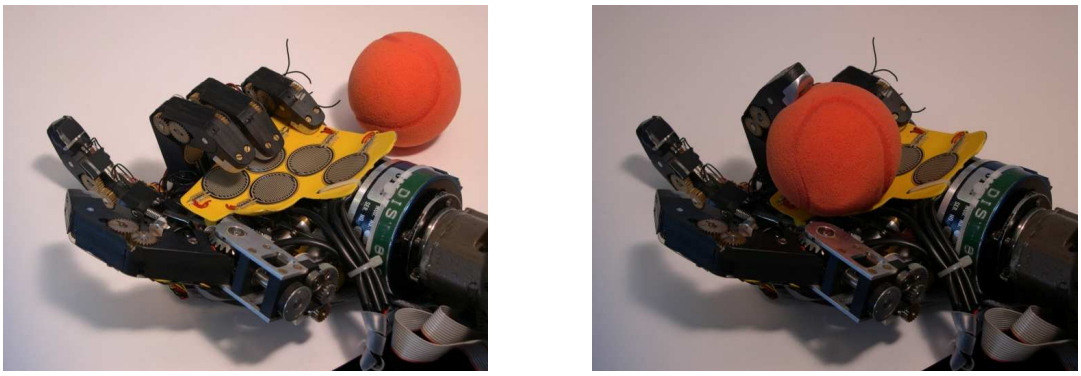
- The mechanical coupling between middle, ring, and small finger is not rigid. The action of the external environment (the object the hand is grasping) can result in different hand postures (see Figure 4).
- Low impedance, intrinsic elasticity. Same motor position results in different hand postures depending on the object being grasped.
- Force control: by measuring the spring displacement it is possible to gauge the force exerted by each joint.



a) b) c)

Figure 3: Mechanical coupling between the second and the third phalanges. The second phalanx of the index finger is directly actuated by a motor. Two gears transmit the motion to the third phalange. The movement is respectively of 90 and 45 degrees.

Hall-effect sensor encoders were employed to measure the strain of the springs coupling the hand's joints. This information together with that provided by the motor optical encoders allows, at least in theory, the estimation of the posture of the hand and the tension at each joint. In addition, force sensing resistor (FSRs) sensors are mounted on the hand to provide tactile feedback. These commercially available sensors exhibit a change in resistance in response to a change of pressure. Although not suitable for precise measurements and prone to failure, their response can be used to detect contact and measure to some extent the force exerted to the object surface. Five sensors have been placed in the palm and three in each finger [apart from the little finger] (see Figure 4).



a) b)

Figure 4: Elastic coupling. a) and b) show two different postures of the hand. Note however that in both cases the position of the motor shafts is the same. In b) the intrinsic compliance of the middle finger allow the hand to adapt to the shape of the object.

Further proprioceptive information is provided to the robot by a strain gauge torque/force sensor mounted at the link between the hand and the manipulator's wrist. This device is a standard JR3 sensor designed specifically for the PUMA arm flange. It can measure forces and torques along three orthogonal axes (see Figure 4).

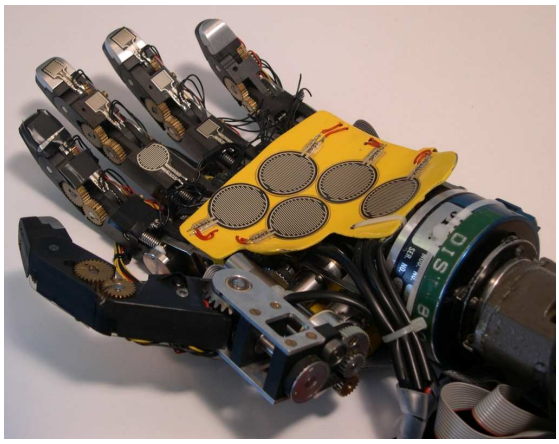


Figure 5: Tactile sensors. 17 Sensors have been placed: five in the palm, three on each finger apart the little finger. In this picture the sensors in the thumb are hidden. The short blue cylinder that links the PUMA wrist to the hand is the J3R force sensor.

3.2 UNIZH robot hand

The tendon driven robot hand (see Figure 6 below) is partly built from elastic, flexible and deformable materials. For example, the tendons are elastic, the fingertips are deformable and between the fingers there is also deformable material. It has 15 degrees of freedom that are driven by 13 servomotors, a bending sensor is placed on each finger as a measure of the position, and a set of standard FSR pressure sensors covers the hand (e.g., on the fingertips, on the back and on the palm). A couple of Hitachi H8 microcontrollers are used, one generates the PWM signals for driving the motors and the other collects the sensory-motor data. The

communication between the computer and the sensory-motor controllers is based on the RS232 standard.

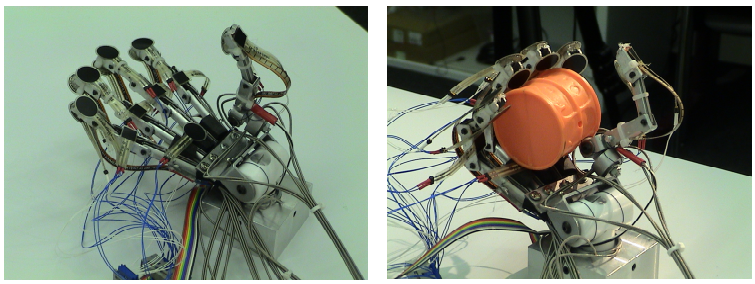


Figure 6: The tendon-driven robotic hand setup.

3.3 Behavioral experiments with infants

3.3.1 Experimental designs for the study of early intersensory integration: haptic and visual stimuli

The following set of pictures shows the design of the experiments with newborn infants. The stimuli used for both shape and texture integration are demonstrated. The description of the experimental preparation and paradigm is reported with full details in D4.1.

Visual objects (experiment SHAPE)

Haptic objects (experiment SHAPE)

Note that the shape is identical for visual and haptic objects but not the size (the newborn is myopic and has a very small hand).

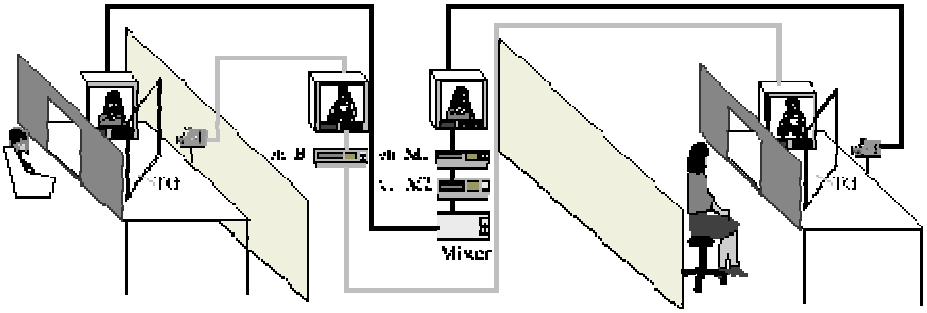
Visual object (experiment TEXTURE)

Haptic objects (experiment TEXTURE)

3.3.2 Experimental designs for the study of early detection of social contingency

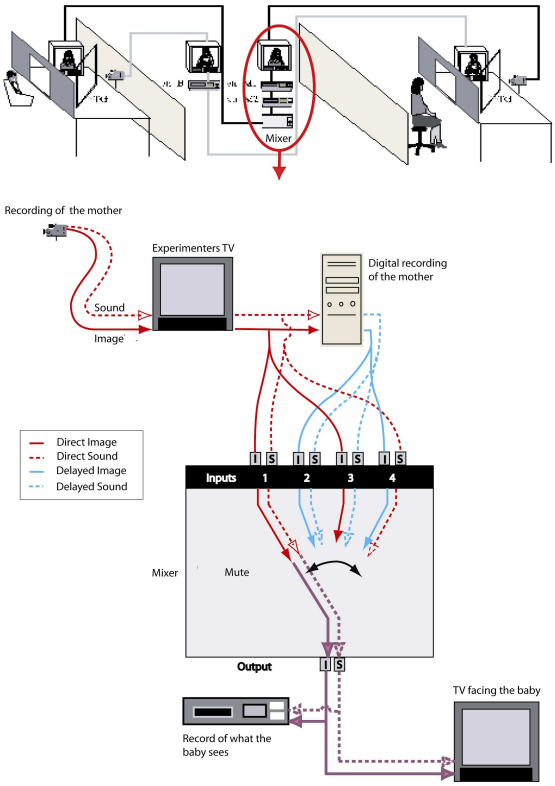
The following presents schematics of the teleprompter device developed at CNRS for the study of the development of interaction with people.

1. Non-contingency teleprompter device (developed by Nadel et al. in 1999) generating a delayed communication with the mother



Infant's room recording room adult's room

2. Novel setup allows decoupling sound with respect to video



4 Results and achievements

In general Adapt results were in good agreement with the planned activities. We had a history of delays especially due to the effort required to the preparation of the new hardware which we believe has been almost completely solved at the end of the project. The integration of the experiments is perhaps somewhat weaker than originally planned but still very compatible with the Technical Annex. We hoped for a full integration of the various results in the robotic architecture. This was not possible within the project timeframe but it will be certainly continued within the framework of other projects (for example, two of the partners of Adapt are now collaborating in a FP6 project with the goal of studying manipulation in greater details). The final demonstration sees the robot reaching and grasping object in real-time. The unsupervised feature analysis happens off-line. Our original idea was to have all learning happening on-line and providing feedback to the controller. The experiments with infants were carried out according to plan and eventually we have more results than what described in the Technical Annex. For example the experiments with texture developed after the first year of research. The research on morphology proceeded as planned although it was not possible so far to apply the same methods on different robotic platforms (at least within Adapt). One of the partners (UNIZH) has still collaboration with the University of Tokyo and University of Indiana – Bloomington where similar techniques are used and elaborated further.

As we mentioned earlier one of the major achievements of Adapt is also the integration of different disciplines. We believe this to be important especially for the long-term outlook of European research. Part of the integration, for instance, is demonstrated in the formulation of a short vocabulary of common terms that can be used by psychologists and roboticists alike.

4.1 Behavioral experiments

4.1.1 Intersensory integration in neonates during interaction with objects

The hypothesis of a primitive unity of senses at birth is held by several researchers (see for instance Maurer, 1997). This leads to postulate that newborns are capable of intermodal transfer from hand to vision as well as from vision to hand. Several studies have demonstrated that neonates can coordinate information between vision and touch. Streri and Gentaz (2003; 2004) have shown that 3-day-old newborns can visually recognize the shape of a previously felt object. It remained to examine the hypothesis of a reverse transfer, i.e. a tactual recognition of an object seen. To this aim, a series of four experiments was led within Adapt.

For the first two experiments, two groups of 12 full-term newborns with an average weight of 3 kilograms were randomly assigned to two habituation conditions. The 12 newborns of Group 1 (mean age: 49 hours) had a haptic habituation phase consisting of successive tactual presentation of a small wooden cylinder or a small prism in newborn's right hand until they reached habituation criterion. The test phase consisted of the visual presentation of familiar and novel objects during four trials. The 12 newborns of group 2 (mean age: 38 hours) had a visual habituation phase consisting of successive visual presentation of the big cylinder or prism until they reached habituation criterion. The test phase consisted of an alternative presentation in newborn's right hand of familiar and novel objects during four trials. The stimuli are those shown in section 3.3.1.
Eleven out of the 12 newborns in group 1 looked longer at the object that they had not held. This accounts for an intermodal transfer from touch to vision, thus replicating on a larger population previous findings by Streri (2000). However, there was no evidence of a reverse transfer from vision to touch in group 2. These results suggest that the acquisition and nature of information about shape gathered by vision and touch are different. This led us to examine cross-modal transfer between vision and touch for another property of object, such as texture. Newborns are able to compare texture density information across modalities (Molina & Jouen, 2003). Shape and texture are both amodal object properties, shared by vision and touch. Whereas shape is a structural property and is essential to object identification, texture is a material property that allows object identification when shapes are similar. A comparison between shape and texture properties in a cross-modal transfer task would allow us to understand how the visual and haptic modalities process information concerning object properties.

For this new test, two groups of 16 full-term newborns were randomly assigned to two habituation conditions. The 16 newborns of Group 1 (mean age: 45 hours) had a haptic habituation phase consisting of successive tactual presentation in the newborn's right hands of a small wooden cylinder or prism with "pearls" on it until newborns reached habituation criterion. The test phase consisted of a visual presentation of familiar and novel objects during four trials. The 16 newborns of group 2 (mean age: 60 hours) had a visual habituation phase consisting in successive visual presentation of the big cylinder or prism with "pearls" on it,

until newborns reached habituation criterion. The test phase consisted of an alternative presentation in newborn's right hand of familiar and novel objects during four trials.

Thirteen out of 16 newborns of group 1 looked longer to the object whose texture they had not previously felt. Thirteen out of 16 newborns of group 2 held longer the object they had not previously seen. These results show a cross-modal transfer of texture from touch to vision and the reverse. This suggests that information about texture is equivalent when gathered by touch and when gathered by vision. Texture might require low-level processing. Taken together, these results support the hypothesis that newborns are able to coordinate information between tactual and visual modalities depending of the object property considered (i.e. it is not a general property of the newborn's sensory system but one that depends on the modality/condition considered).

Previously, we evidenced a reverse cross-modal transfer of texture between vision and touch at birth. However, we presented shapes (cylinder) with texture (granular vs. smooth) and it is difficult to assess if cross-modal transfer observed from vision to touch was due to shape plus texture or to texture alone. We conducted an additional experiment with 16 newborns aged less than 3 days who were presented a wide surface (smooth vs. granular). The visual habituation phase consisted of successive presentation of a granular plate or a smooth plate in slight motion for several trials. The test phase consisted of the alternate presentation in the right hand of two objects, a flat smooth texture and a granular texture for 4 trials. Results revealed no significant difference in holding time between the two objects. There was no evidence of a cross-modal transfer of texture from vision to touch when shape information was reduced. This failure could be explained by the fact that "lateral motion", an exploratory procedure efficient in adult's exploration of texture, is absent in newborns. Newborns display only grasping and sometimes squeeze/release procedures. It is thus more difficult for newborns to process texture information when the object is flat than when volumetric. Further experiments are thus planned, using volumetric objects varying only in texture. This last experiment is very interesting showing again the link between information used for action (volumetric) vs. information used in passive judgment of object properties. This is a common theme of recent neuroscience (Miler and Goodale, 1996) and (Gallese et al., 1996).

4.1.2 Intersensory integration in young infants during interaction with persons

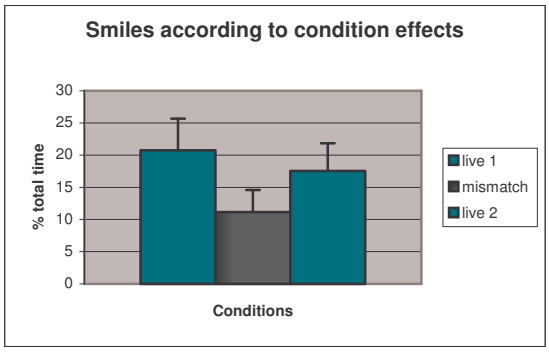
The hypothesis of a primary unity of senses was also tested when interaction concerned persons. Two series of experiments were conducted: one series dealt with the ability of very young infants to perceive, react and understand people as coherent multimodal entities; the second series investigated the hypothesis of a primitive awareness of being imitated in 2-month-old infants interacting with an imitative partner.

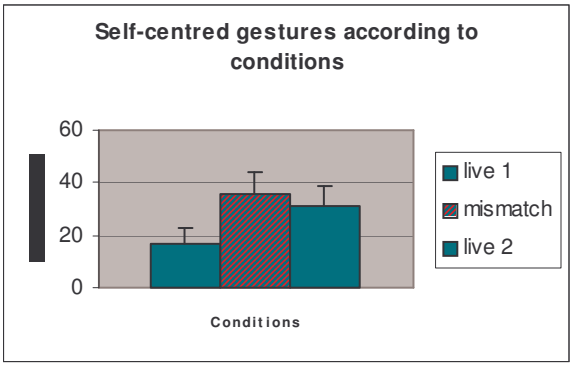
Do infants develop an early awareness of mother as a contingent multimodal entity? Do they capture contingency as a synchronic combination of different sensory modalities converging to produce an online interaction? Do they detect and expect partner's multimodal contingency? These questions have been examined with different methodologies. With classical experimental procedures using static displays and speech, young infants have been shown to be

sensitive to relationships between facial and vocal features of an adult (Bahrick, 2000). For instance, Kuhl and Meltzoff (1982) have shown that 4 month-olds can match a vowel sound with a facial pattern mimicking the sound of vowels. A first methodological advance toward the study of social perception requires shifting from static to dynamic displays. A good example of this shift is given by Walker-Andrews's presentation of two pre-recorded faces expressing different emotions accompanied by a voice matching the emotion of one of the two faces (Walker-Andrews, 1997). Infants around 6 months looked longer to the emotional face matching the voice, showing that they have somehow formed a representation of visual and auditory signals as a coherent multimodal dynamics. What is the influence of this knowledge on unimodal and multimodal interaction with a partner? A step in this direction was provided by the use of face-to-face interaction procedures with TV displays. TV manipulations allow the suppression of a channel, simulate a given disturbance of the partner (Murray& Trevarthen, 1985; Muir & Hains, 1999; Nadel, Carchon, Kervella et al., 1999; Nadel, Soussignan, Canet, Libert & Gérardin, 2005) or present the infant a life-like adult driven by an experimenter that chooses relevant responses to infants' signals in a prerecorded emotional repertoire (Smith & Muir, 2004). What does the use of these various displays tell us? In the course of a TV interaction, if the voice of the mother is turned off but her face remains contingently responsive, infants aged 5-6 months keep gazing and smiling to mother (Hains & Muir, 1996). However, if the voice is altered, smiling decreases and when the mother's face is disturbed, her voice helps the infants in maintaining visual attention (see Muir & Nadel, 1998, for a review). These findings suggest that face alone is sufficient for young infants to communicate but that intact voice is needed to keep or restore a positive emotional state if a disturbance is introduced in the partner's communication. In all cases, the emotional state is modified rather than visual attention.

Some of the above mentioned studies have shown that young infants detect correspondences or disruptions between face and voice; others have shown that infants are likely to exploit the resource of one channel if the other is disturbed. All these studies, however, have used static displays or face-to-face displays that do not maintain the dynamics of an interactive flow via mother's contingent responsiveness. Knowing that infants as young as 2 month detect a non-contingent communication in the course on an on-going dynamic interaction (Murray & Trevarthen, 1985; Nadel, Carchon, Kervella et al., 1999; Nadel, Soussignan, Canet, Libert & Gérardin, 2005), we suspect that processing bimodal communication should involve matching face and voice. What happens if the communication is only partially contingent? How do young infants process two sensory channels that emit simultaneously if one is contingent to their behavior and the other is not? Does it make a difference if the two disconnected channels come from the same source or come from two different sources?

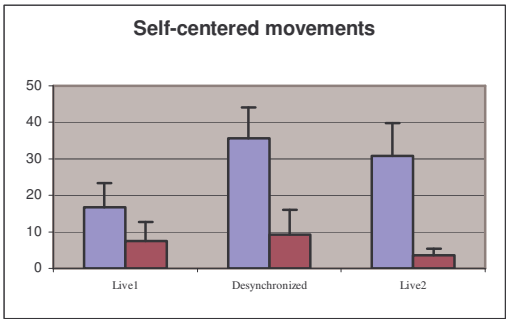
To test these questions, three experiments were set up all providing disconnected visual and auditory inputs to the infant but differed in the provenance of the sources: one source vs. two distinct sources.

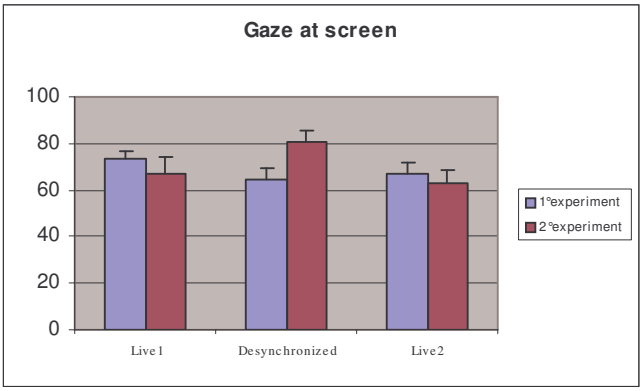




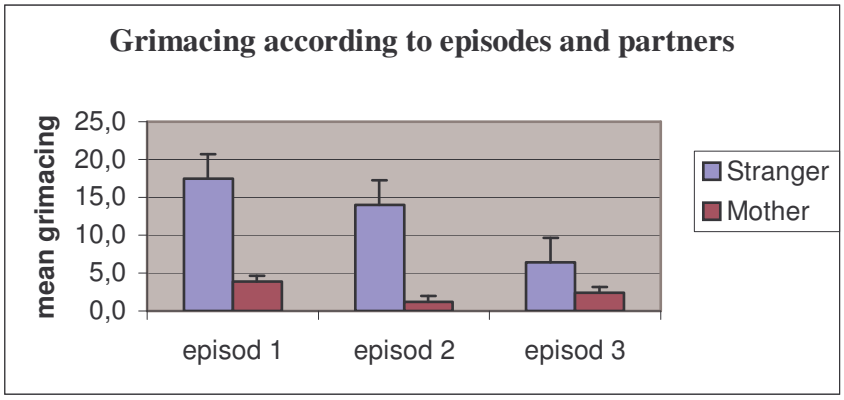
In the first experiment, conducted during the first year of Adapt, infants were presented a three-episode TV interaction with their mother. Via our teleprompter design, we chose to maintain the voice contingent throughout the three episodes of the interaction session and to present a non-contingent (replayed) face of the mother during the experimental episode. Doing so, we did not expect a gaze effect since previous studies have shown that gaze is not modified by various perturbations of the mother's face, but we hypothesized an emotional effect that will indicate a detection of incoherence in mother's facial message compared to the vocal message. Nineteen infants aged 6 months interacted with their mother in 3 conditions: mother on-line, mother's interaction delayed, mother on-line. The presence/absence of "Gaze to the screen", Smile, Grimace and self-centered movements were coded each 40/100th of a second for the three episodes. Results show a significant curvilinear trend for smile: smile decreased significantly during the maternal episode of mismatch between face and voice, and increased significantly when mother was on line again. A significant inverted curvilinear trend was found for self-centered gestures, indicative of stress.

In the second experiment, ten infants aged 6 months participated to the study. The experimental episode was composed of the contingent voice of the mother presented together with the pre-recorded face of a stranger. If there were no changes in the infant's emotional state, we could conclude that the mismatch between the familiar voice and the unacquainted face is not attributed to an incoherent partner but to two distinct sources. The presence/absence of Gaze to the screen, Smile, Grimace and self-centered movements were coded each 40/100th of a second for the three 30-second episodes.





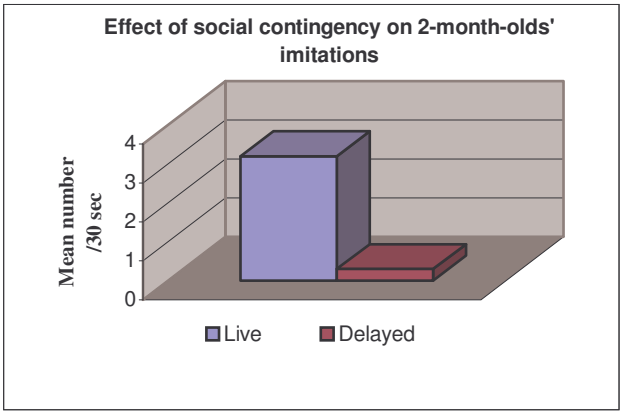
Comparing the results of the two experiments for the four indices, we found no significant difference during the first live interaction. The infants' response to the perturbation episode (episode 2) however, was significantly different for gaze: while infants withdraw from the image of the "dysfunctional mother", they maintained look at the screen in experiment 2, where mother's voice was coupled with a non-contingent stranger's face [$t(27) = 2.13$, $p<.04$]. Infants showed also a significantly higher amount of self-centered movements in experiment 1 [$t(27) = 2.07$, $p<.05$] and a marginally significant higher level of grimacing for experiment 1. Put together, these results show that infants were not disturbed by co-occuring signals coming from two different sources: mother's voice coupled with a stranger's face. They also suggest that infants as young as 6 month have formed the concept of "mother" as a multimodal entity whose co-occurring signals are synchronized. Note however that the mother's face is so familiar that a test with stranger A's voice coupled with stranger B's face is needed to confirm those results. Experiment 3 was mainly aimed at examining whether infants have formed also the concept of mother as a multimodal entity.



In the third experiment, 16 infants aged 6 months participated to the study. They had a three-episode interaction with their mother and a female stranger. The infants were presented first an episode of interaction with one modality only (contingent voice of partner), the screen remaining blank, a second episode with contingent voice of partner and pre-recorded face of a stranger, and a third episode with contingent partner's face and voice. We were interested to see whether the perception of a human partner as a multimodal entity extends to a stranger whose voice and face are unknown. Infants grimaced more when interacting with a stranger during blank screen and during the coupling of voice with another face but not during the third episode when voice and face were matched. This, together with the other results of this study, supports the idea that they have detected the mismatch as well as the final matching between stranger's face and voice. It is thus suggested that 6-month-olds have formed the concept of persons as multimodal agents whose signals are co-occurring synchronously.

Detection of social contingency implies to establish relationships between ones' behavior as perceived via proprioceptive information and the behavior of another, as seen. It requires establishing a relationship between what we see the other doing and what we feel being doing (cross-modal transfer between perception and proprioception). We test the development of this capacity in young infants through our teleprompter device. Mother (or experimenter) and infant (a 2 or 6 month old infant) can hear and see each other through TV monitors. The device

generates a seamless shift from maternal contingency to non-contingency and from non-contingency to contingency again. Thus the infant faces sometimes a contingent mother and sometimes a non-contingent mother. In a version of the method, the non-contingent episode experienced by the child is a replay of a previous contingent communication of the mother. This allows comparing the behavior of the infant facing the same gestural and verbal behavior of mother or experimenter in two conditions: when the partner's behavior is contingent to the infant's behavior, and when it is not.



Fifteen 2 month-olds reacted to non-contingent episode by a decrease of gazing to mother, a disappearance of smile, a dramatic increase of frowning (Nadel et al., 2005) thus replicating Nadel's previous results (Nadel et al., 1999). In exploring which parameters account for such a precocious detection of non-contingency, we found that the infants did not imitate during the non-contingent episode, whilst numerous imitations were coded during the contingent episode. We interpret this results as providing evidence that non-contingent behavior is an obstacle for experiencing a visual and auditory perception of what we are doing (i.e. experiencing other's agency in their mirroring of one's own behavior), which in turn is an obstacle to experience one's own agency in mirroring the other's behavior. Results of this experiment were reported in several conferences and published in Nadel et al., 2004. The data of the twin experiment conducted with an experimenter imitating the infant according to an experimental protocol are currently being coded.

During the third year of the contract we have paralleled with infants aged 6 to 12 months an experiment conducted in Genoa. In this experiment, the robot has to find a relationship between visual information about an object and proprioceptive anticipation of the grasping to operate. Our aim is to follow the development of perception-action coupling leading to reaching strategies that generate an affordant grasping of different kinds of objects. Manual skills increase with the emergence of the capacity to grasp an object, at around 5 months of age. Adaptation of reaching and grasping to object characteristics improves considerably over the next few months (von Hofsten, C., 1979. Development of visually guided reaching: The approach phase. Journal of Human Movement Studies, 5, 160-178). First imitations of actions appear at around 6 months and it is easier when objects are involved, but only in case of affordant relationships between object and action (Nadel, J. & Butterworth, G.1999. Imitation in infancy. Cambridge: CUP). A non-affordant modeling is not expected to lead to imitation in 6 months (following Von Hofsten's data with 6 month-olds in close experimental conditions),

but we expect to observe the beginning of hand preparation according to size and shape. Twenty full-term infants aged 6 month were recruited for this experiment. The infant was sitting on the mother's lap in front of a table. The object was placed in the center at such a distance that he/she has to reach the object first in order to grasp it. The experiment consists in two short episodes: 1) spontaneous grasping: the objects are presented one after the other in a counterbalanced order, and 2) grasping after a model: an experimenter grasps the object in a non-affordant way (i.e.: the bottle as a box; the small ball as a cylinder, etc.). A preliminary analysis of the data has shown 3 groups of infants: those who are not yet able to grasp, those who grasp after trials and without any account of object properties and those who configure their hand according to the object to be grasped. This extended heterogeneity in behavior reveals that the age of 6 months is a cornerstone in the development of a coupling of perception and goal-directed action. We found no example of an effect of modeling a non-affordant grasping on the infants' grasping procedure. The following picture illustrates the calibration, reaching and grasping in a 6-month-old.

(a) calibrating

(b) reaching

(c)grasping

4.2 Robotic experiments

In the following sections we describe the robotic experiments. This is only a summary while full details are reported in D4.4, D5.4, and D3.3. In particular we concentrate on the latest advancements leaving past work as references to the deliverables or published papers. The upper torso humanoid robot we describe in the following is called the Babybot which is simply a contracted word made of "baby" and "robot".

4.2.1 The robot visual system

One of the first steps of any visual system is that of locating suitable interest points in the scene ("salient regions" or events) and eventually directing gaze towards these locations. Human beings and many animals do not have a uniform resolution view of the visual world but rather only a series of snapshots acquired through a small high-resolution sensor (i.e. the fovea). This leads to two questions: i) how to move the eyes efficiently to important locations in the visual scene, and ii) how to decide what is important and, as a consequence, where to look next. Our robot mimics the same high-resolution fovea and low-resolution periphery of the human

retina[§]. The visual attention model we propose in these sections starts from the first stages of the human visual system, using then a concept of salience based on "proto-objects" defined as blobs of uniform color in the images. Subsequently, by exploiting the fact that the robot can act on the world, we were able to do something more: once an object is grasped, the robot could move and rotate it to build a statistical model of its color blobs, thus effectively constructing a representation of the grasped object in terms of proto-objects and their spatial relationships. This internal representation feeds then back to the attention system in a top-down fashion; as an example of this procedure, we demonstrated how it could be used to direct the attention to spot one particular object among several others that are visible on a table in front of the robot. We propose an object-based approach that integrates bottom-up and top-down cues; in particular bottom-up information suggests/identifies possible regions in the image where attention should be directed, whereas top-down information works as a prime for those regions during the visual search task (i.e. when the robot seeks for a known object in the environment).

In short, the attention processing takes input images, and extract blobs by first running an edge detector followed by the so-called watershed transform. Each blob is then tagged with the mean color of the pixels within its area (which leads in turn to a sort of quantized image). The result is blurred with a Gaussian filter and stored: it is averaged then with the next frame to obtain a temporal smoothing and reduce the effects of noise. After an initial startup of 4-5 frames, the number of blobs and their size stabilize. As discussed above, it is known that a feature or stimulus is salient if it differs from its immediate surrounding area. We chose to calculate the bottom-up salience as the Euclidean distance in the color opponent space between each blob and the average color of a ball surrounding it. The radius of the ball (the spot or focus of attention) is not fixed: rather, it changes with the size of the objects in the scene. In the same way the definition of "immediate surrounding area" should be relative to the size of the focus of attention. For this reason the greater part of the visual attention models in the literature uses a multi-scale approach and filters the salience map with suitable filters, or "blob" detectors (Itti & Koch, 2001). These approaches lack continuity in the choice of the size of the focus of attention. We propose instead to vary dynamically the region of interest depending on the size of the blobs. In other words, we compute the salience of each blob in relation to a neighborhood region whose size is proportional to that of the blob itself. In our implementation we use a rectangular region three times the size of the bounding box of the blob. The choice of a rectangular window is not accidental, it was chosen because filters over rectangular regions can be computed efficiently by employing an integral image representation as in (Viola & Jones, 2004). Blobs that are too small or too big are discarded from the saliency computation and will not be considered as possible candidates to be part of objects.

[§] The robot visual system works using logpolar vision (Sandini & Tagliasco, 1980).

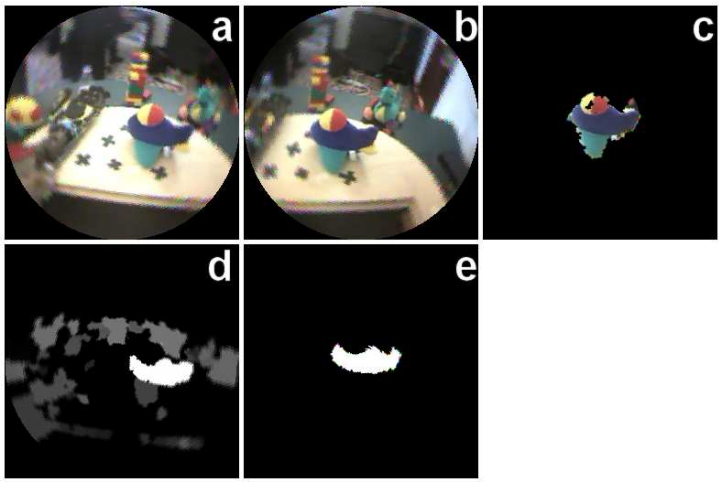


Figure 7: An example of the saccade behavior after the attention system.

The bottom-up saliency is computed as:

$$S_{bottom-up} = \sqrt{\left(\left\langle R^+G^-\right\rangle_{blob} - \left\langle R^+G^-\right\rangle_{surround}\right)^2 + \left(\left\langle G^+R^-\right\rangle_{blob} - \left\langle G^+R^-\right\rangle_{surround}\right)^2 + \left(\left\langle B^+Y^-\right\rangle_{blob} - \left\langle B^+Y^-\right\rangle_{surround}\right)^2} \quad (1)$$

where $\langle\ \rangle$ indicates the average of the image values over a certain area (as in the subscripts). The top-down influence on attention is, at the moment, calculated in relation to the visual search task. When the robot has acquired a model of an object and begins searching for it, it uses the visual information of the object to bias the saliency map. In practice, the top-down saliency map is computed as the distance between the average color of each blob and that of the target:

$$S_{top-down} = \sqrt{\left(\left\langle R^+G^-\right\rangle_{blob} - \left\langle R^+G^-\right\rangle_{object}\right)^2 + \left(\left\langle G^+R^-\right\rangle_{blob} - \left\langle G^+R^-\right\rangle_{object}\right)^2 + \left(\left\langle B^+Y^-\right\rangle_{blob} - \left\langle B^+Y^-\right\rangle_{object}\right)^2} \quad (2)$$

The total salience is simply estimated as the linear combination of the two terms above:

$$S = \alpha \cdot S_{top-down} + \beta \cdot S_{bottom-up} \quad (3)$$

The total salience map S is eventually normalized in the range 0-255, as a consequence the salience of each blob in the image is relative to the most salient one. The target of the next saccade is the center of mass of the most salient blob.

4.2.2 Learning about the self

In humans and biological systems the internal representation of the body is shaped during development and maintained according to the physical modifications occurring in life. In artificial agents (where the body does not change with time) adaptation can spare the tedious operation of manually tuning the system's internal models and their calibration but more importantly to compensate for unmodeled components. The latter might be required to compensate changes in the visual appearance of the body or drift in the sensors (e.g. the motor encoders). In infants this sense of the body emerges a few months after birth; indeed experiments have shown that, for example, five-month-old infants are already able to recognize the movement of their own legs on a mirror (Rochat & Striano, 2000). However this ability is not present at birth but it is acquired during development.

This is a chicken-egg problem: on the one hand the brain uses internal models to recognize the body whereas on the other it has to acquire the body-schema and maintain it up to date. To solve this problem, the brain needs a "bootstrapping" mechanism which allows the identification of the body and, in this way, the acquisition of the internal representation. To distinguish the body from the rest of the world the brain is thought to take advantage of extra information. For example, while a child waves the hand in front of her eyes, her brain "knows" what kind of motion is producing since it has exclusive access to the motor commands it sends to the muscles and the relative proprioceptive feedback. Pattern similarities between this information and other sensory feedback (mainly vision) may allow the brain to identify the hand (or any other body part) and distinguish it from other entities that move differently. The identification of similarities between different sensory channels, that is the perception of *intermodal forms*, is a possible candidate for this purpose. Other factors could be used as well, like *timing* or time coincidence of events (two events happening at the same time are more likely to have been originated by the same source). However detection of intermodal forms seems to play a dominant role whereas timing has a more flexible contribution during development. In other words, events happening in a relatively long time window are often considered by the brain as if they were originated from the same cause. The reason for this is that, probably, coincidence in time is used to detect causalities at different time scales and link more complicated actions with their relative perceptual consequences (consider for example the action of switching on a neon light) (Rochat & Striano, 2000).

We proposed an approach similar to Fitzpatrick and Arsenio (Fitzpatrick & Arsenio, 2004) and Metta and Fitzpatrick (Metta & Fitzpatrick, 2003) for visually segmenting the hand of the robot from the background. Repeated, self-generated actions were performed by the robot during the learning phase. In particular the robot was programmed to execute periodic movements of the wrist. The resulting motion of the hand was detected by computing the image difference between the current frame and an adaptive model of the background. The period of motion of each pixel in the resulting motion image was then computed with a zero-crossing algorithm; similar information was extracted from the proprioceptive feedback of each motor encoder. As a result, the hand of the robot was segmented by selecting, among the pixels that moved periodically, those whose period matched that of the wrist joints. The next step is to build a predictor for the hand position thus avoiding the active generation of periodic movements, which would be impractical in many situation (read slow). At this stage, to gather the training data the robot moved the arm randomly and then waved the hand for about a second; for each spatial location the segmentation of the hand was performed as described in the previous section. For each trial the center of mass of the segmented area was computed along with the best fitting ellipse parameters. The resulting (x,y) coordinates were used to train the first neural network whereas the ellipse parameters (orientation, major and minor axis) constituted the training samples for the second neural networks. It is important to take into account that the position of the hand in the visual field depends both on the posture of the arm and hand (this is not true, for example, for the orientation and size of the hand, if we do not consider translational effects). Unfortunately this enlarges the learning space and increases the time required for exploration (to collect the training set) and learning (higher dimensionality). For this reason the position of the hand was projected into an egocentric reference frame before being used to train the neural network. This last operation significantly reduced the dimensionality of the input space of the neural network. The output of the neural network is

then projected back to the retinocentric reference frame when necessary. Both projections (back and forth from egocentric and retinocentric reference frame) require knowledge of head inverse and direct kinematics. In the experiments reported here they were hardwired in the system, a possible procedure to estimate them is suggested in (Arsenio, Fitzpatrick, Kemp, & Metta, 2003). Figure 8 reports the block diagrams of the two models.

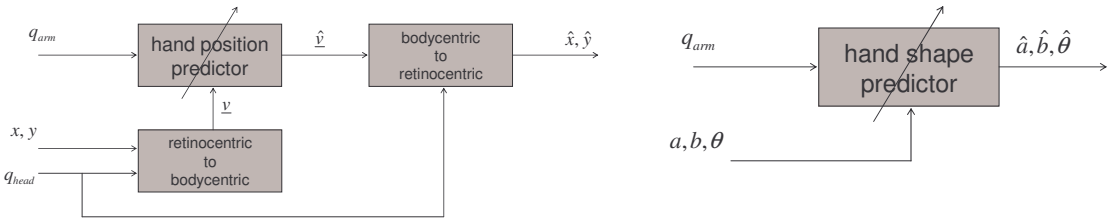


Figure 8: Internal models learned by the procedure described in text.

Learning was performed on-line by using the Schaal et al. model (Schaal & Atkeson, 1998) which was designed for incremental learning (use each sample only once during training). At the end of the exploration phase the robot has trained an internal model of the hand by which it could i) localize its center of mass ii) estimate its orientation and approximate size. These measures were used in numerous ways. The center of mass was employed to close a visual loop to direct gaze towards the hand. For this task the internal model was queried with the proprioceptive feedback of the arm. Another possibility was to query the model with the arm motor command (final joint position) to obtain where the hand would be at the end of the movement. In general this model offers a means of computing a prediction of the position, size and orientation of the hand from a given arm configuration or, in other words, of simulating a motor action.

4.2.3 Reaching

The solution we propose is based on the use of a direct mapping between the eye-head motor plant and the arm motor plant (Metta, Sandini, & Konczak, 1999). Flanders and colleagues (Flanders, Daghestani, & Berthoz, 1999), through an experiment of reaching in humans, suggested that the information about gaze direction might be employed by the brain to establish a reference point for reaching. They analyzed the error when reaching in the dark and showed how this correlates to the error of the gaze (the gaze drifts away from the target in the dark). Accordingly one premise we make is that the position of the fixation point coincides with the object to be reached. In other words, reaching for an object starts by looking at it. Under this assumption, the fixation point can be considered as the "end-effector" of the eye-head system. The position of the eyes with respect to the head, determines uniquely the position of the fixation point in space relative to the shoulder. The arm motor command can be obtained by a transformation of the eye-head motor/positional variables. We called this approach "motor-motor coordination", because the coordinated action is obtained by mapping motor variables into motor variables:

$$q_{arm} = f(q_{head}) \tag{4}$$

where q_{head} and q_{arm} are head and arm posture respectively (joint space).

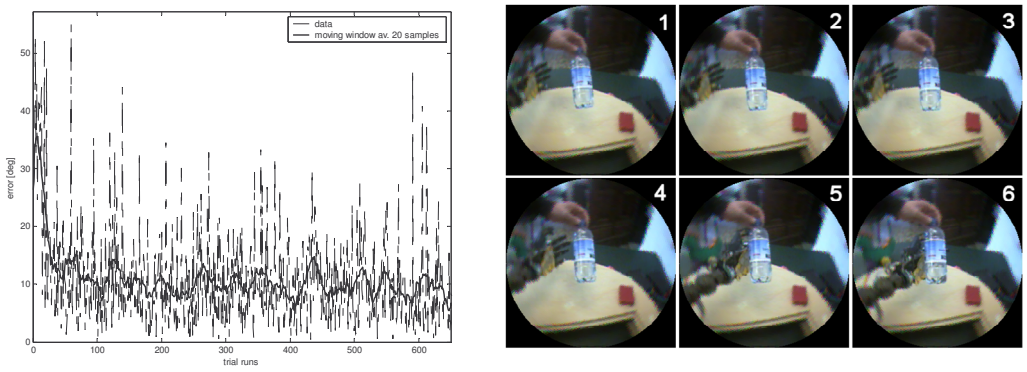


Figure 9. Reaching error (left). As new examples are gathered and presented to the network the performance increases. This improvement is less remarkable; we believe this is due to noise in the training data which affects not only the learning, but also the measure of performance. An exemplar sequence of a reaching action after the learning is reported on the right.

What is interesting in this approach is not equation (4) per se, which, after all, implements the inverse kinematics of the arm, but the mechanisms used to learn it. In fact, this mapping can be easily learnt when the tracking behavior described in the previous section is active. The robot explored the workspace by moving the arm randomly, while simultaneously, it tracked its hand; whenever the eyes fixated the hand a new sample consisting of the arm and head posture was acquired and used to train a neural network approximating equation (4).

We can note that the reaching problem can also be solved in the image plane. Consider the planar case (i.e. no 3D information is available) and suppose to measure the position of the end point in the image plane (x, y). We want to control the arm to reach a target point (x^*, y^*). We can solve the problem by means of a closed loop controller, by following a fairly standard visual servoing approach:

$$\Delta\mathbf{q} = -k \cdot \mathbf{J}(\mathbf{q})\Delta\mathbf{x} \tag{5}$$

where:

$$\Delta\mathbf{x} = \begin{bmatrix} x - x* \\ y - y* \end{bmatrix} \tag{6}$$

$k > 0$ is a scalar and $\mathbf{J(q)}$ is the Jacobian of the transformation between the image plane and the arm joint space. For a 2 dimensional arm $\mathbf{J(q)}$ is a 2 by 2 matrix whose elements are a non-linear function of the arm joint angles. Given the image Jacobian, it is possible to drive the endpoint toward any point in the image plane. At least locally, the Jacobian can be approximated by a constant matrix. In our case:

$$\Delta\mathbf{q} \simeq -k \cdot \hat{\mathbf{J}}\left(\bar{\mathbf{q}}\right)\Delta\mathbf{x} = -k \cdot \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\Delta\mathbf{x} \tag{7}$$

Convergence is guaranteed if the following condition is met:

$$J^{-1}(q)\hat{J}\left(\bar{q}\right) > 0 \tag{8}$$

Since following the procedure of section 4.2.2 the robot has learnt a direct transformation between the arm joint angles and the image plane, it can now recover the position of the endpoint in the image plane from a given joint configuration:

$$\begin{bmatrix} x \\ y \end{bmatrix} = f(q) \tag{9}$$

Indeed, to compute a local approximation of the Jacobian, a random sampling of the arm joint space around a given point $(\bar{\mathbf{x}}, \bar{\mathbf{q}})$ can be performed:

$$\mathbf{q_i} = \bar{\mathbf{q}} + \Delta\mathbf{q_i} \tag{10}$$

with

$$\Delta\mathbf{q_i} = \boldsymbol{\eta}(\mathbf{0}, \boldsymbol{\sigma}) \tag{11}$$

and where $\boldsymbol{\eta}(\mathbf{0}, \boldsymbol{\sigma})$ follows a normal distribution of zero mean and standard deviation of 5 degrees.

For each sample, by applying equation (9) we obtain a new value $\mathbf{x_i} = \bar{\mathbf{x}} + \Delta\mathbf{x_i}$ that can be used to estimate the Jacobian around $\bar{\mathbf{q}}$ with a least squares procedure:

$$\Delta\mathbf{q}_i = \begin{bmatrix} \Delta\mathbf{x}_i^T & \mathbf{0} \\ \mathbf{0} & \Delta\mathbf{x}_i^T \end{bmatrix} \cdot \begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \end{bmatrix} \tag{12}$$

$\hat{\mathbf{J}}(\bar{\mathbf{q}})$ can then be used in the closed loop controller to drive the arm toward a specific position in the image plane. However, there is no need to close the loop with the actual visual feedback. By using the map in equation (9), in fact, we can substitute the actual visual feedback with the internal simulation provided by the model. From the output of the closed loop controller we can estimate the position of the arm at the next step, by assuming a pure kinematic model of the arm; in this way the procedure can be iterated several times to obtain the joint motor command required to perform a reaching movement. The flowchart below explains this procedure.

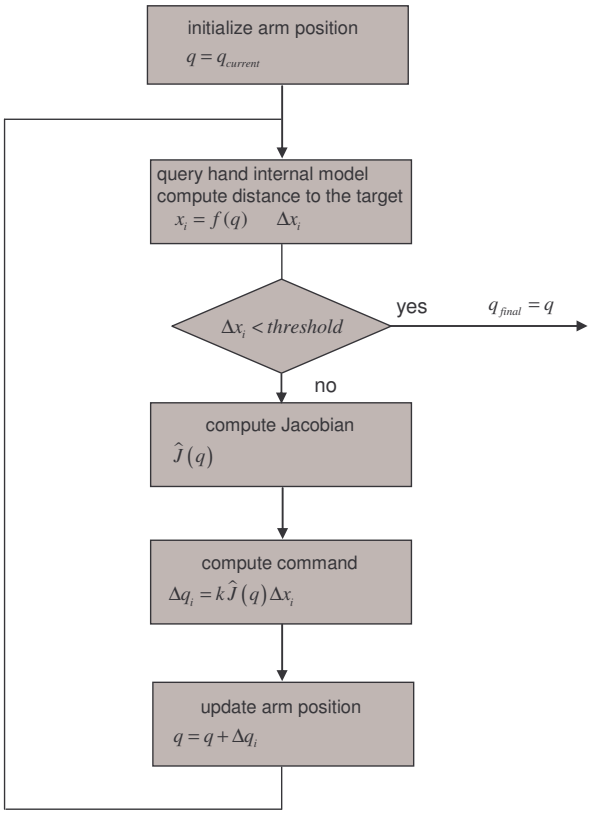


Figure 10. Closed-loop approach to reaching, flowchart. See text for further details.

The main limitation of this approach is that we do not make use of three-dimensional visual information; while this is a clear limitation of this implementation, the same approach can be easily extended to the full 3D case. The implementation is consistent with the hand internal model which provides the position of the hand in the image plane of one of the eyes only (left). Since in the Babybot the hand position is uniquely described by three degrees of freedom (the first three joints of the Puma arm), this technique was necessarily used to control only two of them (arm and forearm). Given the kinematics of the Puma arm this allowed to perform movements on the plane defined by the shoulder joint.

Let us summarize what we have described in this section. We have introduced two approaches to solving the inverse kinematics of the manipulator. The first method uses a mapping between the posture of the head (whose fixation point implicitly identifies the target) and the arm motor commands; it allows controlling the arm to reach any point fixated by the robot[**]. The second approach uses the hand internal model to compute a piecewise constant approximation of the inverse Jacobian and simulate small movements of the arm in the neighborhood of the desired target. The procedure is iterated several times to compute the motor command required for reaching the target.

[**] During the learning of the motor-motor map, the robot tracks the palm of the hand.

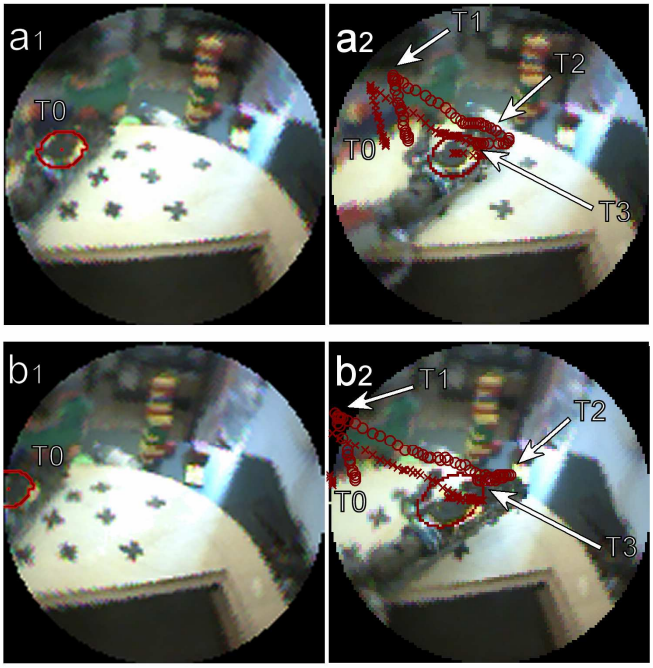


Figure 11. Arm trajectories for two reaching actions (a) and (b). T0 marks the position of the hand at the beginning of the action. Crosses correspond to the position of the palm; circles show the position of the fingers. The action is divided in three phases. From T0 to T1 arm prepositioning. From T1 to T2, reaching: in this case the motor-motor map is used to move the palm towards the center of the visual field (the target). A small adjustment with the arm Jacobian is performed to position the fingers on the target (T2 to T3).

4.2.4 Learning about objects

In this section we describe a method for building a model of the object grasped by the robot. We assume for a moment that the robot has already grasped an object; this can happen because a collaborative human has given the object to the robot (as we describe in the next section) or because the robot has autonomously grasped the object. In this case the robot may spot a region of interest in the visual scene and apply a stereotyped action with the arm and the hand to catch it. Both solutions are valid bootstrapping behaviors for the acquisition of an internal model of the object. When the robot holds the object it can explore it by moving and rotating it.

In short, the idea is to represent objects as collections of blobs generated by the visual attention system and their relative positions (neighboring relations). The model is created statistically by looking at the same object several times from different points of view. At the same time the system estimates the probability of each blob to belong to the object by counting the number of times each blob appears during the exploration.

In the following, we use the probabilistic framework proposed by Schiele and Crowley (Schiele & Crowley, 1996a, , 1996b). We want to calculate the probability of the object O given a certain local measurement M. This probability $P(O|M)$ can be calculated using Bayes' formula:

$$P(O \mid M) = \frac{P(M \mid O)P(O)}{P(M)}$$
$$O_{MAP} = \arg\max_{O,\sim O} \{P(O \mid M), P(\sim O \mid M)\} \quad (13)$$

where $P(O)$ is the *a priori* probability of the object O, $P(M)$ the *a priori* probability of the local measurement M, and $P(M|O)$ is the probability of the local measurement M when the object O is fixated. In the following experiments we carried out only a single detection experiment, there are consequently only two classes, one representing the object and another representing the background. $P(O)$ and $P(\sim O)$ are simply set to 0.5 because this choice does not affect the maximization. Since a single blob is not discriminative enough, we considered the probabilities of observing pairs of blobs; the local measurement M becomes the event of observing both a central (i.e. fixated) and surrounding blobs:

$$P(M \mid O) = P(B_i \mid B_c \text{ and } (B_i \text{ adjacent } B_c)) \quad (14)$$

where B_i is the i^{th} blob surrounding the central blob B_c which belongs to the object O. That is, we exploit the fact the robot is fixating the object and assume B_c to be constant across fixations of the same object – this is guaranteed by the fact the object is being hold by the hand. In practice this corresponds to estimating the probability that all blobs B_i adjacent to B_c (which we take as a reference) belong to the object. Moreover the color of the central blob B_c will be stored to be used during visual search to bias the salience map. This procedure, although requiring the "active participation" of the robot (through gazing) is less computationally expensive compared to the estimation of all probabilities for all possible pairs of blobs of the fixated object. Estimation of the full joint probabilities would require a larger training set than the one we used in our experiments. The probabilities $P(M|\sim O)$ are estimated during the exploration phase with the blobs not adjacent to the central blob. The local measurements were considered independent, because they refer to different blobs, so the total probability $P(M_1,\ldots,M_N|O)$ can be factorized in the product of the probabilities $P(M_i|O)$. An object is detected if the probability $P(O|M_1,\ldots,M_N)$ is greater than a fixed threshold.

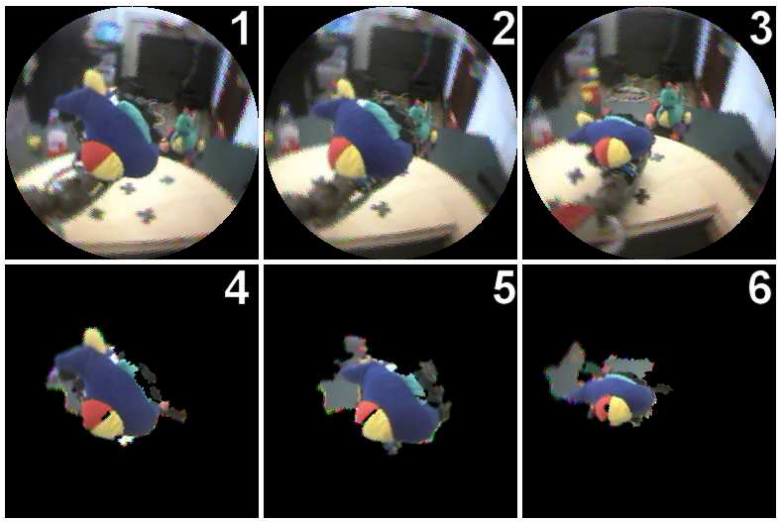


Figure 12. Object exploration and corresponding segmentation 1-3 and 4-6 respectively. The segmentation consists in the object central blob together with the relative adjacent ones. Notice that fixation is maintained on the object by using the hand localization module as explained in Section 4.2.2.

Our requirement was that of building the object model with the shortest possible exploration procedure. Unfortunately, the small training set might give histograms *P(M|*)* with many empty bins zero counts bins. To overcome this problem a probability smoothing method was used. A popular method of zero smoothing is Lidstone's law of succession (Lidstone, 1920):

$$P(M \mid O) = \frac{count(M \wedge O) + \lambda}{count(O) + v\lambda} \tag{15}$$

for a *v* valued problem. With $\lambda=1$ and a two valued problem ($v=2$), we obtain the well-known Laplace's law of succession. Following the results of Kohavi et al. (Kohavi, Becker, & Sommerfield, 1997), we choose $\lambda=1/n$ where *n* is equal to the number of frames utilized during the training. Then our probability estimator becomes:

$$P(M \mid O) = \frac{count(M \wedge O) + 1/n}{count(O) + v/n} \tag{16}$$

When an object is detected after visual search, a possible figure-ground segmentation is attempted, using the information gathered during the exploration phase. Each blob is segmented from the background if it is adjacent to the central blob and if its probability to belong to the object is greater than 0.5. This probability is approximated using the estimated probability as follows:

$$P\left(B_i \in O \mid B_c \text{ and } \left(B_i \text{ adiacent } B_c\right)\right) \simeq P\left(B_i \mid B_c \text{ and } \left(B_i \text{ adiacent } B_c\right)\right) \tag{17}$$

As an example Figure 13 shows the result of the segmentation procedure.

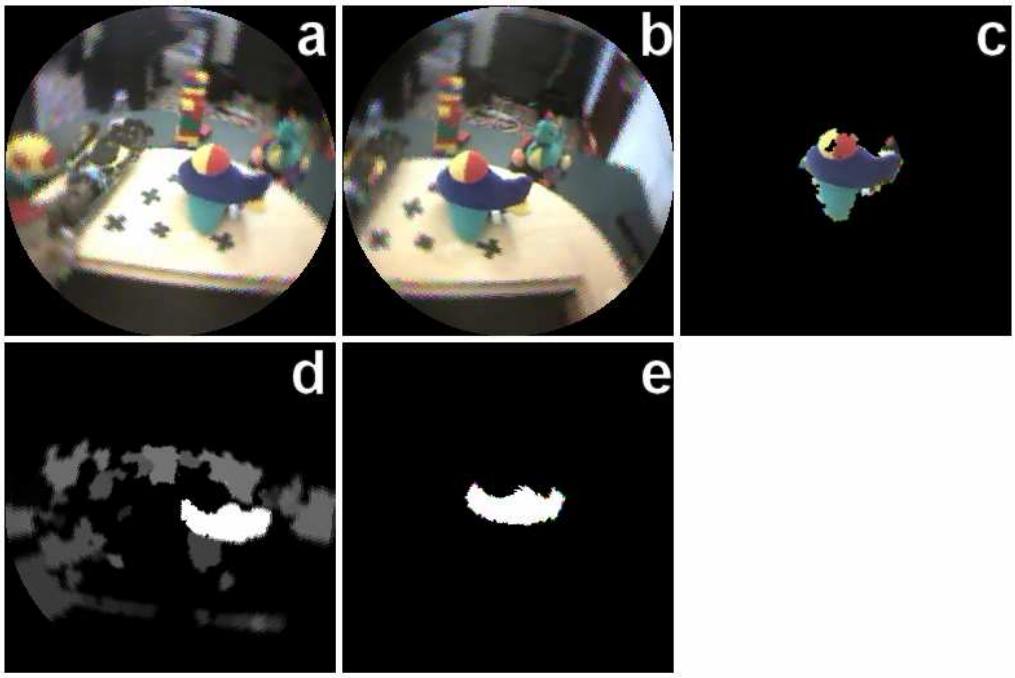


Figure 13. Visual search. The robot has acquired a model of the airplane toy during an exploration phase (not shown); this information primes the attention system. The blue blob at the center of the airplane is selected and a saccade performed. (a) and (b) show the visual scene before and after the saccade. (d) and (e) show the output of the visual attention system synchronized with (a) and (b) respectively. The result of the segmentation after the saccade is in (c).

In table 1, results are shown of using a toy car and a toy airplane as target objects; 50 training sessions were performed for each object. The first column shows the recognition rate, the second the average number of saccades (mean ± standard deviations) it takes the robot to locate the target in case of successful recognition, and the third in case of unsuccessful recognition.

Object	Recognition rate	Number of saccades when recognized	Number of saccades when not recognized
Toy car	94%	3.19±2.17	3±1.41
Toy airplane	88%	3.02±2.84	3.5±0.76

Table 1. Performance of the recognition system measured from a set of 50 trials.

4.2.5 Grasping

The modules described in the previous sections can be integrated to achieve a meaningful grasping behavior. Figure 14 can be used as a reference for the following discussion. The action starts when an object is placed in the robot's hand and the robot detects pressure in the palm (frame 1). This elicits a clutching action of the fingers; the hand follows a preprogrammed trajectory, the fingers bend around the object toward the palm. If the object is of some appropriate size, the intrinsic elasticity of the hand (as described in Section 3.1) facilitates the action and the grasping of the object. The robot moves the arm to bring the object close to the cameras and begins its exploration. The object is placed in four positions with different orientations and background (frames between 2 and 6). During the exploration, the robot tracks the hand/object; when the object is stationary and fixation is achieved, a few frames are acquired and the model of the object trained as explained in Section 4.2.4. At the end of the exploration the object is released (frame 4). At this point the robot has acquired the visual model of the object and starts searching for it in the visual scene. To do this, it selects the blob whose features better match those of the object's main blob and perform a saccade. After the saccade the model of the object is matched against the blob that is being fixated and its surrounding. If the match is not positive search continues with another blob, otherwise grasping starts (frames 7-8-9). At the end of the grasp the robot uses haptic information to detect whether it is holding the object or the action failed. In this process the weight of the object and its consistence in the hand is checked (the shape of the fingers holding the object). If the action is successful the robot waits for another object, otherwise it performs another trial (search and reach).

It is fair to say that part of the controller was preprogrammed. The hand was controlled with stereotyped motor commands. Three primitives were used: one to close the hand after pressure was detected, and two during the grasping to pre-shape the hand and actually clasp the object. The robot relied on the elasticity of the hand to achieve the correct grasping. To facilitate grasping, the trajectory of the arm was also programmed beforehand; waypoints relative to the final position of the arm were included in the joint space to approach the object from the top.

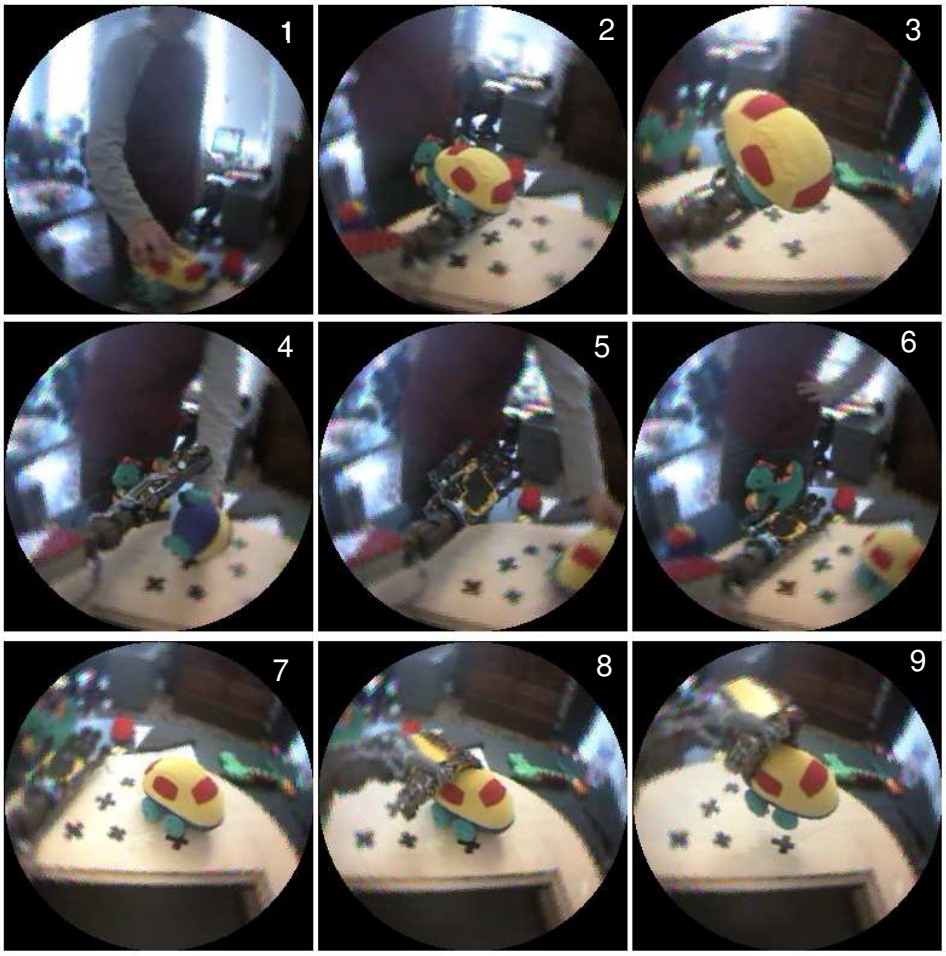


Figure 14. A sequence of the robot grasping an object. The action starts when an object is placed on the palm (1). The robot grasps the object and moves the eyes to fixate the hand (2). The exploration starts in (3) when the robot brings the object close to the camera. The object is moved in four different positions while maintaining fixation; at the same time the object model is trained (3-6). The robot drops the object and starts searching for it (7). The object is identified and a saccade performed (7-9). The robot eventually grasps the toy (10-12).

4.2.6 Semi-supervised learning

The motivation for learning a body schema and reaching behaviors were discussed in details in D5.1. In particular, we referred to this procedure as Self-Supervised Learning (SSL) since it employs supervised learning techniques but frees the experimenter from the burden of preparing the training set for the learning machinery. In this schema, the training data are gathered online and autonomously by the robot through an exploration procedure (see, for example, section 4.2.2). In D5.1 we also motivated the use of unsupervised (or partially supervised) techniques for extracting conspicuous features automatically from unlabeled data. The model was further elaborated in D5.2, D5.3 and D5.4. D5.3 D5.3 details the semi-supervised learning model which combines aspects from unsupervised and supervised learning.

All data for the semi-supervised learning experiments were recorded on the Babybot using the procedures and hardware described in D4.4 and outlined in the previous section: that is, the robot acting, actively looking and grasping objects. Three datasets have been used:

- 5000 log-polar color images: The images do not form a continuous sequence but resulted from the fixations to salient objects as driven by the attention system;
- Grasping sequences under the control of a preprogrammed grasping module; includes all modalities.
- See-feel: objects placed in the visual field and felt by the hand (grasp reflex initiated by touching the palm with the object). Only vision, touch and hand configuration at the end of grasp can be expected to carry useful information. Eight different objects were used and ten samples of each were recorded.

The main reason for first using visual modality alone was to collect enough image statistics to be able to compress the huge amounts of visual data (76,608 pixels/sample) by unsupervised learning. The grasping sequences are interesting because the robot perceives the results of its own movements, thus having the possibility to distinguish its own body from the environment. In addition, the ability to visually recognize the shape and position of a hand would be useful not only for visuomotor control of grasping but for interpreting the actions of others, too. The third dataset may have the potential to develop visual shape features under the guidance of proprioceptive information about shape. With this data, this may well turn out to be impossible because the amount of the data is small compared to the difficulty of the problem. However, we expect that the experience gained with this data will help in designing future experiments exploring the issue more deeply.

Unsupervised approach to learning has been quite successful in modeling the very first stages of human visual processing. Basically this means that human vision seems to make use of the statistical structure found in natural images. In our case, it is necessary to reduce the amount of visual data because the other modalities alone don't provide enough information to guide the development of useful features.
We have used principal and independent component analyses (PCA and ICA) to reduce the dimensionality of the visual data and to extract prominent features. These turned out to be mostly:

- edge features without color selectivity and
- color features without orientation selectivity.

which are shown in Figure 15 below.

Figure 15: Several ICA components extracted from a particular location in the retinal image. Note the edge-like and 4-color basis features. It is worth noting that this is only a subset of the full decomposition. Other components do not show this nice selectivity resulting in a mixture of color and edges.

Interestingly, this seems to agree quite well with what is known about the human visual system. The agreement may be partly due to the structure of the camera eye which has three color sensitive receptors like the human eyes, and partly because natural images have a statistics that is clearly the same no matter who is the observer (the sensor structure might affect the measurement though). It should be emphasized that unsupervised learning was able to autonomously learn many interesting aspects about the data such as:

- Each pixel in the camera was sensitive to one color (red, green or blue) but the edge features appeared to selectively discard the information about the wavelengths and only retain information about light intensity.
- The geometry of the camera pixels was learned autonomously. Due to the log-polar geometry of the cameras (see D4.4), edges appear to be curved. This was reflected in the developed edge features which correspond to curved edges on retina but straight edges in the outside world (and in the retina).
- The detectors were denser and with smaller receptive fields in the fovea, reflecting denser sampling.

Although the image dataset used to estimate the features did not have any temporal structure, the features can convey temporal information if their outputs are temporally filtered. Also, although ICA discovers a linear mapping, it is possible to include nonlinearities (such as the absolute value of the edge features) which make further processing stages useful (see D5.4 for more discussion).

At the time of writing, the analysis of the multi-modal representations by semi-supervised learning is still underway. The plan is to study the development of visual features under the guidance of proprioceptive information using the model described in D5.3. Two sets of experiments have been planned:

- Using proprioceptive information about hand position and posture to guide the development of visual hand features.
- Using proprioceptive information about finger configuration to guide the development of visual shape features.

4.2.6.1 Future research directions in semi-supervised learning

The development of the model for multi-modal integration and feature extraction is still work in progress but it is already apparent that the approach is useful (with several real-world applications ranging from climate data analysis to mobile network signal detection) and has inspired new hypotheses about human perceptual learning. A rather straightforward and biologically motivated extension to the model will be to include nonlinear top-down prediction to guide learning. So far we have used linear predictions but it is nowadays thought that the apical dendrites integrate their inputs in a nonlinear fashion (see D5.4, Section 2.3, for discussion about the role of apical dendrites). This may open up new, simple and robust learning methods for nonlinear mappings because, unlike normal supervised learning algorithms, it is not necessary to accurately predict the magnitude of feature activations. It is enough to find a correlate for the activations. This should obviate the need for prediction error computations and could instead rely on simple correlation-based techniques akin to Hebbian

learning. Another potentially very fruitful research direction that came up as a result of the ADAPT project is a new hypothesis about the role of attention in perceptual learning (see Valpola, 2004 and 2005). It is well established through psychophysical research that attention has a major role in perceptual learning. However, the underlying mechanisms are poorly understood. Our model of perceptual learning relies on feedback information to guide the development of feature extraction. Anything that can modulate this information should be able to modulate learning. Attention is clearly a process which modulates the flow of information and, moreover, it is strongly influenced by motivation and goals. This seems to put attention in a good position to mediate the guidance from motivation and goals to perceptual learning.

4.2.7 Morphology and information theory analysis of manipulation

This section presents a subset of the experiments performed to study the influence of morphology into data self-structuring especially for what regards manipulation. Manipulation entails manual haptic perception, which is the ability to gather information about objects by using the hands. Haptic exploration is a task-dependent activity, and when people seek information about a particular object property, such as size, temperature, hardness, or texture, they perform stereotyped exploratory hand movements. In fact, spontaneously executed hand movements are best at maximizing the availability of relevant sensory information gained by haptic exploration (Lederman and Klatzky, 1990). The same holds for visual exploration. Eye movements, for instance, depend on the perceptual judgment that is requested by the task, and the eyes are typically directed toward areas of a visual scene or an image that deliver useful and essential perceptual information. To reason about the organization of saccadic eye movements, Lee and Yu (1999) proposed a theoretical framework based on information maximization. The basic assumption of their theory is that due to the small size of our foveas, our eyes have to continuously move to maximize the information intake from the world. Differences between tasks obviously influence the statistics of visual and tactile inputs, as well as the way the brain acquires information for object discrimination, recognition, and categorization.

Clearly, the common denominator underlying our perceptual abilities seems to be a process of sensorimotor coordination which couples perception and action. It follows that coordinated movements must be considered part of the perceptual system (Thelen and Smith, 1994), and whether the sensory stimulation is visual, tactile, or auditory, perception always includes associated movements of eyes, hands, arms, head and neck (Ballard, 1991; Gibson, 1988). Sensorimotor coordination is important, because (a) it induces correlations between various sensory modalities (such as vision and touch) that can be exploited to form cross-modal associations, and (b) it generates structure in the sensory data that facilitates the subsequent processing of those data (Lungarella and Pfeifer, 2001; Lungarella and Sporns, 2004; Nolfi, 2002; Sporns and Pegors, 2003).

One of our goals is to quantitatively understand what sort of coordinated motor activities lead to what sort of information. We also aim at identifying “fingerprints” (or patterns of sensory or sensorimotor activation) characterizing the agent-environment interaction. Our approach builds on top of previous studies on category learning (Pfeifer and Scheier, 1997; Scheier and Pfeifer, 1997), as well as on work on the information-theoretic and statistical analysis of sensory and motor data (Lungarella and Pfeifer, 2001; Sporns and Pegors, 2003; Te Boekhorst et al., 2003).

We first introduce some of the quantities that are calculated as a measure of sensorimotor coordination. Correlation quantifies the amount of linear dependency between two random variables *X* and *Y*, and is given by the following formula:

$$Corr(X,Y) = (\sum_{x \in X} \sum_{y \in Y} p(x,y)(x - m_X)(y - m_Y) / \sigma_X \sigma_Y \quad (18)$$

where $p(x;y)$ is the second order (or joint) probability density function, m_X and m_Y are the means, and σ_X and σ_Y are the standard deviations of *x* and *y* computed over *X* and *Y*. Note that the analysis was performed by fixing the time lag between the two time series to zero. The entropy of a random variable *X* is a measure of its uncertainty, and is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \quad (19)$$

where $p(x)$ is the first order probability density function associated with *X*. In a sense entropy provides a measure for the sharpness of $p(x)$. The joint entropy between variables *X* and *Y* is similarly defined as:

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p(x,y) \log p(x,y) \quad (20)$$

Mutual information measures the statistical independence of two random variables *X* and *Y* (Cover and Thomas, 1991; Shannon, 1948). Using the joint entropy $H(X,Y)$, we can define the mutual information between *X* and *Y* as:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad (21)$$

In comparison with correlation, mutual information provides a better and more general criterion to investigate statistical dependencies between random variables (Steuer et al., 2002). For entropy as well as for mutual information, we assumed the logarithm in base 2. Correlation, entropy and joint entropy were computed by first approximating $p(x)$ and $p(x;y)$. The most straightforward approach is to use a histogram-based technique, described, for instance, in (Steuer et al., 2002). Because the sensors had a resolution of 5 bits, we estimated the histograms by setting the number of bins to 32 (which led to a bin-size of one). Having a unitary bin size allowed us to map the discrete value of the sensory stimulus directly onto the corresponding bin for the approximation of the joint probability density function. Because of the limited number of available data samples, the estimates of the entropy and of the mutual information were affected by a systematic error (Roulston, 1999). We compensated for this bias by adding a small corrective term *T* to the computed estimates: $T=(B-1)/2N$ to the entropy estimate (where *N* is the size of the temporal window over which the entropy is computed, and *B* is the number of states for which $p(x_i)$ is not zero), and $T=(B_x+B_y-B_{x,y}-1)/2N$ to the mutual information estimate (where B_x, B_y, $B_{x,y}$, and *N* have an analogous meaning to the previous case).

We have performed experiments with three different experimental setups:

- hand-eye coordination setup;

- anthropomorphic robot hand setup;
- simulated mobile robot setup.

For the scope of Adapt, clearly, the anthropomorphic hand is the most interesting experimental setup. Figure 16 shows a 13 degree of freedom robotic hand equipped with bending and pressure sensors (Gomez et al., 2005).

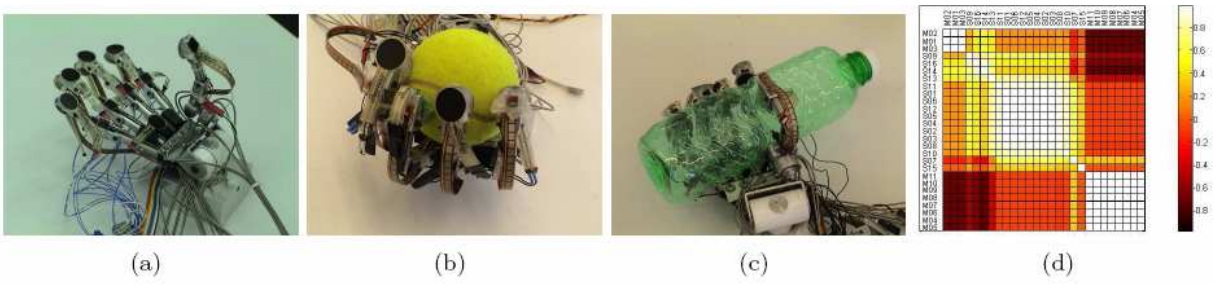


Figure 16: Tendon driven robot hand in (a). (b) and (c) grasping different objects. (d) correlation matrix obtained from the pair-wise coorelation of the bending sensors, pressure sensors, and motor encoders for one particular experimental run.

The hand is controlled by a neural network based system that allows exploring its movement capabilities when the hand is in contact with objects (of different shapes and materials) and for some of our work we focused on how the neural network could exploit the hand's anthropomorphic morphology to speed up the learning of grasping. Furthermore, the robustness of the evolved neural controller was tested by making systematic changes in the robot's morphology (e.g., position, number and types of the sensors, stronger motors, covering materials in order to increase the friction with objects) to investigate how the neural controller reacts to unforeseen perturbations.

Deliverable 3.3 presents a more detailed discussion on the application of the information theory analysis of sensorimotor data. For example, it is possible to note that different behaviors generate different "correlation" patterns. The idea here is to exploit these "patterns" to distinguish between different states the robot encounters and decide on which part of the state space (which might consist of several sensors) concentrate the learning resources (which are a finite amount in any reasonable autonomous agent).

4.2.8 Multisensory integration using information theory

In robotics, we can conceive crossmodal perception as an extension of the active-vision/perception-action paradigm. The crossmodal perceptual agent can employ multisensorial cues to reinforce its explorative perception and to create actively synchronized multisensorial inputs (e.g. by hitting repeatedly an object on the ground producing a change in both the visual field and the aural input). We implemented this type of perception by trying to solve, in the context of a humanoid robotic architecture, two problems: a) object segmentation using multisensorial cues, and b) sound classification for attentional priming. We will show that a combination of speech recognition techniques and statistics can be used to create a crossmodal perceptual architecture that can create associations between the images of toys and the sounds

the toys produce; and, in a second stage, evocate the toy's visual image by recognizing the sound associated to the toy, and consequently, have the potential to exploit this visual expectation in additional explorative movements.
For the experiment, we used a set of three baby toys. Figure 17 shows the group of toys as seen by the robot. Figure 17(a) is a deformable yellow plastic duck; it produces a high frequency sound when squeezed with the hand. The hollow hard plastic toy pigs shown in Figure 17(b) Figure 17(c) are the same toy; the differences are: they have different colors and we have filled them with different materials. Therefore, the sound produced by each toy pig was slightly different.

Figure 17: Three objects as seen from the robot: (a) A deformable plastic yellow duck, (b) a hollow hard plastic blue pig filled with plastic bottle caps, and (c) a hollow hard plastic red pig filled with chickpeas.

The complete architecture is show in the block diagram in Figure 18 and we will make frequent reference in the following description of the implemented system. The first step in the processing of the signal is to parameterize the auditory input to obtain a low dimensional representation of sound. In the speech recognition literature this module is known as the signal-processing front-end. The idea is to have a sequence of measurements of the input signal, usually the output of some type of spectral analysis technique that yields a "pattern" that represents the sound; though we prefer the term sound template for this representation. This sound template is a sequence of spectral vectors. Each of these vectors represents the frequency transformation of the sound in a short period of time; in our system, this period of time is 40 milliseconds long. Therefore, the sound template is a representation of the sound both in time and in frequency.
One of the most popular techniques used in speech recognition and, based on well-enstablished line of research, is method called mel-frequency cepstral coefficients (MFCC). The MFCC algorithm can create a compact representation of sound into a vector of few parameters. We tested the MFCC algorithm in the Matlab environment using the auditory toolbox developed by Malcolm Slaney (1998) and then we implemented a C++ version based on his algorithm for the robot environment.
In short, the traditional approach to spectral analysis of the sound signal consists in applying a set of filter-banks (see Rabiner and Juang, 1993). According to (Rabiner and Juang, 1993), the filter bank computation can be conveniently implemented by applying first a short-time Fourier transform (STFT) to the incoming data:

$$S_n(e^{jw_i}) = \sum_m s(m)w(n-m)e^{-jw_i m} \qquad (22)$$

where *s(m)* is the sound sequence, and *w(n−m)* is in our case a Hamming window. The STFT produces a representation of the sound stream both in time and frequency domains that facilitates the application of the filter-bank in the frequency domain. Rabiner proposes that the filter-bank can be implemented by varying adequately the frequency in the exponential term of equation (22); in the simplest case, this frequency has an uniform distribution choosing *fi = i(Fs/N)*, where *Fs* is the sampling frequency.

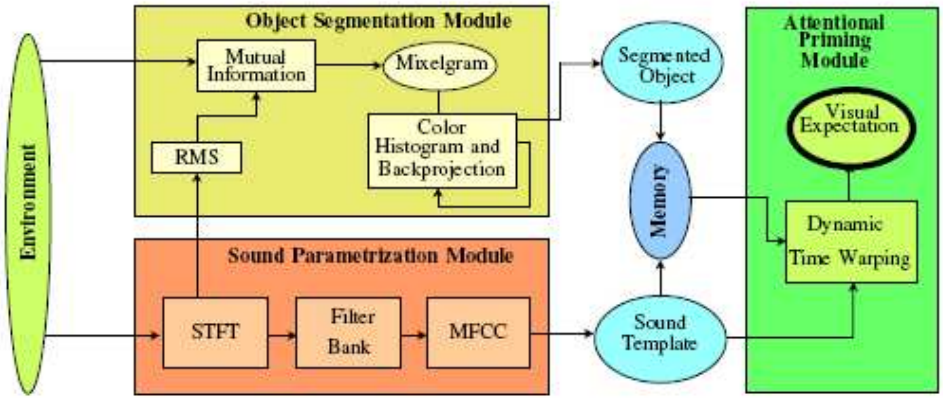


Figure 18: Block diagram of the crossmodal architecture.

However, non-uniform frequency distributions can be used; in particular, neurophysiological studies propose various models of the human auditory system. One of those is the mel-frequency scale where the filter-banks are distributed linearly in low frequencies and then they decrease logarithmically in higher frequencies. As suggested in (Slaney, 1998), we constructed the filter-bank using 13 linearly-spaced filters (133.33 Hz between center frequencies) followed by 27 log-spaced filters (separated by a factor of 1.0711703 in frequency). The mel-frequency cepstral transform is computed as follows:

$$c_i = \frac{2}{N}\sum_{k=1}^{N} Y_k \cos[i(k+0.5)\frac{\pi}{N}], \quad i = 1,2,...,M \qquad (23)$$

where c_i is the cepstral coefficient, and Y_k are the outputs of the filter-bank discussed in the previous section. In our system, the MFCC transform reduces the dimensionality by transforming the output of 40 filter-banks into a compact representation of 13 cepstral coefficients. Figure 19 shows a graphical 3D representation of a MFCC transform applied to the sound produced by toy Figure 17(a).

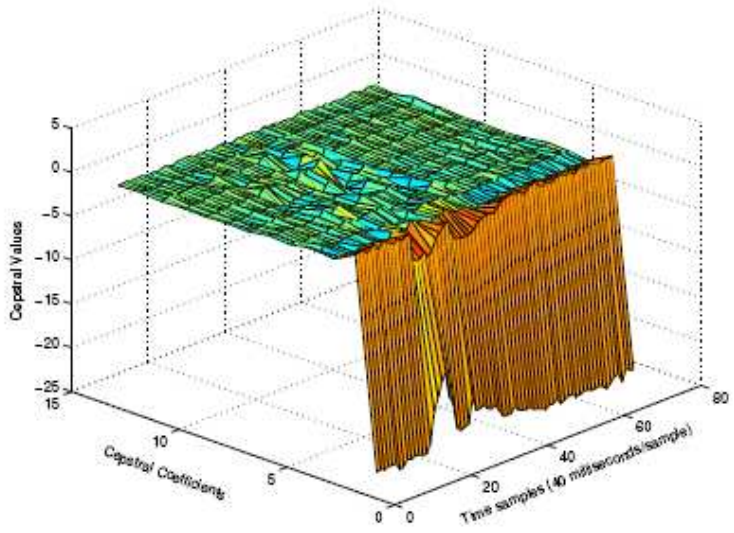


Figure 19: Three dimensional representation of a MFCC transform.

After applying equation (23) we packed the cepstral coefficients in the sound template data structure. This template contains the cepstral coefficients associated to a sound produced by a toy. To detect the presence of an object producing a sound, we measure empirically the background sound level and we use it as a threshold to activate the template recording procedure. Once the sound is parameterized, the level of synchrony of the sound and visual data streams needs to be measured. For this purpose, we use the method suggested by Hershey and Movellan based on the mutual information (Hershey and Movellan, 2000). They define the temporal synchronization of video and sound channels as an estimate of the mutual information between both streams. Their algorithm was originally applied to the problem of finding a vocalizing person in a video sequence. Let $a(t)$ be a vector describing the acoustic signal at time t and $v(x,y,t)$ be a vector describing the video signal at the same time instant. Still from (Hershey and Movellan, 2000), the authors assume that these vectors form a set S of audio-visual vectors and that these vectors are independent samples from a joint multivariate Gaussian process. Under these assumptions, Hershey and Movellan affirm that an estimate of the mutual information can be calculated as:

$$I(A(t_k);V(x,y,t_k)) = \frac{1}{2}\log_2 \frac{|\sum_A(t_k)||\sum_V(x,y,t_k)|}{|\sum_{A,V}(x,y,t_k)|} \qquad (24)$$

where the sum over $|A(t_k)|$ is the determinant of the covariance matrix of the audio stream, $|V(x,y,t_k)|$ is the determinant of the covariance matrix of a pixel of the image (e.g. the RGB values), and $|A,V(x,y,tk)|$ is the joint covariance of both the audio and visual signals (see [6] and [20] for details about how to derive (24)). To compute equation (24) different sound and image parameterizations can be used. In a first experiment, we used 13 mel-frequency cepstral coefficients (the parameters of covariance matrix $A(t_k)$) and three RGB values of the pixel (the parameters of the covariance matrix $V(x,y,t_k)$) during 0.6 seconds ($S = 15$). Consequently, the

combined audio-vision covariance matrix $A,V(x,y,t_k)$ comprises 15x15 elements. The computation of the determinants of these matrices exhibits a considerable computational cost, because the determinants are calculated for each pixel in the image. This produces a considerable degradation of the system performance. Although this algorithm can be improved by having a distributed computation, we decided to use a simplified version of the mutual information. This is the special case when the data streams are in a one dimensional representation (i.e. n = m = 1). Then, the mutual information can be expressed as:

$$I(A(t_k);V(x,y,t_k) = -\frac{1}{2}(1-\rho^2(x,y,t_k)) \quad (25)$$

where $\rho^2(x,y,t_k)$ is the Pearson correlation coefficient between $A(t_k)$ and $V(x,y,t_k)$. To obtain this one dimensional representation, we used for the sound the root mean square (RMS) of the short-time Fourier transform coefficients (see the arrow connection between the STFT box and the RMS box in figure 1) and a gray level value of the color RGB components. Notice that the MFCC transform was still used to form the sound template representation.

To conceptualize the output of the mutual information between sound and vision, Prince et al. (2004) introduced the *mixel*; that is a combination of the words mutual and pixel. They proposed that the mixels form a topographic representation called mixelgram. These can form shapes that are perceptually relevant for human observers. Therefore, the mixelgram is to be considered a common space representation for both visual and audio sensorial channels. Figure 4 depicts an example of the mixelgram of the toy Figure 17(a). It is possible to distinguish the shape of the duck.

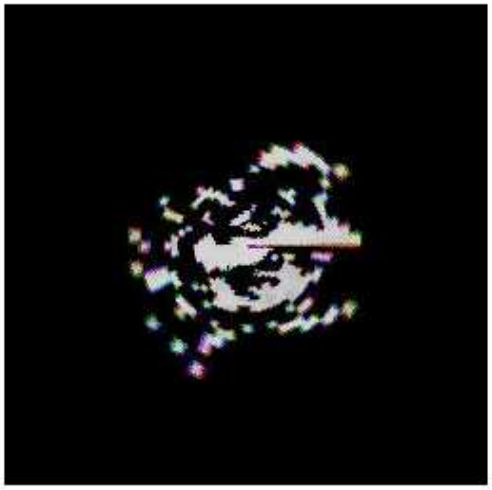


Figure 20: the mixelgram of one of the toys.

The original image and the mixelgram maintain a direct geometric correspondence; therefore the mixelgram can be used to segment the object by segmenting the pixel in the original image whose position corresponds to an activated mixel. However, the segmentation obtained with this method has a low quality because many pixels of the object are not segmented at all. To improve the object segmentation, we used a technique based on color segmentation and with the additional assumption that the activated mixels belongs to a uniformly colored object. After

an object is segmented, the segmentation results are stored in a dynamic lookup table. Each element in the lookup table contains the segmented image and the sound template associated to that object. To create the memory we present the object in front of the robot several times squeezing or shaking it with different speeds and strengths. By this procedure, we produced slightly different sounds that were associated to the same object in the memory. This provided some robustness to the process of recognizing the sound.
This module performed basically a pattern classification for sound identification. When the system hears an unknown sound, the sound is parameterized using the MFCC algorithm explained earlier. Then, the sound template is compared with the memorized sound templates using a measure of similarity (distance). To compare the sound templates it is necessary to compute both a local distance measure between the spectral vectors, and a global time alignment procedure. To compute the local distance, we used the truncated cepstral distance; time alignment is obtained by the dynamic time warping algorithm (Rabiner & Juang, 1993).

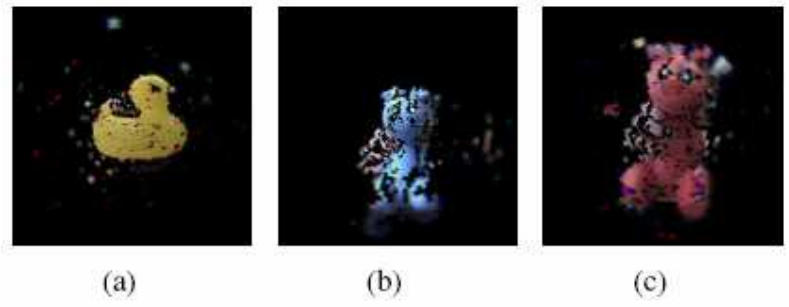


Figure 21: The segmented toys.

The following table presents the experimental results obtained from a data set of 100 trials per object. In the case of segmentation, the table shows the percentage of segmentation trials with similar results of those presented in Figure 21. Since color segmentation is used, lighting conditions influence the segmentation. The results presented were obtained with under sufficient illumination. In the case of the sound, experiments were carried out in a quiet environment with only the background noise generated by computers and the running robot. The results in both situations degrade significantly in noisy conditions, as for example, with people talking in the room. For the recognition module we used only the $c_1 . . . c_{12}$ cepstral coefficients. The use of the c_0 cepstral coefficient degraded the capacity of the system to distinguish between similar objects. This was the case with the two pig toys that are made of the same material. Perhaps c_0 can be employed to recognize between different classes of objects. This may be convenient when the classification needs to be done among a large number of different sounds.

Experiment	Duck	Blue pig	Red pig
Segmentation	64%	70%	75%
Classification	99%	88%	83%

4.2.9 Conclusions

The previous sections described a set of experiments that attacked the core aspects of representation from both the developmental psychology and the computer sciences point of view. Overall, it is perhaps clear that these are but bits of information in the global picture that we are tried to uncover in Adapt: that is, the nature of representation. Nonetheless, we can see some progress and, we can note the potential of a multidisciplinary approach to scientific problems. Information technology and its derivates, next to the realization of the tools making modern neuroscience possible, can now actively participate into the experimental phase, into the design of the experiments, into a fruitful discussion. This is not completely new since it started already with some of the pioneering work on artificial neural networks, but it might require now a new effort. One of the big changes in our opinion is represented by the possibility of building real hardware (and software) in support of models, something that was not possible years ago. The penetration of robotics into our daily life, as for example the AIBOs) is certainly a sign of the advances made by robotic technology.

During the execution of the project there were many slight adjustments to the directions taken. A certain degree of non-homogeneity is thus to be expected and this is fully reflected in the history of deliverable reports. As often happens we had to focus research to specific areas to make progress and consequently other aspects had to be neglected: For example, the issue of motivations into the architecture, though very important, has been considered only tangentially and aspects of the autonomous development of features were not fully integrated into the robotic system.

At partial justifications of this difference between the plan and the actual implementation we would like to stress the fact that these are issues that were never solved by the AI community first, computer vision and robotics later on. We believe we have given a valid contribution and certain parts of our models have been validated in the various robotic setups. This validation through robotic implementation is something that was not available years ago and, thus, it makes a clear difference with previous work.

A somewhat anecdotal example regards the Babybot's attention system. We started with a space variant visual system and moving cameras whereas the most part of the vision community typically assumes attention over one static image (a sort of a fixed background) with gaze moving over this uniform representation of space. This is plainly misleading: the geometry of our attention system is influenced by our motor system (Craighero et al. 2004). In our case, it has been even difficult to compare a robot with space variant vision with a static implementation of the attention system. But even in the non-foveated case (e.g. regular rectangular images), moving the cameras requires a full elaboration of the new visual input (after the movement). This clearly requires some attention and poses constraints to the attention system which are not otherwise included in the static case. We think this approach being as close as possible to the real world bears the potential for major breakthroughs in the years to come.

4.2.10 References

Arsenio, A. (2004). Cognitive-Developmental Learning for a Humanoid Robot: A Caregiver's Gift, PhD Thesis, (CSAIL, MIT, Boston, 2004)

Bahrick, L.E. (2000). Increasing specificity in the development of intermodal perception (pp.117-136). In D. Muir &.A. Slater (Eds.). Infant Development: The Essential Readings, Oxford: Blackwell Publishers.

Ballard, D. (1991). Animate vision. Artificial Intelligence, 48(1):57-86.

Cover, T. and Thomas, J. (1991). Elements of Information Theory. New York: John Wiley.

Crary, J. (1992). Suspension of Perception. Cambridge (Mass), MIT Press.

Craighero L., Fadiga L., Nascimben M. (2004). Eye Position Affects Orienting of Visuospatial Attention ELSEVIER SCIENCE - Current Biology, Vol. 14, 331–333, February 17.

Dretske, F. (1995). Naturalizing the Mind. Cambridge (Mass), MIT Press.

Fitzpatrick, P. and A. Arsenio. (2004). Feel the beat: using cross-modal rhythm to integrate perception of objects, others and self, in: Fourth International Workshop on Epigenetic Robotics, Genoa, Italy. (Lund University Cognitive Studies Publisher).

Flanders, M., L. Daghestani, and A. Berthoz. (1999). Reaching beyond reach, Experimental Brain Research, 126(1) 19-30.

Gallese V., Fadiga L., Fogassi L., Rizzolatti G. (1996). Action recognition in the premotor cortex, Brain, 119:593-609.

Gibson, J. J. (1952). "The Visual Field and the Visual World." Psychological Review LIX: 148-151.

Gibson, J. J. (1977). The theory of affordances. Perceiving, acting, and knowing: Toward an ecological psychology. R. E. Shaw and J. Bransford. Hillsdale (NJ), Lawrence Erlbaum Associates.

Gibson, J. J. (1979). The Ecological Approach to Visual Perception. Boston, Houghton Mifflin.

Gibson, E. (1988). Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge. Annual Review of Psychology, 39:1-41.

Gomez, G. and Eggenberger Hotz, P. (2004a). An evolved learning mechanism for teaching a robot to foveate. In Proc. of the 9th Int. Symp. on Artificial Life and Robotics (AROB-9), pages 655-658.

Gomez, G. and Eggenberger Hotz, P. (2004b). Investigations on the robustness of an evolved learning mechanism for a robot arm. In Proc. of the 8th Int. Conf. on Intelligent Autonomous Systems (IAS-8), pages 818-827.

Gomez, G., Hernandez, A., Eggenberger Hotz, P., and Pfeifer, R. (2005). (in press). an adaptive learning mechanism for teaching a robot to grasp. In To appear in Proc. of AMAM 2005.

Gomez, G., Lungarella, M., and Tarapore, D. (2005). Information-theoretic approach to embodied category learning. In Proc. of the 10th Int. Symp. on Artificial Life and Robotics (AROB10), Beppu, Oita, Japan., pages 332-337.

Hains, S., & Muir, D.W. (1996). Effects of Stimulus Contingency in Infant-Adult Interactions. Infant Behavior and Development, 19, 49-61.

Hershey, J. and J. Movellan. (2000). Audio-vision: Using audiovisual synchrony to locate sounds. Advances in Neural Inforomation Processing Systems, 12.

Itti, L., C. Koch, and E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11) (1998) 1254-1259.

James, W. (1890/1950). The Principles of Psychology. New York, Dover.
Jones, K. S. (2003). "What Is an Affordance?" Ecological Psychology 15(2): 197-114.
Kohavi, R., B. Becker, and D. Sommerfield. (1997). Improving simple Bayes, in: European Conference on Machine Learning.
Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. Science, 218, 1138-1141.
Lederman, S. J. and Klatzky, R. L. (1990). Haptic exploration and object representation. In Goodale, M., editor, Vision and Action: The Control of Grasping, pages 98{109. New Jersey: Ablex.
Lee, T. S. and Yu, S. X. (1999). An information-theoretic framework for understanding saccadic behaviors. In Proc. of the First Intl. Conf. on Neural Information Processing. Cambridge, MA: MIT Press.
Lidstone, G. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities, Transactions of the Faculty of Actuaries, 8 182-192.
Lungarella, M. and Pfeifer, R. (2001). Robots as cognitive tools: Information-theoretic analysis of sensory-motor data. In Proceedings of the 2nd International Conference on Humanoid Robotics, Waseda, Japan.
Lungarella, M. and Sporns, O. (2004). Methods for quantifying the informational structure of sensory and motor data. Neuroinformatics. in preparation.
Maurer, Daphne. (1997). "Neonatal synaesthesia: implications for the processing of speech and faces." In S. Baron-Cohen and J. Harrison (Eds.); Synaesthesia: Classic and Contemporary Readings; Oxford, England: Blackwell. Pp. 224-242.
Metta, G., G. Sandini, and J. Konczak. (1999). A Developmental Approach to Visually-Guided Reaching in Artificial Systems, Neural Networks, 12(10) 1413-1427.
Metta G. and P. Fitzpatrick. (2003). Early Integration of Vision and Manipulation, Adaptive Behavior, 11(2) 109-128.
Milner, D. and M. A. Goodale. (1996). The Visual Brain in Action (Oxford Psychology Series, No. 27). Oxford University Press, Oxford, UK.
Molina M, Jouen F. (2003). Haptic intramodal comparison of texture in human neonates. Developmental Psychobiology. May 2003;42(4):378-85.
Muir, D. & Nadel, J. (1998). Infant social perception. In A. Slater (Ed.). Perceptual Development: Visual, Auditory and Speech Perception in Infancy (pp. 247-286). Psychology Press: East Sussex, UK.
Muir, D. & Hains, S. (1999). Young infants' perception of adult intentionality: Adult contigjency and eye direction. In P. Rochat (Ed.). Early Social Cognition: Understanding Others in the First Months of Life (pp. 155-188).
Murray, L., & Trevarthen, C. (1985). Emotional regulations of interactions between two-month-olds and their mothers. In T. M. Field & N. A. Fox (Eds.), Social perception in infants (pp. 177-197). Norwood, NJ: Ablex.
Nadel, J., Carchon, I., Kervella, C. et al. (1999). Expectancies for social contingency in 2-month-olds. Developmental Science, 2, 164-174.
Nadel, J. et al. (2004). Toward communication: first imitations in infants, children with autism and robots. Interdisciplinary Journal of interaction studies, 1, 45-75.
Nadel, J., Soussignan, R., Canet, P., Libert, G., & Gerardin, P. (2005). Two –month old infants' emotional state after non-contingent interaction. Infant Behavior and Development.
Nolfi, S. (2002). Power and limit of reactive agents. Neurocomputing, 49:119-145.

Pfeifer, R. and Scheier, C. (1997). Sensory-motor coordination: The metaphor and beyond. Robotics and Autonomous Systems.

Pfeifer, R. and Scheier, C. (1999). Understanding Intelligence. MIT Press.

Prince, C.G., G. J. Hollich, N. A. Helder, E. J. Mislivec, A. Reddy, S. Salunke, and N. Memon. (2004). Taking synchrony seriously: A perceptual-level model of infant synchrony detection. In Proceedings of the Fourth International Workshop on Epigenetic Robotics.

Rabiner, L. and B-H Juang. (1993) Fundamentals of Speech Recognition. Prentice Hall Signal Processing Series. Prentice Hall.

Rochat, P. and T. Striano. (2000). Perceived self in infancy, Infant Behavior & Development, 23 513-530.

Roulston, M. (1999). Estimating the errors on measured entropy and mutual information. Physica D, 125:285-294.

Sandini G. and V. Tagliasco, An Anthropomorphic Retina-like Structure for Scene Analysis, Computer Vision, Graphics and Image Processing, 14(3) (1980) 365-372.

Schaal, S. and C. G. Atkeson. (1998). Constructive Incremental Learning from Only Local Information, Neural Computation, (10) 2047-2084.

Scheier, C. and Pfeifer, R. (1997). Information theoretic implications of embodiment for neural network learning. In ICANN 97, pages 691-696.

Schiele, B. and J. L. Crowley. (1996). Where to look next and what to look for, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, (Osaka, 1996).

Shannon, C. (1948). A mathematical theory of communication. Bell System Tech. Journal, 27.

Slaney, M. (1998). Auditory toolbox. version 2. Technical Report 1998-010, Interval Research Corporation.

Smith L., & Muir, D. (2004). Infant perception of dynamic faces: emotion & eye direction effects. In O. Pascalis & Slater, A. (Eds.). The development of face processing in infancy and early childhood: current perspectives. New York: Nova Science Publishers.

Sporns, O. and Pegors, J. (2003). Generating structure in sensory data through coordinated motor activity. In Proc. of Intl. Joint Conf. on Neural Networks, page 2796.

Steuer, R., Kurths, J., Daub, C., Weise, J., and Selbig, J. (2002). The mutual information: detecting and evaluating dependencies between variables. Bioinformatics, 18:231-240. Suppl.2.

Streri, A., & Gentaz, E. (2003). Cross-modal recognition of shape from hand to eyes in human newborns. Somatosensory & Motor Research, 20(1), 11-16.

Streri, A. & Gentaz,E.(2004). Cross modal recognition of shape from hand to eyes and handedness in human newborns. Neuropsychologia, 42, 1365-1369.

Tarapore, D., Lungarella, M., and Gomez, G. (2004). Fingerprinting agent-environment interaction via information theory. In Proceedings of the 8th Int. Conference on Intelligent Autonomous Systems (IAS-8), Amsterdam, The Netherlands, pages 512-520.

Tarapore, D., Lungarella, M., and Gomez, G. (2005). Quantifying patterns of agent-environment interaction. Robotics and Autonomous Systems(accepted for publication).

Te Boekhorst, R., Lungarella, M., and Pfeifer, R. (2003). Dimensionality reduction through sensory-motor coordination. In Kaynak, O., Alpaydin, E., Oja, E., and Xu, L., editors, Proc. of the Joint Inlt. Conf. ICANN/ICONIP, pages 496{503. LNCS 2714.

Thelen, E. and Smith, L. (1994). A Dynamic Systems Approach to the Development of Cognition and Action. Cambridge, MA: MIT Press. A Bradford Book.

Valpola, H. (2004). Behaviourally meaningful representations from normalisation and context-guided denoising. AI Lab technical report, University of Zurich.
Valpola, H. (2005). Development of representations, categories and concepts: a hypothesis. In Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2005), Espoo, Finland.
Viola, P. and M. J. Jones. (2004). Robust Real-Time Face Detection, International Journal of Computer Vision, 57(2) 137-154.
von Hofsten, C. (2004) An action perspective on motor development. Trends in Cognitive Sciences. 8(6): 266-272.
Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: Differentiation of multimodal information. Psychological Bulletin, 121, 437-456.

4.3 Learning in Adapt – after the review meeting

In Adapt we planned for the implementation of a cognitive architecture approximately modeled after the proposal of (Doya, 1999) who described how different regions of the brain might be in fact employing different learning paradigms. Although, this is probably an over-simplification when it comes to brain theory, it helps in nailing down the details and in organizing the macro-modules of the architecture. In Babybot we expressly designed a support software framework (http://yarp0.sf.net) to facilitate the implementation of the architecture on a cluster of computers.

We then analyzed which parts need most work and started designing and implementing the relevant machine learning algorithms. We tested some of them on robotic experiments. It is fair to say that we did not integrate all these modules on a single robotic setup. The most advanced form of this integration within Adapt was carried out on the Babybot including though contributions from all partners.

In this respect, results are:

- general plan for the cognitive architecture
- hand segmentation using self-elicited movements
- attention-related work
- semi-supervised learning (including general machine learning applications and Babybot data in particular)

4.3.1 Cognitive architecture

The general plan for architecture included components of reward-based reinforcement learning (RL), self-supervised learning and semi-supervised sensory processing. One difference with respect to Doya's proposal is that we replaced supervised learning for semi-supervised processing which is a situation where supervised learning techniques can be applied without requiring the manual annotation of the data (e.g. input-output pairs are acquired automatically).

The overall structure combines:

- reward-based learning and decisions (action selection and combination)
- self-supervised learning based on learned or hard-wired reflexes

- unsupervised learning and perception of sensory inputs with supervised selection of relevant information

Reward-based learning tries to achieve a goal (see for example the introductory text of Sutton & Barto, 1998). The outcome of learning is useful behavior (this view is consistent with some of the work of Maturana and Varela, 1980 and their concept of autopoiesis). In general this can be thought of as an "action selection" mechanism (Shanahan, 2005) whose role is to combine in various forms sets of primitive or learned behaviors.
Self-supervised learning is a form of supervised learning where the training set is acquired automatically via the specification of a meta-criterion for when to sample input-output pairs. This is possible since in many cases the input-output pairs are directly measurable. One typical example of self-supervised learning is for learning sensorimotor coordination internal models (Kawato et al., 1987).
Statistical learning captures regularities in the data. As a result of learning, it is possible to fill in missing data. An important special case is prediction (future is missing at the time of prediction). Statistical learning can also be used to extract useful features from generic and unlabelled data.
Interaction with the environment is required to do anything useful. Machine learning methods can only learn if there is data. The methods set limits to what can be learned but the data ultimately define what will be learned. The data are generated in interaction with the environment, not by passive observation. Regularities and structure is actively generated; the agent structures its own inputs. One example of the quantification of the structure and regularity in the presence of sensorimotor coordination is described in the paper by Tarapore, Lungarella, and Gomez, 2004 (see list of published papers or the accompanying CD). More details on the concepts are described in the following sections.

4.3.1.1 Reward-based learning

In machine learning, reward is usually defined as a scalar signal which carries information about the success of the agent. The name "reward signal" implies that in response to the signal, the agent will modify its behavior so as to maximize the reward. This is called reinforcement learning. The reason why reward has to be a scalar, a single number, stems from *decision theory*: the reward is basically equivalent to *utility* which is a scalar valued quantity measuring the preference of different outcomes of actions. In the process of making decisions, there can be competing criteria and multi-dimensional evaluations of options. In the end, however, different options need to be ordered, making it necessary to evaluate actions on a one-dimensional scale.
In the process of learning, the agent has to find out the "parameters" which define the behavior. The term parameter is here used in a very wide sense; it could be a synaptic connection, a lookup table, a parameter of a functional mapping, and so on. One way or another, explicit or implicit, there must be parameters which define the behavior.
The process of learning can therefore be viewed as a transformation of information in the teaching signals to information in the parameters of the agent. Since the reward signal is scalar valued it cannot contain a lot of information. Worse still, there is often a significant delay between the relevant behavioral choices and the ensuing rewards and a degree of randomness in reward delivery, meaning that the amount of useful information in the reward signal is necessarily low. Consequently, not many parameters of the agent can be accurately tuned in a

reasonable time. A key element in reward-based learning is exploration. Learning advances through trial and error in much the same way as evolution is driven by random mutations and survival of the fittest. Indeed, natural evolution – and its artificial counterpart genetic algorithms – can be considered as a form of reward-based learning. The term learning usually refers to adaptation that takes place during the life-time of an organism but evolutionary adaptation can be considered as the learning process of species. In that case the reward signal derives from breeding the offspring (or survival of genes in general) and the parameters to be evolved are the genetic code of the organism. The similarity between evolution and reward-based learning led some authors to adopt the term *neural Darwinism*, implying that learning in neural networks can be based on selection of successful neuron assemblies and subnetworks as opposed to gradual adaptation of existing weights. However, it is important to recognize that reward-based learning does not scale up: as the behavioral repertoire expands and the brain of the agent (biological or artificial) grows larger, the relative significance of reward-based learning is bound to decrease.

Experimentally, we showed that an artificial evolutionary system whose task is to track a light source is able not only to evolve and grow a neural network, but can also evolve learning mechanisms. In more recent experiments (Gomez, Eggenberger Hotz, 2004), the evolved neural network was then transferred to a robotic system consistent of a camera mounted on the gripper of a robot arm with results comparable to the ones achieved by the simulation. We further continued to test the evolved controller in the real-world increasing the sensorimotor capabilities and the robot's task as follows: (a) The visual system was enhanced to detect color and movement in the environment and a proprioceptive system was added to have feedback of the arm movements. (b) The number of degrees of freedom (DOF) of the robot was increased from two to three. (c) The position of the cameras was fixed and the same underlying principles were used to teach the robot arm in front of the cameras to move a colored object from an initial location at the periphery of the visual field to the center of it. The arm could solve the task not only for two DOF, but also for three DOF.

4.3.1.2 Anticipation

However difficult reward-based learning or adaptation is, it is still important because it is the most versatile form of learning. It can be used for learning basically any type of mapping or even for evolving algorithms and structures. There are several ways in which other types of learning can be usefully combined with reward-based learning. One way or another, they improve the agent's capability to anticipate: predicting the consequences of actions, future rewards, which actions need to be taken in the future and so on. Instead of random exploration, the learning agent can then select among promising candidate actions. In evolutionary learning this would correspond to targeted mutations, intelligent design.

In a sense, prediction of future rewards gives the agent the capacity to form subgoals because in the process of learning, those states of the world that predict reward can become rewarding in their own right. Ability to form subgoals is particularly important in complex environments where a long sequence of actions needs to be performed in order to affect the primary reward signal. The problem in such cases is that it is very difficult to know which particular actions among all the performed ones are responsible for the reward (the problem of credit assignment). The problem can be alleviated by replacing the primary reward signal by the change in predicted future reward, a secondary reward signal. Since the secondary reward is delivered much earlier than the primary reward, less actions have been already executed and

credit assignment becomes easier. The ability to acquire new goals may seem to be beyond the reach of simple organisms, but in fact, since evolution has modeled the brain, all rewarding stimuli (sweet taste, warmth, etc.) can be considered to be secondary reward signals. The sequence of actions that leads to the survival of genes is so complex that it could not possibly guide learning during the lifetime of an organism. From evolutionary viewpoint, the stimuli that the organisms find rewarding predict the survival of the genes and are therefore learned subgoals. In this case learning has taken place during evolution. The organism can further learn tertiary and higher-order subgoals during its lifetime.

The secondary reward signal depends on the state of the world that results from the actions taken by the agent. It is not necessary to predict in advance what effect the actions will have. This makes the prediction more reliable because uncertainty about the effects of the actions does not compromise learning. However, it also means that then it is not possible to plan. Planning requires a prediction of the future rewards for each candidate action separately. In other words, the prediction needs to be a function of the action in addition to the state of the world. This marks the difference between learning from one's own mistakes and avoiding the mistakes in the first place. Note that the evaluation of candidate actions can be both a conscious, sequential act of considering different options and a subconscious, parallel process, but in each case the success critically depends on the ability to predict future rewards as a function of the action. This component was not implemented in the Adapt cognitive architecture and it is subject of future developments.

4.3.1.3 Self-supervised control

When augmented with other types of learning, reward-based learning becomes useful in goal-directed selection (of actions, attended objects, plans, etc.). Selection can be considered to be a discrete process with a finite number of outcomes as opposed to control which has a continuous valued output. As the amount of information available in reward signals is necessarily limited, reward-based learning and decision making is better suited to selection than control and even then it is useful to limit the rate of decisions as much as possible.

Control cannot usually be based on on-line evaluation of rewards because there is not enough information in the reward signal to control all the actuated degrees of freedom simultaneously. It is, however, possible to gradually learn the parameters of sensorimotor mappings. Note that the term sensorimotor mapping might apply to various types of controllers, such as genetically determined, hardwired sensorimotor mappings but since they too have been “learned” by evolution, we will call any automatic behavior *sensorimotor mapping* whether they are in place at birth or learned through trial and error.

It would usually take a very long time to accurately tune the parameters of a new mapping by using reward-based learning, but again it is possible to use other learning schemes to perfect the mapping once a rough initial reflex exists. The main idea is to use the initial reflex as a teacher for a more refined sensorimotor mapping that anticipates the output of the teacher. Through practice, the second controller can learn to execute the motor commands that the teacher reflex would have executed in the end. Initial slow, inaccurate and possibly jerky movements become smooth, swift and accurate. Note that the goal of this type of learning is usually not simply to predict the teacher reflex but to make it unnecessary. In other words, the initial reflex provides an error signal, not the target of prediction. During learning, the anticipatory sensorimotor mapping may take over almost completely and the initial reflex will no longer be needed in the execution of the movement. Usually the initial reflex is still needed

in order to keep the learned mapping from diverging. Also, if something changes, the teaching reflex which has been lying dormant will again take its original role in pushing the anticipatory sensorimotor mapping in the right direction (in certain pathologies initial reflexes become apparent again).
A well-studied example of this type of learning is gaze stabilization. In this case the teaching reflex is the so-called optokinetic reflex (OKR) which maps retinal optic flow to eye-muscle movements. A typical problem in such feedback system is that the required gain is typically unknown and very difficult to hardwire because it depends on the exact morphology of the eyes and changes throughout the life. Second, since there are substantial sensory delays in the computation of retinal optic flow, the control is stable only when the gain is relatively small. These shortcomings can be compensated by a second, anticipatory reflex which adapts in the guidance of the first reflex. The second reflex can use information from all senses and even about the agents own planned movements. One of the most important predictors of required gaze stabilization movements is the signal from the balance (vestibular) organ. The second reflex learns to evoke eye movements in response to the vestibular signals before any information from the eye arrives. This is called the vestibulo-ocular reflex (VOR).
If the second reflex, VOR, has too small gain, there will be residual optic flow and the first reflex, OKR, activates. This instructs VOR to increase its gain. If VOR has too large gain, there will be, on average, optic flow in the opposite direction, activating OKR in the opposite direction and instructing VOR to decrease its gain. The system learns smooth, swift and stable control as long as the adaptive reflex takes its teaching signal from the activity of the teaching reflex over a short time window in the future (this is exactly the mechanism implemented into the Babybot).
The above learning scheme can be called self-supervised because the teaching signal is internally generated. From theoretical viewpoint, however, the information for learning derives mainly from the environment and the dynamics of the body. The teaching reflex only specifies the direction of adaptation and the rest will be learned through interaction with the environment. The problem of learning a controller is often considered an inverse problem in traditional control theory. The forward problem is how the control signal affects the state of the system (and thus the sensory inputs) while here the problem is to map sensory inputs to motor output. The teaching controller can be thought of as a feedback controller which solves the inverse iteratively. The second controller is a forward controller which learns to imitate the solution obtained by the first controller.
In classical control theory, feedback controllers have usually been augmented by predictors which anticipate the sensory inputs. In the case of gaze stabilization, for instance, it would have been possible to predict the retinal optic flow signals from other signals such as the vestibular signals and the motor commands sent to the eye muscles and then use these predictions as the input for OKR. This control strategy called the Smith predictor can alleviate problems related to sensory delays but it does not directly address the problem related to unknown gains. There are several papers on the Babybot and the application of self-supervised techniques to sensorimotor coordination (see for a review http://www.liralab.it).

4.3.1.4 Sensory processing

Unsupervised learning is based on finding structure in data. From a theoretical perspective, unsupervised learning is actually quite close to supervised learning because in both cases learning is based on statistical dependencies between variables as opposed to maximizing

utility as in reward-based learning. In practice the similarities mean that most supervised learning methods can be used for unsupervised learning and vice versa. In supervised learning, the data are divided to inputs and outputs. The statistical dependencies are used to find a mapping from inputs to outputs. In contrast, in unsupervised learning there is no division and all dependencies are usually considered important. Given enough resources, unsupervised learning therefore seems to be preferable because any part of the data can be predicted or reconstructed given any other part of the data. It is not necessary to decide in advance which parts of the data are predicted from the other part. In practice there are almost always limited resources: there are many more dependencies in the data that can be modeled. The predictors discussed so far can be learned in a supervised manner. The outputs or targets of prediction can be internally generated rewards or reflexes or selected external stimuli as in the Smith predictor. The inputs can in principle be any neural activity carrying information which makes the prediction possible. In such a supervised learning method, the amount of information available for learning can be orders of magnitude higher than in reward-based learning. For prediction of reward, the information available in the target signal is obviously the same as in reward-based learning but for prediction of reflexes, there is roughly speaking as many times more information as there are reflexes.

Nevertheless, supervised learning has limitations. Perhaps the most severe problem is the difficulty to form new concepts and categories. Multi-layer models can in some sense learn internal representations in a supervised manner because the neural activities in the hidden neurons, those between the inputs and the outputs can be considered representations. However, learning deep multi-layer models is notoriously difficult. Since the teaching signals for learning derive from the outputs, it is difficult to learn anything useful close to the inputs if the mapping is complex. For instance, it is very difficult to learn useful low-level visual features if the learning signal derives from motor reflexes.

Unsupervised learning is often used as a preprocessing stage, in tasks analogous to finding useful low-level visual features. Finding efficient representations which compress the essential information of the data into a smaller representation simplifies the second, supervised or reward-based learning stage. Interestingly, it has been shown that the representations found in the first stages of processing in mammalian visual cortex can be learned from natural images in unsupervised manner. This shows that, indeed, the cortex seems to form efficient representations which rely on the statistical dependences in the sensory inputs (Doya, 1999).

However, unsupervised learning methods have had much less success in finding meaningful higher-level representations. Ultimately the goal of these representations is to facilitate the prediction of rewards and reflexes; the representations bridge the gap between raw sensory stimuli and behaviorally relevant predictions. The gap can be between modalities (e.g., retinal stimuli and motor reflexes) and time instants. The problem with unsupervised learning is that since all dependencies are considered equally important, most of them will turn out to be useless. If the preprocessing stage finds too many features (too high-dimensional representation) from the inputs, it actually makes the learning task in the second stage more difficult.

In summary, unsupervised learning works well on the input side and supervised learning on the output side but methods that would work in between, for more abstract concepts and categories, which are far both from inputs and outputs, are still needed.

In ADAPT, we developed semi-supervised learning methods. They combine aspects of supervised and unsupervised learning; they utilize the statistical structure of the inputs to find

efficient representations but also use supervisory signals to bias learning and to select those candidate representations which seem to have most information relevant to predicting the outputs (reward and motor information).
Selection of information is needed dynamically, too. Different tasks, for instance grasping an object or navigating, require different types of information, different input-output mappings. Attention is a process which selects relevant sensory inputs and, more generally, can also be understood to include different input-output mappings if selection of the outputs is taken into account, too. Since the dynamic selection of sensorimotor mappings is very relevant to the topics studied in ADAPT, part of machine-learning research effort was dedicated to developing mechanisms for attention. In all machine learning research in ADAPT, we have kept in mind that sensory processing serves specific behavioral goals, is needed for the prediction of rewards and reflexes, and is part of an agent which develops through interaction with its environment. One example of sensory processing that uses unsupervised learning is reported in this document in section 4.2.6.

4.3.1.5 Attention

Attention is a selective process by which certain sensory stimuli gain access to higher levels of processing (e.g., working memory, consciousness). There are several reasons why attention is useful. First, because there are limited representational resources, objects need to be represented as combinations of individual features. Such coding is very efficient if only one object is present. However, if there are several objects at once, there has to be an indication of which features belong together. This so-called binding problem can be solved by representing one object at a time and alternating through the objects in the scene thus giving each object a short time slot in the representation. Effectively, each object is attended to for a short interval of time.
Second, the body cannot engage in several conflicting movements. It is, for instance, not possible to grasp two objects simultaneously and therefore it is necessary to select a single target. Attention is therefore not just about sensory selection but involves competing sensorimotor networks. Motor control becomes significantly easier if distractors do not enter the representations which guide the motor system.
Third, reward-based learning is possible only if the rate of decision making is low enough. Several automated movements can be ``running in the background'' but the number of conscious, voluntary movements must be very limited if something is to be learned from their ensuing rewards. In the development of attention, reward-based learning may be needed for learning which types of sensory stimuli need to be attended to.
The selection of attended objects is affected both by bottom-up sensory cues and by top-down expectations. Both aspects have been considered in ADAPT as described in section 4.2.1.

4.3.1.6 Semi-supervised learning

The task of sensory processing is to provide useful information for motor control and decision making. Senses (vision, audition, proprioception, etc.) can be considered as inputs and motor signals or reward predictions of the outputs. Overall, the task is therefore a supervised learning one, finding a mapping from inputs to outputs.
As stated earlier, the problem with supervised learning is that, while there are many methods that work well for simple mappings (information in the inputs is “close” to the outputs), it is very difficult to learn complex mappings. An unsupervised pre-processing stage can help by

providing sensible candidate representations which make the required mapping easier. The problem is that unsupervised learning has no "sense of direction", that is, it does not use information about outputs in the development of representations. The solution arrived at in ADAPT was to combine an unsupervised feature extraction stage with a supervised selection stage. These stages together form a semi-supervised module which can be linked together to span the mapping from inputs to outputs.

In principle these modules could have been designed to be symmetric in the sense that it does not matter which side is input and which is output. However, it is reasonable to assume that the representations linking different low-level areas, such as primary visual and primary motor areas, should be more abstract or invariant. For instance, the concept of a chair is useful because it links low-level visual features (how the chair looks like) with low-level motor features (which muscle-movement sequences correspond to sitting) in many different circumstances. Therefore, the semi-supervised modules were designed to form a mapping between a more specific, low-level representation and a more invariant, higher-level representation.

The following sections introduce the basic ingredient of the semi-supervised module. In short, the idea is to use top-down predictions to select which low-level primitives are chunked together to form a higher-level primitive. Information flows in both directions in the module. Since the higher-level representations do not exist in the beginning of development, learning in the bottom-up flow resembles unsupervised learning. Once the higher-level representations exist, the top-down direction can adapt in a supervised manner. Semi-supervised learning has been described earlier in this document and it is also reported in Sarela and Valpola (2005).

4.3.1.7 Unsupervised feature extraction

Unsupervised learning makes use of regularities, dependencies, redundancy or other such structure found in the data and builds a more compact or otherwise useful representation where the prominent information is more readily available. Many of the techniques use neural networks or other such models which consist of a number of simple interconnected nodes. Unsupervised learning in such models usually requires some type of competition which assures that during learning, all neurons or nodes start to represent different aspects or features of the inputs. The neurons (or nodes) usually have adaptable parameters which describe the type of input which best activates the neuron. Learning is Hebbian which means that each time a neuron activates, its parameters are adapted towards the input.

Exact details of the methods vary but usually the above mentioned elements can be found. For instance, competition is not necessarily explicit but one way or another excess redundancy in the resulting representation needs to be avoided. If all neurons would activate independent of the activities of the other neurons, they could all get adapted towards the same inputs and by time all of their parameters could adapt to the same values.

In ADAPT, we used independent component analysis (ICA) techniques which can find simple primitives with which the inputs can be represented compactly. For instance, it has been shown that when given natural images as inputs, ICA can find simple edge features, primitives which can be used to represent natural images compactly. There are other methods which find similar features but ICA has the advantage that it can process a lot of information in parallel. In ICA, neurons compete with each only if they represent similar information. This is important since raw sensory inputs have a lot of information which cannot be summarized in the activity of a single neuron.

Without knowing the geometry of the sensor array, ICA can find meaningful image features. The learned edge features are useful primitives which reflect the statistical structure of the inputs. Interestingly, there was also a separation between color and shape features. The response properties of these neurons (which can also be called detectors, primitives, nodes, etc.) resemble those of so-called simple cells found at the first stages of processing in the primary visual cortex.

Mathematically, each neuron computes a linear projection. If that would be the end of story, it would not make any sense to stack many modules together: two consecutive linear mappings can always be represented by one linear mapping. In order to form complex mappings which are needed in tasks such as visuomotor control, nonlinearities are needed. We used a fixed activation function which maps the input projection into output. More specifically, we used a nonlinear function which asymptotically saturates to zero for negative inputs and approaches an identity mapping for positive inputs; reported in one of the papers (see Sarela and Valpola, 2005). Many other types of nonlinearities could have been used, too, such as sharpening of tuning curves for competing neurons. This would correspond to the perceptual phenomenon known as hyperacuity.

4.3.1.8 Selection by prediction

In bottom-up processing, the first stage extracts features, primitives which efficiently represent the inputs. The next step combines these features into more invariant features. Several primitives with the same meaning are chunked into a more abstract primitive. A classical example of this stage is the transition from simple cells to so-called complex cells in the primary visual cortex. While the simple cells are sharply tuned to an edge with a certain orientation and location, complex cells have much less specific about the exact location. In other words, complex cells are more invariant to translation. In practice, a complex cell achieves this invariance by sampling many simple cells which the same orientation but slightly different location tuning. The transition from simple to complex cells in the primary visual cortex is just the first step in achieving abstract representations, but it seems that the basic principle works all the way from low-level sensory representations to complex categories and concepts. Each stage combines lower-level primitives into more abstract, higher-level primitives which then act as the new building blocks for the next level.

The benefit of increased abstraction is the ability to generalize. For instance, if an agent has translation, scaling and rotation invariant visual representation, it will be able to recognize the object in many different situations after having learned how the object looks like in one particular situation. Abstract representation allow for making longer-term predictions and planning and coordination behavior in terms of more complex sequences of actions.

The generality of abstract representations is achieved by losing information. Complex cells, for instance, lose information about the exact location of edges. Since losing information will obviously generate problems, one might argue that it would be better to retain all information and let the later stages decide what they need. The problem with this approach is that since the world is extremely complex, the later stages would be overwhelmed by the amount of information. Any useful bit would be lost in the flood of irrelevant trivia.

The need to lose information brings forward two obvious problems. First, which information should be retained? This is where the top-down flow of information enters the stage and we will discuss this in detail. Second, what if the information is needed after all? The answer is that if need be, the information is still there, in the lower levels. Visuomotor control, for

instance, will probably need much lower-level information than planning. To serve such needs, there can be direct associations between lower levels in addition to the associations at the higher levels. The mechanisms of attention will likely contribute to the channeling of information, but will not consider this problem here.

In ADAPT, we focused on solving the first problem, how to select the information that should be retained. The solution was to use top-down signals (or, more generally, context) to pinpoint those primitives which carry important information and to decide which primitives have the same "meaning" when they are grouped together. The basic principle is very similar to the hand-segmentation example shown earlier, where motor signals were used to identify the pixels which correspond to the hand. In that case selection was dynamic, but the same principle works on longer timescales, too. At the heart of the selection process is a predictor. The targets of prediction are the bottom-up signals (the activity levels of the neurons or primitives) and the inputs are the top-down signals. In short, those primitives will be selected for later processing whose activity can be predicted. The algorithmic details of selection based on prediction are given in the next section. Here we describe the basic idea and properties of the method.

The main difference between this type of learning and traditional supervised learning is that the supervisory signals do not determine the representations which are developed; they only guide the selection and grouping of information. Since selection requires much less information than determining the synaptic weights of the primitives, less accurate top-down suffices and overall learning is semi-supervised rather than supervised. While supervised learning works in practice only close to the outputs (i.e. propagation of error signals through multiple processing levels is difficult), semi-supervised learning can always fall back to unsupervised learning which works closer to the inputs.

Usually, when learning predictors, the choice of input variables is not particularly critical because the system will learn to use those inputs that turn out to be useful. To some extent this is the case here, too: there can be completely irrelevant input signals to the predictor and the system still works. However, now the predictions are used to learn the selection and grouping of the bottom-up signals and therefore the input signals for the predictor do influence the resulting representations.

Theoretical justification for using top-down predictions for guiding learning is that mutual information is symmetric: if the top-down signals contain information about a set of bottom-up signals, then the opposite must be true, too. The predictions can therefore be used for selecting relevant features. Moreover, it makes sense to group together primitives which cannot be distinguished based on the prediction. For instance, if top-down signals are able to predict the orientation of an edge but not its exact location, it is useful to group together primitives which respond to a particular orientation but in different locations. Likewise, if only the location but not the orientation can be predicted, primitives with a given location but different orientations can be grouped together.

Since the resulting representations are supposed to be useful for predicting motor movements and rewards, these modalities should have a privileged role in guiding the development of sensory representations. The input for the predictor is therefore not necessarily just "top-down signals", but more generally any signals which may need to be predicted.

Low-level motor signals will probably not have much in common with low-level visual inputs (with the exception of oculomotor signals). It therefore makes sense to abstract motor signals before using them for guiding visual signals. For this purpose, that part of motor information

should be represented which can be predicted by visual information because, due to the symmetry of mutual information, that part is also most useful for guiding visual learning.

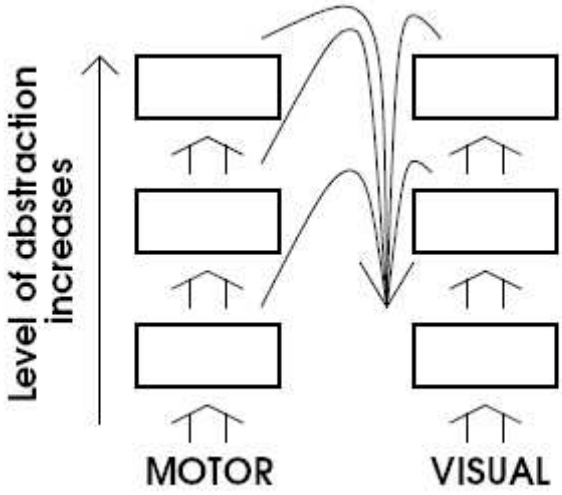


Figure 22: A schematic depiction of a visuomotor hierarchy. Motor and visual signals enter the system from below. The boxes represent semi-supervised learning blocks. The thick arrows denote bottom-up processing and the thin arrows denote supervisory top-down and contextual signals. For clarity, supervisory signals are shown only for the visual stream but the system can be completely symmetric, channeling supervisory signals from visual to motor stream.

Figure 22 shows one possible architecture where two hierarchies, visual and motor, have been built from semi-supervised learning modules. Only the supervision from motor to visual hierarchy is shown for clarity but the architecture can be completely symmetric. The supervisory signals crossing the hierarchies are pulling the representations together, helping to bridge the gap between visual and motor information. The gap should be narrower at the higher levels of abstraction where the representations are more invariant.

In real systems there would usually be more than two modalities and the hierarchy would then include more than one stream. In that case it is also possible and perhaps sensible to make branches to the streams, directing each branch towards different modalities. For instance, if there were three modalities, visual, motor and reward, then it might make sense to divide the visual stream into a branch which receives supervision from the motor stream and another branch which is supervised by the reward stream. The human visual system consists of tens of areas whose specialization may partly be influenced by such segregation of top-down information received by different areas.

4.3.1.9 Decorrelation and denoising source separation

In order to build a semi-supervised learning module, the mechanism by which top-down predictions can select and group bottom-up primitives is needed. The other parts (learning primitives and making predictions) can be handled by existing machine learning techniques for unsupervised and supervised learning, respectively. In ADAPT, we developed a new framework which extends existing independent component analysis algorithms such that they can use supervisory information. In the framework, source separation algorithms are

constructed around a denoising function and the framework was accordingly named denoising source separation (DSS).

DSS framework is very flexible because many different types of denoising procedures can be used. If only the bottom-up signals are used in denoising, for instance by utilizing their temporal structure or marginal distribution, the resulting algorithm is an unsupervised source separation technique (e.g., ICA). Conversely, if external supervisory signals alone are used, the resulting algorithm is purely supervised. In between, there are many options to use both bottom-up and external signals for denoising and DSS algorithms therefore span the whole spectrum from unsupervised to supervised algorithms with many possibilities for semi-supervised algorithms in between.

The key idea in DSS framework is to decorrelate and normalize the inputs in order to allow the denoising procedure to guide the development of features (i.e., in ICA terminology, the extraction of sources). Decorrelation and normalization of inputs result in a distribution where projections of inputs yield the same variance for all projection directions. For this reason the procedure is also known as sphering. After the sphered inputs are denoised, different projection directions again yield components with different variances. By denifition, denoising removes relatively more of the unwanted noise than signals. Therefore interesting components can be identified from their higher variance after sphering and denoising. In practice this can be done with principal component analysis (PCA). There are various methods for implementing PCA, among them biologically motivated neural PCA which uses simple Hebbian learning.

The intuitive justification of DSS is that different types of features have often quite different strengths in the sensory inputs and without sphering the strongest features would dominate, making it very difficult to use weaker but potentially important features. For instance, in visual input, global features such as the overall illumination or gross differences in illumination across the visual field are typically several magnitudes stronger than finer details. Sphering gives all features an equal chance of getting selected, making it possible to use other criteria than the strength of the feature for selection.

In the brain, mechanisms suited for decorrelation and normalization are common. They are found in sensory periphery (e.g., retina), thalamus and cortex to name a few examples. Figure 15 visualizes a sphering transformation for image patches taken from natural images. It turns out that the components of this transformation resemble on-center-off-surround cells found in the retina and thalamic relay cells between retina and neocortex. The cells respond to features in different locations of retina but only if they are local. Such spatial high-pass filtering is essentially a sphering transformation for natural images. Normally localized features are weaker than global ones, but after the transformation all features have similar strengths.

After sphering has equalized the strengths of all input features, minute modulations of learning rates are able to guide Hebbian learning. In its simples form, Hebbian learning amounts to changing the synaptic weights proportional to the product of input and output activities (firing rates of neurons). One way to modulate learning rate is therefore to modulate the output activities. If, for instance, the activities are low-pass filtered in time, learning is biased towards extracting components which have slow dynamics because such components will have the highest variance after low-pass filtering. Moreover, the filtering can be very modest, only changing the activities by a tiny fraction, because after sphering any projection is equally good solution for PCA. It is useful to think about the modulation of the activities as denoising, because the algorithm extracts the features that are emphasized by modulation.

Ordinary ICA can be implemented by using relatively vague information about the statistics of the bottom-up signals to denoise the activities. In many real-world cases the activities are sparse (technically, they have a high kurtosis). Compared to Gaussian distribution, such distributions have more samples close to zero but also more values which are much higher than the standard deviation would predict. Edges in natural images are a typical example of sparsely distributed feature: there are often large smooth image areas almost without edges but then there are sharp boundaries where an edge-detector gives a strong response. In DSS framework, such sparsely distributed features can be extracted by using so-called shrinkage denoising: activities close to zero are put even closer to zero but activities departing significantly from zero are left untouched.

For semi-supervised learning, we have used top-down predictions of activities to modulate their amplitude. During learning those components get extracted whose variance will be highest. With this type of modulation, the variance is highest when the top-down prediction matches bottom-up inputs. The learning process is therefore driven towards finding primitives which are best predicted from top-down inputs. In order to enforce selection and grouping of primitives, the output dimension is required to be smaller than the number of bottom-up primitives extracted in the unsupervised learning stage. Due to sphering of the bottom-up inputs (activities of the primitives), weak modulation is enough, letting bottom-up inputs determine the output activities almost completely but still letting top-down predictions guide learning.

DSS has turned out to be a versatile framework and it has found applications in various fields. Here we focus on experiments about robot grasping (as in D5.5 and the presentation in the review meeting). Technical details are given in (Sarela and Valpola, 2005).

4.3.2 Future perspectives

Development of a complete cognitive architecture is a grand challenge and a very active research topic in machine learning. We have contributed with a unifying perspective and specific methods inspired by our developmental viewpoint. We have began experimenting with the various parts of the system but final integration of the sub-systems was out of the scope of the project and will be an important focus of future research.

Combination of attention and semi-supervised learning should be easy because the dynamic selection of active features in attention uses basically the same mechanisms as development of features though on different timescales. The main difference between the two is the magnitude of effect that top-down signals have on the activation of features. Attention requires (or rather is equated with) stronger modulation than learning but nothing prevents learning from working with strong modulation, too. The similarities between the algorithms for attention and learning suggest that decorrelation and normalization, which allow learning to select on weak input features if they turn out to be relevant, might also be useful in attention, allowing attention to be grabbed by weaker bottom-up stimuli which nevertheless have high top-down relevance.

There is also a need for further experiments about integrating sensory processing (learning and attention) with motor control. It is straight-forward to use sensory input features for the predictors used in self-supervised and reinforcement learning of motor control, but feedback from motor and emotional systems to sensory systems is critical for useful guidance of sensory attention and learning. The architecture outlined in the previous section is designed for this type of supervision.

Finally, a cognitive architecture will also need other ingredients which we did not discuss here. At least a type of episodic memory is required, making it possible for the agent to remember routes, places and events from a single experienced episode. In humans, the hippocampus seems to be a critical component of episodic memory. Also, mechanisms for planning are required. In the previous section, top-down prediction was an important ingredient in attention and learning. The same prediction can also support imagination, making it possible for the system to predict the consequences of actions without actually taking the actions. In order to be useful, there needs to be a gating mechanism which selects the plans which are actuated. In humans, basal ganglia seem to be the system controlling this type of selection both in external motor and internalized cognitive control.
For further references and technical details on these preceding sections please see the published papers listed in section 5.1.

4.3.3 Relevance to psychology

The physical world is composed of two kinds of multimodal entities: objects and agents. Both provide different though co-occurring sensory messages coming from a unique source. Attending simultaneously several stimuli from different sensory channels might lead to a chaotic representation of the world and consequently of the self. Paradoxically, it leads to the 'sense of being there' as a unifying process.

ADAPT interdisciplinary meetings and associated interdisciplinary interactions have led to stress the common aspects ruling the interaction with these two kinds of multimodal entities through the mechanisms of intermodal transfer.
Concerning the interaction with multimodal objects, behavioral experiments with human infants have tested the hypothesis of a primitive unity of senses that can be theorized, modeled and designed by AI partners via artifacts. Grasping has been studied as a main and basic feature of visual-tactile intermodality in newborns and robots. Conversely, the study of the robot's reaching and grasping was paralleled with infant studies. Concerning interpersonal interaction, early detection of agents as coherent multimodal entities was tested through a device that was especially designed to violate the natural synchronization of co-occurring messages coming from a unique source.

Mutual influence due to interdisciplinary collaboration for the ADAPT program has led to innovative procedures and design. They have favored the interest of ADAPT AI researchers for human cognitive development as shown by their continuation with a new project called RobotCub (Unit-E5 Cognition), and the interest of ADAPT psychologists for AI models and artifacts as shown by their participation to a Specific Targeted Project (MATHESIS) concerning observational learning in cognitive agents. Thus ADAPT appeared not only to kick off a multicentric approach of the sense of presence via interaction with objects and agents, but also to open new interdisciplinary avenues for most participants.

4.3.4 Further references

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks, 12, 961-974.

Kawato, M., Furukawa, K., & Suzuki, R. (1987). A hierarchical neural network model for control and learning of voluntary movement. Biological Cybernetics, 57, 169-185.

Maturana, R. H., & Varela, F. J. (1980). Autopoiesis and Cognition: The Realization of the Living. Dordecht: D.Reidel Publishing Co.

Shanahan, M. P. (2005). Emotion, and imagination: A brain-inspired architecture for cognitive robotics. Paper presented at the AISB 2005 Symposium on Next Generation Approaches to Machine Consciousness.

Sutton, R. S., & Barto, A. (1998). Reinforcement Learning: an Introduction. Cambridge: MIT Press.

4.4 European-level implications of Adapt

Adapt workplan does not include a real marketing or product development phase and as such the implications are mainly scientific rather than technological. On the other hand, robotics might become an extremely important market in the near future, and it is exactly in this direction that many Japanese companies started to get prepared. This market might span applications ranging from the assistance to the elderly, to medical (surgical, diagnosis, assistance), and to personal mobility or remote presence/remote operation. The applications, once the robots are sufficiently reliable, are only limited by fantasy.

On the other end of the spectrum, "brain sciences" research is likewise crucial. Beside the specific link we exploited in Adapt (i.e. linking brain and AI), understanding the brain will have a beneficial fallout on the quality of life with respect to mental illness for example, but also to rehabilitation. Media are the next possible target: new applications specifically designed to convey immersive and realistic experience could be developed by starting from the knowledge of the functioning of perception. In this sense, we can see where Adapt's contribution lays in practice.

These two big new developments, brain sciences and robotics, might one day converge into the understanding of consciousness but, in the meanwhile, there is a clear path that leads to the design of applications exploiting the results of the general R&D effort in these fields. Employment would receive then the most immediate benefits.

5 List of deliverables

Number	Title	Type	Due month
D1.1	Project presentation	Docs + web site	3
D1.2	Dissemination and use plan	Document	6
D1.3	Management report	Document	6
D1.4	Periodic progress report Y1	Document	12
D1.5	Management report	Document	18
D1.6	Periodic progress report Y2	Document	24
D1.7	Management report	Document	30
D1.8	Technology implementation plan	Document	32
D1.9	Final report	Document	36

D2.1	A tentative theory of intentionality and the sense of being there	Document	7
D2.2	A validated theory of intentionality and the sense of being there	Document	36
D2.3	A common psycho-physical vocabulary	Document	36
D3.1	Definition and implementation of a human-like robotic setup	Document	12
D3.2	Hardware and software in place to run experiments on changing morphologies (e.g. changing resolution and motor precision)	Prototype	15
D3.3	A set of formal methods for the analysis of the interplay of morphology, materials and control	Document	30
D4.1	Definition of experimental paradigm	Document	12
D4.2	Definition and implementation of setup for the investigation on child development	Prototype	12
D4.3	Results of behavioral experiments with the babies	Document	30
D4.4	Results of behavioral experiments with the robot	Document	30
D5.1	System's architecture specifications and design	Document	6
D5.2	Basic unit design and implementation	Prototype	9
D5.3	Initial implementation of the integration model	Prototype	12
D5.4	Initial experiments with multiple sensory modalities integrations (delayed from Y2)	Document	18
D5.5	Validation of multisensory representations	Document	33

Submitted [yellow]. Relative to the last 12 months [gray].

5.1 List of publications

2003

Giorgio Metta, Giulio Sandini, Lorenzo Natale, Riccardo Manzotti. Artificial Development Approach to Presence. In Presence 2003. Aalborg, DK. Oct 6-8th, 2003.

Max Lungarella, Giorgio Metta, Rolf Pfeifer, Giulio Sandini. Developmental Robotics: A Survey. Connection Science. 15(4), pp. 151-190. 2003.

Streri, A., & Gentaz, E. (2003). Cross-modal recognition of shape from hand to eyes in human newborns. Somatosensory & Motor Research, 20(1), 11-16.

2004

Gómez, G. and Eggenberger Hotz, P. "Investigations on the robustness of an evolved learning mechanism for a robot arm". In proceedings of the 8th conference on Intelligent Autonomous Systems, 2004.

Tarapore, G., Lungarella, M. and Gómez, G. "Fingerprinting Agent-Environment Interaction Via Information Theory". In proceedings of the 8th conference on Intelligent Autonomous Systems, 2004.

Gómez, G. and Eggenberger Hotz, P. "An Evolved Learning Mechanism for Teaching a Robot to Foveate". In proceedings of AROB 9th Artificial Life and Robotics, Jan 28th-30th, 2004, Japan.

Valpola, H. (2004). Behaviourally meaningful representations from normalisation and context-guided denoising. Technical Report, Artificial Intelligence Laboratory, University of Zurich.

L.Natale, G.Metta, G.Sandini. Learning haptic representation of objects. In International Conference on Intelligent Manipulation and Grasping. Genoa - Italy July 1-2, 2004.

Eggenberger Hotz, P. and Gómez, G. (2004). "The transfer problem from simulation to the real world in artificial evolution". In Bedau, M., Husbands, P., Hutton, T., Kumar, S., and Suzuki, H. (Eds.) Workshop and Tutorial Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (Alife IX). pp. 17-20

Gómez, G., Lungarella, M., Eggenberger Hotz, P., Matsushita, K., and Pfeifer, R. (2004). "Simulating development in a real robot: on the concurrent increase of sensory, motor, and neural complexity". In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C. (Eds.) Proceedings of the Fourth International Workshop on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems. Lund University Cognitive Studies, 117. pp. 119-122.

H. Valpola and J. Särelä. Accurate, fast and stable denoising source separation algorithms. In Proceedings of the 5th International Conference on Independent Component Analysis and Blind Signal Separation (ICA 2004), Granada, Spain, pp. 65-72, 2004.

Streri, A. & Gentaz,E. (2004). Cross modal recognition of shape from hand to eyes and handedness in human newborns. Neuropsychologia, 42, 1365-1369.

Nadel, J. et al.. (2004). Toward communication: first imitations in infants, children with autism and robots. Interdisciplinary Journal of interaction studies, 1, 45-75.

2005

L.Natale, G.Metta, G.Sandini. A Developmental Approach to Grasping. In Developmental Robotics. A 2005 AAAI Spring Symposium. March 21-23rd, 2005. Stanford University, Stanford, CA, USA.

F. Orabona, G. Metta, G. Sandini. Object-based Visual Attention: a Model for a Behaving Robot. In 3rd International Workshop on Attention and Performance in Computational Vision within CVPR, San Diego, CA, USA. June 25, 2005.

L.Natale, F. Orabona, G. Metta, G. Sandini. Exploring the world through grasping: a developmental approach. In 6th CIRA Symposium, Espoo, Finland, June 27-30, 2005.

Gomez, G., Lungarella, M. and Tarapore, D. (2005) Information-theoretic approach to embodied category learning. Proc. of 10th Int. Conf. on Artificial Life and Robotics. (AROB 10): Proceedings of the 10th Int. Symp. on Artificial Life and Robotics, Beppu, Oita, Japan. pp. 332-337.

H. Valpola. Development of representations, categories and concepts--a hypothesis. In Proceedings of the 6th IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA 2005), Espoo, Finland, 2005.

J. Särelä and H. Valpola. Denoising source separation: a novel approach to ICA and feature extraction using denoising and Hebbian learning. In AI 2005 special session on correlation learning, Victoria, British-Columbia, Canada, pp. 45-56, 2005.

J. Särelä and H. Valpola. Denoising source separation. Journal of Machine Learning Research, 6:233-272, 2005.

NADEL, J., SOUSSIGNAN, R., CANET, P., LIBERT, G., & GERARDIN, P. (2005). Two – month-old infants's emotional state after non-contingent interaction. Infant Behavior and Development

NADEL, J., PREPIN, K., & OKANDA, M. (2005). Experiencing contingency and agency: first step toward self-understanding? Interaction studies: Social Behaviour and Communication in Biological and Artificial Systems, 6, 3, 447- 462.

Manzotti R, An outline of an alternative view of conscious perception, TSC2005, Copenhagen, 2005.

Manzotti R. Villamira M. The "What" problem: the emergence of new goals in a robot, 6th CIRA Symposium, Espoo, Finland, June 27-30, 2005.

Manzotti R., Tagliasco V., From "behaviour-based" robots to "motivations-based" robots, in "Robotics and Autonomous Systems", 51 (2-3), 175-200, 2005.

Manzotti R. Tagliasco V, The What Problem: Can a Theory of Consciousness be Useful?, in "Yearbook of the Artificial", (3), Peter Lang (Ed.), Berna, 2005.

C. Beltran-Gonzalez, G. Sandini. Visual Attention Priming Based on Crossmodal Expectations. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2005), Edmonton, Alberta, Canada, August 2-6 2005

In press, submitted, etc.

Gomez, G., Hernandez, A., Eggenberger Hotz, P., and Pfeifer, R. (in press). An adaptive learning mechanism for teaching a robot to grasp. To appear in Proc. of AMAM 2005.

Tarapore, D., Lungarella, M. and Gomez, G. (2005). Quantifying patterns of agent-environment interaction. Robotics and Autonomous Systems (accepted for publication).

L. Natale, F. Orabona, G. Metta, G. Sandini. Exploring the world through manipulation: a developmental approach. Submitted to International Journal of Humanoid Robotics. 2005.

L. Natale, F. Orabona, F. Berton, G. Metta, G. Sandini. From sensorimotor development to object perception. Submitted to the International Conference of Humanoids Robotics. 2005.

G. Metta, P. Fitzpatrick, L. Natale. YARP: yet another robot platform. Submitted to International Journal on Advanced Robotics Systems Special Issue on Software Development and Integration in Robotics. 2005.

SOUSSIGNAN, R., NADEL, J., CANET, P., & GERARDIN, P. (submitted). Sensitivity, not tiredness, accounts for 2-month-olds emotional changes during non-contingent maternal behavior.

PREPIN, K., SIMON, M., MAHE, A-S, CANET, P., SOUSSIGNAN, R., & NADEL, J. (submitted). The effect of maternal mismatch between face and voice in 6-month-old infants.

REVEL, A., & NADEL, J. (in press). How to build an imitator. In K. Dautenhahn & C. Nehaniv (Eds), Imitation in animals and artefacts. Cambridge: Cambridge University Press.

Pascal Kaufmann and Gabriel Gomez (submitted). Developing Virtual Neural Tissue for Real-Time Applications: Growth and Dynamics of Spiking Neurons. Submitted to Neural Networks.

Pascal Kaufmann and Gabriel Gomez (submitted). Brains for Robots: Virtual Neural Tissue for Real-Time Applications. Submitted to the 9th International Conference on Intelligent Autonomous Systems (IAS-9).

6 Potential impact of project results

Questions about project's outcomes	Number	Comments
1. Scientific and technological achievements of the project (and why are they so ?)		
Question 1.1. Which is the 'Breakthrough' or 'real' innovation achieved in the considered period	N/A	
2. Impact on Science and Technology: Scientific Publications in scientific magazines		
Question 2.1. Scientific or technical publications on reviewed journals and conferences	14 +10 under review	Partners: ALL[††] See section 5.1 (complete table for the three years of the project)
Question 2.2. Scientific or technical publications on non-reviewed journals and conferences	0	Title and journals/conference and partners involved[‡‡]
Question 2.3. Invited papers published in scientific or technical journal or conference.	0	Title and journals/conference and partners involved[§§]
3. Impact on Innovation and Micro-economy		

[††] Please submit these information in an 'excel' sheet with title of publication/authors/journal or conference/date etc.
[‡‡] Please submit these information in an 'excel' sheet with title of publication/authors/journal or conference/date etc.
[§§] Please submit these information in an 'excel' sheet with title of publication/authors/journal or conference/date etc.

A - Patents		
Question 3.1. Patents filed and pending	0	When and in which country(ies): Brief explanation of the field covered by the patent:
Question 3.2. Patents awarded	0	When and in which country(ies): Brief explanation of the field covered by the patent* (if different from above):
Question 3.3. Patents sold	0	When and in which country(ies): Brief explanation of the field covered by the patent* (if different from above):
Questions about project's outcomes	**Number**	**Comments or suggestions for further investigation**
B - Start-ups		
Question 3.4. Creation of start-up	No	If YES, details: - date of creation: - company name - subject of activity: - location: - headcount: - turnover: - profitable : yes / no / when expected
Question 3.5. Creation of new department of research (ie: organisational change)	No	Name of department:
C – Technology transfer of project's results		

Question 3.6. Collaboration/ partnership with a company ?	YES	Which partner : UGDIST Which company : Telerobot SRL, Genoa, Italy What kind of collaboration ? Realization of robotic parts
4. Other effects		
A - Participation to Conferences/Symposium/Workshops or other dissemination events		
Question 4.1. Active participation*** to Conferences in EU Member states, Candidate countries and NAS. (specify if one partner or "collaborative" between partners)		Names/ Dates/ Subject area / Country:
Question 4.2. Active participation to Conferences outside the above countries (specify if one partner or "collaborative" between partners)		Names/ Dates/ Subject area / Country:
B – Training effect		
Question 4.3. Number of PhD students hired for project's completion	5	In what field : Robotics AI Developmental psychology
Questions about project's outcomes	**Number**	**Comments or suggestions for further investigation**

*** 'Active Participation' in the means of organising a workshop / session / stand / exhibition directly related to the project (apart from events presented in section 2).

C - Public Visibility		
Question 4.4. Media appearances and general publications (articles, press releases, etc.)	0	References: (Please attach relevant information)
Question 4.5. Web-pages created or other web-site links related to the project	3	References: http://www.liralab.it/adapt Also: http://www.ifi.unizh.ch/ailab/people/gomez/roboticHand/index.htm Also: http://www.consciousness.it (Please attach relevant links)
Question 4.6. Video produced or other dissemination material	0	References: (Please attach relevant material)
Question 4.7. Key pictures of results	0	References: (Please attach relevant material .jpeg or .gif)
D - Spill-over effects		
Question 4.8. Any spill-over to national programs	No	If YES, which national programme(s):
Question 4.9. Any spill-over to another part of EU IST Programme	Yes	If YES, which IST programme(s): RobotCub, IST-004370 http://www.robotcub.org
Question 4.10. Are other team(s) involved in the same type of research as the one in your project ?	Yes	If YES, which organisation(s): MANY. *Developmental robotics* is becoming a well-established research field with links into humanoid robotics and neurosciences.

7 Future outlook

Adapt team members are continuing collaboration in several ways. Certainly there is the intention to continue on similar lines of investigation at the European level. In this sense, the project was an excellent means for knowing each other's work in details.
As we have already mentioned, two of the partners (UGDIST and UNIZH) are already involved in a joint FP6 project started about 10 months ago. This is a 5 year long project supported by the European Commission through the Unit E5 (Cognition). This can be seen, in a sense, as the continuation of Adapt since includes for example manipulation as one of its crucial topics.
UGDIST team and CNRS are also still involved in the organization of the Epigenetic Robotics workshop that is planned to be held in Paris next year. This will certainly allow an even stronger collaboration in the future.
We believe that the study of "representation" still remains a fertile research ground which requires further enquiry and experimentation with the right contributions of philosophy, psychology, neurosciences and information technology in general. This mutual rapprochement of various disciplines is probably the only way to address the difficult questions of the "brain" and the "machines".

8 Management report

8.1 Specific objectives for the reporting period

This management report presents the last six months since D1.7 already covers the preceding period. Adapt's objectives during this last period were the consolidation of results, which are described in details in this same document, and the finalization of all workpackages.

8.2 Overview of the progress

See section 4.

8.3 Deliverables

All deliverables have been submitted either before or together with this final report of the project. The complete list of deliverables is shown in section 5 including those delivered during year 3. The same documents are available from the Adapt website for download in PDF format: http://www.liralab.it/adapt

8.4 Comparison between planned and actual work

The comparison between planned and actual work for the reporting period is contained in section 4 and 4.2.9. The main objective during the last 12 months was the continuation of the experimental work along the lines presented in D1.6 and D1.7 (progress reports).

8.5 Milestones

Number	Title	Delivery date (month)
M1	Tentative Theory formulation	7
M2	Validated theory and common vocabulary	36
M3	Different robotic setups to test the effect of morphology	12
M4	Formal analyses and first setup of conclusions	30
M5	Final evaluation of morphology changing experiments	30
M6	Human like robotic setup	15
M7	Experimental setup and paradigm	12
M8	Result of behavioral experiments	30
M9	Modeling of coherent representations	33
M10	Basic units design and implementation	12
M11	Multi sensory modalities integrations	21
M12	Artificial intentional architecture	33

[yellow] reached, [grey] relative to reporting period.

8.6 State of the art update

State of the art updates are no longer applicable to Adapt since the project is at the end.

A major component that is still missing on current manipulative devices is the skin. Sensitizing a skin-like surface while maintaining the mobility of the robot and making room for the large number of connections is clearly daunting, probably requiring new technology (e.g. nanotechnology, micromachining).

Research is clearly underway as shown by this feature article in the Guardian:
http://www.guardian.co.uk/life/feature/story/0,13026,1550736,00.html
and, clearly, the artificial skin would be a very nice feature to add in the next generation of robots especially if they have to work in human populated environment. Other aspects are clearly in need of improvement: actuators, compliance, visual system, etc.

8.7 Actions taken after Y2 review

The second year review report was positive. The major recommendation was to double our effort in trying to recover the delay that plagued one of the workpackages throughout the duration of the project. This was partially caused by the delay in preparing the experimental setups. In particular, the realization of the robot hand(s) has taken longer than expected, and

also starting new behavioral experiments with infants requires a lot of pre-planning (e.g. ethical issues, approval by the local university committees, etc.). We also corrected the format of certain deliverables as requested.

8.8 Planned work and status of experiments

This section does not apply since this is the last reporting period. The experiments are described in details in various submitted deliverables and in part in this document (see section 4).

9 Project management and coordination

Project management during the last 12 months progressed mainly through electronic means (email and website for exchanging data and other information) or phone calls. We did not schedule an extra formal meeting since the team met in various occasions either because of other projects or for conferences and workshops. For instance the coordinators met twice with people from CNRS (in Ferrara – Italy for a workshop on the origin of language and at ICDL05 in Osaka, Japan) and at least three times with people from UNIZH (project meetings and at the IEEE CIRA05 conference in Espoo, Finland). We had papers submitted at the same conferences (e.g. Epigenetic Robotics and CIRA). We estimated the level of communication to be sufficient for the goal of the project. UGDIST is also still planning and developing a common platform with UNIZH among others with the goal of sharing it with a broader community of scientists with an interest in brain sciences and robotics: i.e. the robot as a tool.

Management required also the organization of the Adapt booth at CeBit last March in Hanover, and the final review meeting in London next September 22nd.

Finally, Adapt and the FET were advertised mainly in scientific publications and presentations/posters at conferences. We published 37 papers in about 3 years (some are still in press and/or submitted).

As planned, Adapt will contribute to the Handbook of Presence Research.

10 Cost breakdown

Note: the figures reported here are only indicative. Complete calculation has been performed after the end of the project (September 30th, 2005) and submitted in the form of the final cost statements.

Participant Code	One person-month corresponds to N hours
C1 – DIST	141
P2 – UNIZH	179
P3 – CNRS	135
P4 –UPMC	

Work-Package ID	Title		Reporting period
WP1	Project management		1.10.2004 – 30.09.2005
Participant Code	Spent (person-months)	Planned (person-months) Total	Start date / End date Month 1 / Month 36
C1 – DIST	0.7	3	
P2 – UNIZH[1]	1.8	1 (1)	
P3/P4 – CNRS/ UPMC	0.5	1.2	

Work-Package ID	Title		Reporting period
WP 2	Theory of intentionality and the sense of being-there		1.10.2004 – 30.09.2005
Participants Code	Spent (person-months)	Planned (person-months) Total	Start date / End date Month 1 / Month 36
C1 – DIST	4	12	
P2 – UNIZH[1]	0	10 (5)	
P3/P4 – CNRS/ UPMC	0.3	4	

Work-Package ID	Title		Reporting period
WP 3	Embodiment and body morphology		1.10.2004 – 30.03.2005
Participants Code	Spent (person-months)	Planned (person-months) Total	Start date / End date Month 1 / Month 30
C1 – DIST	3	12	
P2 – UNIZH[1]	12	24 (10)	
P3/P4 – CNRS/ UPMC	4.3	12	

Work-Package ID	Title		Reporting period
WP 4	Development of Coherent Representations		1.10.2004 – 30.04.2005
Participants Code	Spent (person-months)	Planned (person-months) Total	Start date / End date Month 1 / Month 31
C1 – DIST	6	14	
P2 – UNIZH[1]	10	25 (10)	
P3/P4 – CNRS/ UPMC	9.7	26	

Work-Package ID	Title		Reporting period
WP5	System's architecture		1.10.2004 – 30.06.2005
Participants Code	Spent (person-months)	Planned (person-months) Total	Start date / End date Month 1 / Month 33
C1 – DIST	4	12	
P2 – UNIZH[1]	3.6	12 (3)	
P3/P4 – CNRS/ UPMC	1	4	

The number between brackets report the persons/month spent by permanent staff at UNIZH and not charged to the project.

Title				Reporting period		
Cumulative effort				1.10.2004 – 30.09.2005		
Participants Code	SPENT HOURS	Spent (person-months)	Planned hours 3nd year	Planned person-months 3nd year	Planned hours (TOTAL)	Planned person-months (TOTAL)
C1 – DIST	2496	17.7	2496	17.7	7488	53
P2 – UNIZH	5083	28.4	4296	24	12888	72 (29)
P3/P4 – CNRS/UPMC	2124	15.8	2124	15.8	6372	47.2

11 Information dissemination and exploitation of results

See section 9.

11.1 Publications

See section 5.1.