MIRROR

IST–2000-28159

Mirror Neurons based Object Recognition

Deliverable Item 4.6

Results on comparison between “artificial” and “real” neurons

Delivery Date: November 15th, 2004

Classification: Internal

Responsible Person: Prof. Giulio Sandini – University of Genova

Partners Contributed: ALL

Contract Start Date: September 1st, 2001 Duration: 36 Months

Project Coordinator and Partners:
DIST - University of Genova (Prof. Giulio Sandini and Dr. Giorgio Metta)
Department of Biomedical Sciences – University of Ferrara (Prof. Luciano Fadiga)
Department of Psychology – University of Uppsala (Prof. Claes von Hofsten)
Instituto Superior Técnico – Computer Vision Lab – Lisbon (Prof. José Santos-Victor)

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

Content list

1. Artificial and real *mirror* neurons

1.1. Introduction

The MIRROR project had two interrelated goals: (i) advancing the understanding of the mechanisms used by the brain to learn and represent gestures and, inter-alia, (ii) building an artificial system that learns to communicate by means of body gestures. We adopted a three-pronged approach based on the results of neurophysiological and developmental psychology experiments, on the construction of models from the recording of human movements, and on the implementation of these models on various robotic platforms interacting in a natural environment.

The biological motivation of this study is the discovery of the so-called mirror neurons in the primates' premotor cortex. These neurons behave as a "motor resonant system", activated both during execution of goal directed actions and during the observation of similar actions performed by others. One hypothesis is that this unified representation serves the acquisition of goal directed actions during development and the recognition of motor acts, when executed by somebody else.

MIRROR's main scientific contribution is a plausible explanation of the development of mirror neurons. This explanation was constructed by means of mathematical models, engineering and neural sciences.

This deliverable draws a comparison between properties found in mirror neurons, through biological experiments, and the artificial models and implementations created throughout the project with the aim of better understanding the role possibly played by these visuo-motor neurons in action understanding. Needless to say, this comparison will be kept at an abstract level (functional) since achieving a neuronal-like implementation was not one of the goals of the MIRROR project.

Section 1.2 describes the biological basis of MIRROR which represents simultaneously our source of inspiration and a fundamental guideline throughout the whole implementation of biological principles on our robotic platforms. Section 1.5 describes the experimental platforms employed for the execution of the project: we devote some more text here to the description of the platforms specifically developed for MIRROR. Section 1.3 summarizes the model of F5's mirror neurons. Section 1.4 gives a plausible account of the ontogenesis of the "mirror" system and, finally, section 1.6 summarizes the overall achievements of the project.

Section 2 provides a somewhat more detailed description of the experiments that were carried out during the last year of the project and which correspond to the most integrated experiments.

1.2. Biological context: neurophysiology of area F5

Area F5 forms the rostral part of inferior premotor area 6 (Figure 1). Electrical microstimulation and single neuron recordings show that F5 neurons discharge during planning/execution of hand and mouth movements. The two representations tend to be spatially segregated with hand movements mostly represented in the dorsal part of F5, whereas mouth movements are mostly located in its ventral part. Although not much is known about the functional properties of "mouth" neurons, the properties of "hand" neurons have been extensively investigated.

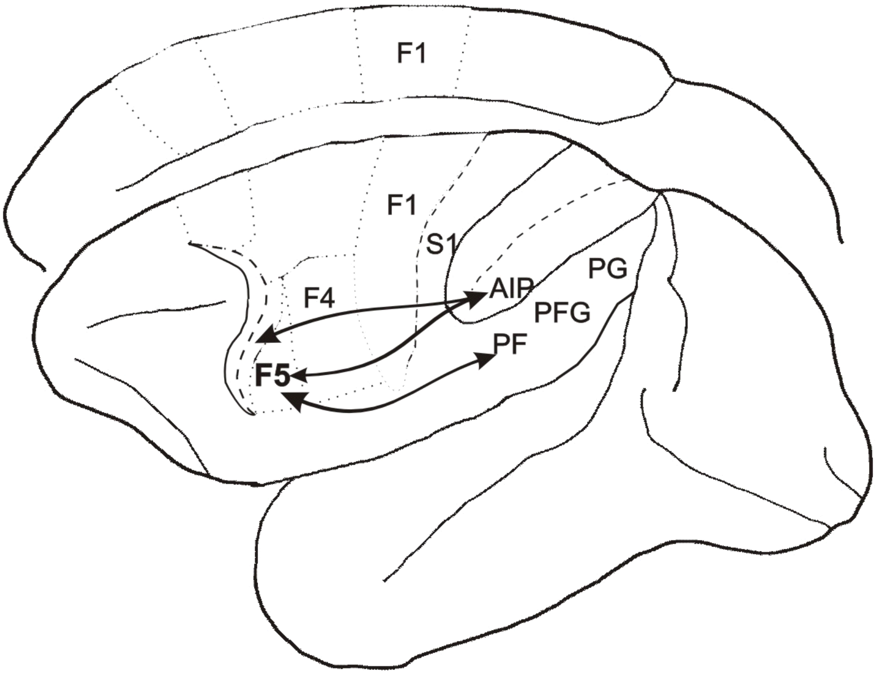

Figure 1: Lateral view of monkey right hemisphere. Area F5 is buried inside the *arcuate sulcus* (posterior bank) and emerges on the convexity immediately posterior to it. Area F5 is bidirectionally connected with the inferior parietal lobule (areas AIP, anterior intraparietal, PF and PFG). Areas F5 sends some direct connections also to hand/mouth representations of primary motor cortex (area F1) and to the cervical enlargement of the spinal cord. This last evidence definitely demonstrates its motor nature.

1.2.1. Motor neurons

Rizzolatti and colleagues (Rizzolatti et al., 1988) found that most of the hand-related neurons discharge during goal-directed actions such as grasping, manipulating, tearing, and holding. Interestingly, they do not discharge during finger and hand movements similar to those effective in triggering them, when made with other purposes (e.g., scratching, pushing away). Furthermore, many F5 neurons are active during movements that have an identical goal regardless of the effector used to attain them. Many grasping neurons discharge in association with a particular type of grasp. Most of them are selective for one of the three most common monkey grasps: precision grip, finger prehension, and whole hand grasping. Sometimes, there is also specificity within the same general type of grip. For instance, within the whole hand grasping, the prehension of a sphere is coded by neurons different from those coding the prehension of a cylinder. The study of the temporal relation between the neural discharge and the grasping movement showed a variety of behaviors. Some F5 neurons discharge during the whole action they code; some are active during the opening of the fingers, some during finger closure, and others only after the contact with the object. A typical example of a grasping neuron is shown in Figure 2. In particular, this neuron fires during precision grip (Figure 2, top) but not during whole hand grasping (Figure 2, bottom). Note that the neuron discharges both when the animal grasps with its right hand and when the animal grasps with its left hand.

Taken together, these data suggest that area F5 forms a repository (a vocabulary) of motor actions. The words of the vocabulary are represented by populations of neurons. Each indicates a particular motor action or an aspect of it. Some indicate a complete action in general terms (e.g., take, hold, and tear). Others specify how objects must be grasped, held, or torn (e.g., precision grip, finger prehension, and whole hand prehension). Finally, some of them subdivide the action in smaller segments (e.g., fingers flexion or extension).

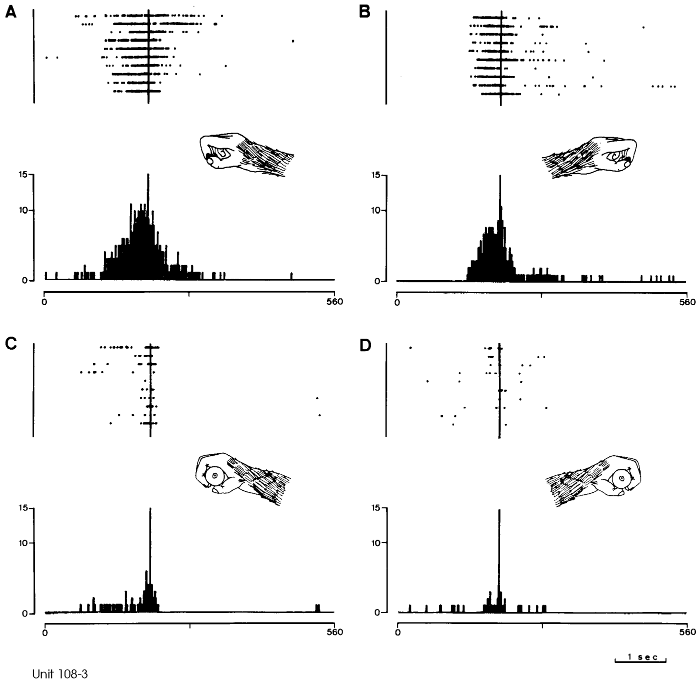

Figure 2: F5 grasping neurons. In the uppermost part of each panel eight successive trials are represented. Each dot represents an action potential. In the lowermost part the sum histogram is drawn. Trials are aligned with the moment at which the monkey touches the object (vertical lines across histograms). Ordinates: spikes/second; Abscissa: time (20 ms bins); from (Rizzolatti et al., 1988).

1.2.2. Visuomotor neurons

Some F5 neurons in addition to their motor discharge, respond also to the presentation of visual stimuli. F5 visuomotor neurons pertain to two completely different categories. Neurons of the first category discharge when the monkey observes graspable objects ("canonical" F5 neurons, (Murata et al., 1997; Rizzolatti et al., 1988; Rizzolatti & Fadiga, 1998)). Neurons of the second category discharge when the monkey observes another individual making an action in front of it (Di Pellegrino, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 1996; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). For these peculiar "resonant" properties, neurons belonging to the second category have been named "mirror" neurons (Gallese et al., 1996).

The two categories of F5 neurons are located in two different sub-regions of area F5: "canonical" neurons are mainly found in that sector of area F5 buried inside the arcuate sulcus, whereas "mirror" neurons are almost exclusively located in the cortical convexity of F5 (see Figure 1).

1.2.3. Canonical neurons

Recently, the visual responses of F5 "canonical" neurons have been re-examined using a formal behavioral paradigm, which allowed testing the response related to object observation both during the waiting phase between object presentation and movement onset and during movement execution (Murata et al., 1997). The results showed that a high percentage of the tested neurons, in addition to the “traditional” motor response, responded also to the visual presentation of 3D graspable object. Among these visuomotor neurons, two thirds were selective to one or few specific objects.

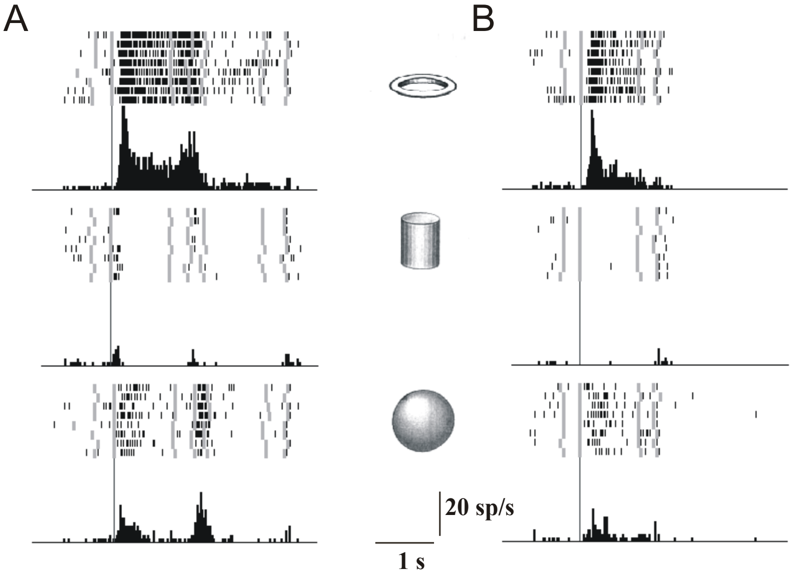

Figure 3: Responses of a visuomotor “canonical” neuron of area F5. Each panel shows the neuron activity recorded during the observation and grasping (A) or the mere observation (B) of three different three-dimensional objects. The alignment of the single trials coincides with the moment in which the object becomes visible (thin line through histograms). In A, the first gray marker following the alignment bar represents the appearance of the signal which commands the beginning of grasping movement. In B, the monkey had to merely observe the object and the first gray bar after alignment represents the moment at which the animal had to release a bar to receive reward. The conventions used in the visualization of the responses are the same as those used in Figure 2. Modified from (Murata et al., 1997).

Figure 3A (grasping in light) shows the responses of an F5 visually selective neuron. While observation and grasping of a ring produced strong responses, responses to the other objects were modest (sphere) or virtually absent (cylinder). Figure 3B (object fixation) shows the behavior of the same neuron of Figure 3A during the fixation of the same objects. In this condition the objects were presented as during the task in 2A, but grasping was not allowed and, at the go-signal, the monkey had simply to release a key. Note that, in this condition, the object is totally irrelevant for task execution, which only requires the detection of the go-signal. Nevertheless, the neuron strongly discharged at the presentation of the preferred object. To recapitulate, when visual and motor properties of F5 neurons are compared, it becomes clear that there is a strict congruence between the two types of responses. Neurons that are activated when the monkey observes small sized objects, discharge also during precision grip. On the contrary, neurons selectively active when the monkey looks at large objects discharge also during actions directed towards large objects (e.g. whole hand prehension).

1.2.4. Mirror neurons

Mirror neurons are F5 visuomotor neurons that discharge when the monkey both acts on an object and when it observes another monkey or the experimenter making a similar goal-directed action (Di Pellegrino et al., 1992; Gallese et al., 1996). Recently, mirror neurons have been found also in area PF of the inferior parietal lobule, which is bidirectionally connected with area F5 (Fogassi, Gallese, Fadiga, & Rizzolatti, 1998). Therefore, mirror neurons seem to be identical to canonical neurons in terms of motor properties, but they radically differ from the canonical neurons as far as visual properties are concerned (Rizzolatti & Fadiga, 1998). The visual stimuli most effective in evoking mirror neurons discharge are actions in which the experimenter's hand or mouth interacts with objects. The mere presentation of objects or food is ineffective in evoking mirror neurons discharge. Similarly, actions made by tools, even when conceptually identical to those made by hands (e.g. grasping with pliers), do not activate the neurons or activate them very weakly. The observed actions which most often activate mirror neurons are grasping, placing, manipulating, and holding. Most mirror neurons respond selectively to only one type of action (e.g. grasping). Some are highly specific, coding not only the type of action, but also how that action is executed. They fire, for example, during observation of grasping movements, but only when the object is grasped with the index finger and the thumb.

Typically, mirror neurons show congruence between the observed and executed action. This congruence can be extremely precise: that is, the effective motor action (e.g. precision grip) coincides with the action that, when seen, triggers the neurons (e.g. precision grip). For other neurons the congruence is somehow weaker: the motor requirements (e.g. precision grip) are usually stricter than the visual ones (any type of hand grasping). One representative of the highly congruent mirror neurons is shown in Figure 4.

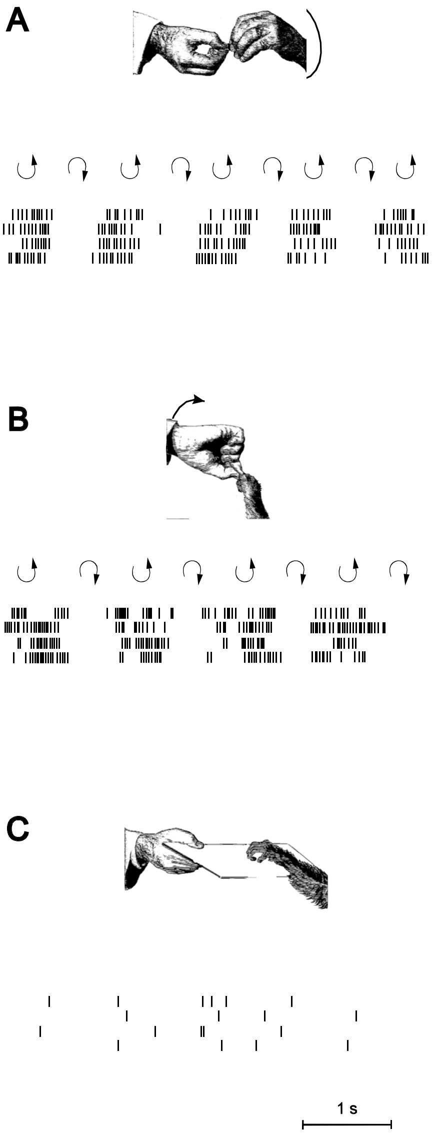

Figure 4: Highly congruent F5 visuomotor "mirror" neuron. Behavioral situations are schematically represented in the upper part of each panel above a series of consecutive rasters and relative response histograms. A, the monkey observes the experimenter who rotates his hands around a raisin alternating clockwise and counterclockwise movements. The response is present only in one rotation direction. B, the experimenter rotates a piece of food held by the monkey who opposes the experimenter movement making a wrist rotation movement in the opposite direction. C, monkey grasps food using a precision grip. Four continuous recordings are shown in each panel. Small arrows above the records indicate the direction of rotations. Note that in C the response is almost absent demonstrating a high degree of motor specificity. From (Rizzolatti et al., 1996).

1.2.5. Discussion

From this short review of the basic properties of F5 neurons, it appears that this area stores potential actions or, as we previously described it, a "vocabulary of actions" (Rizzolatti et al., 1988). The activation of F5 neurons does not necessarily imply an actual action execution. It seems to evoke only its representation. If other contingencies are met, this potential action becomes an actual motor act (see (Rizzolatti & Luppino, 2001)). F5 potential actions can be activated endogenously or exogenously. Exogenous (visual) activation is caused by the observation of objects (canonical neurons) or by the observation of actions made by others (mirror neurons).

To complete this description, we must note that another cortical area (PF) shows a 'mirror-like' response (Fogassi et al., 1998; Gallese, Fogassi, Fadiga, & Rizzolatti, 2002). This area forms the rostral part of the inferior parietal lobule. PF receives input from STS, where there are many neurons that discharge during the observation of the hand (Perrett et al., 1989), and it sends output to area F5. Neurons in area PF are functionally heterogeneous. Most of them (about 90%) respond to sensory stimuli (Fogassi et al., 1998; Gallese et al., 2002;

Hyvarinen, 1982; Leinonen & Nyman, 1979). About 50% of them discharge also in association with the monkey active movements. Neurons responding to sensory stimuli have been subdivided into three categories: “somatosensory” neurons (33%), “visual” neurons (11%), and “bimodal” somatosensory and visual neurons (56%). Among the neurons with visual responses (“visual neurons” and “bimodal neurons”), 41% respond to the observations of actions made by another individual. One third of them, however, similarly to STS neurons, do not appear to have motor-related activity. The other two-third discharge also during the monkey movement and, in most cases, showed the visuo-motor congruence typical of mirror neurons (Gallese et al., 2002).

From the very first moment of the discovery of mirror neurons it was suggested that they could play a role in action understanding. The core of this proposal is the following:

> *When an individual acts he selects an action whose motor consequences are known to him. The mirror neurons allow this knowledge to be extended to actions performed by others.*

Each time an individual observes an action executed by another individual, neurons that represent that action are activated in his or her premotor cortex. Because the evoked motor representation corresponds to that internally generated during active action, the observer understands the observed action (see (Rizzolatti & Luppino, 2001)).

This action recognition hypothesis was recently tested by studying mirror neuron responses in conditions in which the monkey was able to understand the meaning of the occurring action, but without the visual stimuli that typically activate mirror neurons. The rationale of the experiments was the following: if mirror neurons are involved in action understanding, their activity should reflect the meaning of the action and not the specific sensory contingencies. In a series of experiments the hypothesis was tested by presenting auditory stimuli capable of evoking the ‘idea’ of an action (Kohler et al., 2002).

F5 mirror neuron activity was recorded while the monkey was either observing an action producing a sound (e.g. ripping a piece of paper), or hearing the same noise without visual information. The results showed that most mirror neurons that discharge to the presentation of actions accompanied by sound, discharge also in response to the sound alone (“audio-visual” mirror neurons). The mere observation of the same “noisy” action without sound was also effective. Further experiments showed that a large number of audio-visual mirror neurons respond selectively to specific sounds (linked to specific actions). These results strongly support the notion that the discharge of F5 neurons correlates with action understanding and not simply with the stimuli that determine it. The effective stimulus might be visual or acoustic. The neuron fires as soon as the stimulus has specified the meaning.

Another series of experiments studied mirror neurons’ responses in conditions where the monkey was prevented from seeing the final part of the action (and listening to its sound), but were provided with clues on what the action might be. If mirror neurons are involved in action understanding they should discharge also in this condition. This experiment was recently carried out by (Umilta et al., 2001). The experimental paradigm consisted of two basic conditions. In the first condition, the monkey was shown a fully visible action directed toward an object (“full vision” condition). In the second condition, the monkey watched the same action but with its final critical part hidden (“hidden” condition). Before each trial the experimenter placed a piece of food behind the screen so that the monkey knew that there was an object behind it. The main result of the experiment was that more than half of the tested neurons discharged in hidden condition. Some did not show any difference between hidden and full vision condition, others responded stronger in full vision. In conclusion, both these experiments showed that F5 mirror neuron activation correlates with action representation rather than with the properties of the stimulus leading to it. This finding strongly supports the notion that F5 activity plays a fundamental role in the understanding of the meaning of action.

We can then ask why action is so important in interpreting "purely" sensory stimuli (visual, acoustic, or both). Ultimately, it must be noted that animals are actors in their environment, not simply passive bystanders. They have the opportunity to examine the world using causality, by performing probing actions and learning from the response. In other words animals can act and consequently observe the effects of their actions. Effects can be more or less direct, e.g. I feel my hand moving as the direct effect of sending a motor command, or they can be eventually ascribed to complicate chains of causally related events producing what we simply call "a chain of causality". For example, I see the object rolling as a result of my hand pushing it as a result of a motor command. Tracing chains of causality from motor action to perception (and back again) is important to understand how the brain deals with the problem of sensorimotor coordination and, it might be relevant in implementing those same functions in an artificial system (imagine the MIRROR humanoid robot).

Table 1 shows four levels of causal complexity. The simplest causal chain that an actor – whether *robotic or biological* – may experience is the perception of its own actions. The temporal aspect is immediate: visual information is tightly synchronized to motor commands. Once this causal connection is established, it/he/she can go further and use this knowledge about its own body to actively explore the boundaries of the environment (specifically objects). In this case, there is one additional step in the causal chain, and the temporal nature of the response may be delayed since initiating a reaching movement does not immediately elicit consequences in the environment. Finally, we argue that extending this causal chain further will allow the actor to make a connection between its own actions and the actions of another individual. This is clearly reminiscent of what has been observed in the response of the monkey's premotor cortex (area F5).

Type of activity	*Nature of causation*	*Time profile*
Sensorimotor coordination	Direct causal chain	Strict synchrony
Object probing	One level of indirection	Fast onset upon contact, potential for delayed effects
Constructing mirror representation	Complex causation involving multiple causal chains	Arbitrary delayed onset and effects
Object recognition	Complex causation involving multiple observations	Arbitrary delayed onset and effects

Table 1 Degrees of causal indirection. There is a natural trend from simpler to more complicated tasks. The more time-delayed an effect the more difficult it is to model.

An important aspect of the analysis of causal chains is the link with objects. Many actions are directed towards objects, they act on objects, and the goal eventually involves to some extent an object. For example, Woodward (Woodward, 1998), and Wohlschlager and colleagues (Wohlschlager & Bekkering, 2002) have shown that the presence of the object and its identity change the perception and the execution of an action.

While an experienced adult can interpret visual scenes perfectly well without acting upon them, linking action and perception seems crucial to the developmental process that leads to that competence. We can construct a working hypothesis: that action is required whenever the animal (or our robot in the following) has to develop autonomously. Further, as we argued above, the ability to act is also fundamental in interpreting actions performed by a conspecific. Of course, if we were in the standard supervised learning setting, then action would not be required since the trainer would do the job of pre-segmenting the data by hand

and providing the training set to the machine. In an ecological context, some other mechanism has to be provided. Ultimately this mechanism is the body itself and the ability of being the initiator of actions that by means of interaction (and under some suitable developmental rule) generate percepts informative to the purpose of learning.

Grossly speaking, a possible developmental explanation of the acquisition of these functions can be framed in terms of tracing/interpreting chains of causally related events. We can distinguish four main conceptual functions (similar to the schema of Arbib et al. (Arbib, 1981)): reaching, grasping, mimicry, and object recognition. These functions correspond to the four levels of causal understanding introduced in Table 1. They form also an elegant progression of abilities which emerge out of very few initial assumptions. All that is required is the interaction between the actor and the environment, and a set of appropriate developmental rules specifying what information is retained during the interaction, the nature of the sensory processing, the range of motor primitives, etc.

The results briefly outlined in the previous sections can be streamlined into a developmental sequence roughly following a dorsal to ventral gradient. Unfortunately this is a question which has not yet been investigated in detail by neuroscientists, and there is very little empirical support for this claim (apart from (Bertenthal & von Hofsten, 1998) and (Kovacs, 2000)).

What is certainly true is that the three modules/functions can be clearly identified. If our hypothesis is correct then the first developmental step has to be that of transporting the hand close to the object. In humans, this function is accomplished mostly by the circuit VIP-7b-F4-F1 and by PO-F2-area 5. Reaching requires at least the detection of the object and hand, and the transformation of their positions into appropriate motor commands. Parietal neurons seem to be coding for the spatial position of the object in non-retinotopic coordinates by taking into account the position of the eyes with respect to the head. According to (Pouget, Ducom, Torri, & Bavelier, 2002) and to (Flanders, Daghestani, & Berthoz, 1999) the gaze direction seems to be the privileged reference system used to code reaching. Relating to the description of causality, the link between an executed motor action and its visual consequences can be easily formed by a subsystem that can detect causality in a short time frame (the immediate aspect). A system reminiscent of the response of F4 can be developed by the same causal mechanism.

Once reaching is reliable enough, we can start to move our attention outwards onto objects. Area AIP and F5 are involved in the control of grasping and manipulation. F5 talks to the primary motor cortex for the fine control of movement. The AIP-F5 system responds to the “affordances” of the observed object with respect to the current motor abilities. Arbib and coworkers (Fagg & Arbib, 1998) proposed the FARS model as a possible description of the processing in AIP/F5. They did not however consider how affordances can be actually learned during interaction with the environment. Learning and understanding affordances requires a slightly longer time frame since the initiation of an action (motor command) does not immediately elicit a sensory consequence. In this example, the initiation of reaching requires a mechanism to detect when an object is actually touched, manipulated, and whether the collision/touch is causal to the initiation of the movement.

The next step along this hypothetical developmental route is to acquire the F5 mirror representation. We might think of canonical neurons as an association table of grasp/manipulation (action) types with object (vision) types. Mirror neurons can then be thought of as a second-level associative map, which links together the observation of a manipulative action performed by somebody else with the neural representation of one's own action. Mirror neurons bring us to an even higher level of causal understanding. In this case the action execution has to be associated with a similar action executed by somebody else. The two events do not need to be temporally close to each other. Arbitrary time delays might occur.

The conditions for when this is feasible are a consequence of active manipulation. During a manipulative act there are a number of additional constraints that can be factored in to simplify perception/computation. For example, detection of useful events is simplified by information from touch, by timing information about when reaching started, and from knowledge of the location of the object.

Subsequently fully blown object recognition can develop. Object recognition can build on manipulation in finding the physical boundaries of objects and segmenting them from the background. More importantly, once the same object is manipulated many times the brain can start learning about the criteria to identify the object if it happens to see it again. These functions are carried out by the infero-temporal cortex (IT). The same considerations apply to the recognition of the manipulator (either one's own, or another's). In fact, the STS region is specialized for this task. Information about object identity is also sent to the parietal cortex and contributes to the formation of affordances. No matter how object recognition is carried out, at least all the information (visual in this case) pertaining to a certain object needs to be collected and clustered during development so that a model of the object can be constructed.

1.3. Modeling of the mirror system

Our model of area F5 revolves around two concepts that are likely related to the evolution and development of this unique area of the brain. Firstly, we posit that the mirror neuron system did not appear brand new in the brain but evolved from a pre-existing structure devoted solely to the control of grasping actions. The reason for this claim is the finding of a large percentage of motor neurons in F5 (70%) compared to those that have also visual responses. Secondly, if we pose the problem in terms of understanding how such a neural system might actually be autonomously developed (shaped and learned by/through experience during ontogenesis), then the role of canonical neurons – and in general that of contextual information specifying the goal of the action – has to be reconsidered. Since purely motor, canonical, and mirror neurons are found together in F5, it is very plausible that local connections determine at least in part the activation of F5. For explanatory purpose, the description of our model of the mirror system can be further divided in two parts. The first part describes what happens in the actor's brain, the second what happens in the observer's brain when watching the actor (or another individual). As we will see the same structures are used both when acting and when observing an action.

We consider first what happens from the actor's point of view (see Figure 5): in her/his perspective, decision to initiate a particular grasping action is attained by the convergence in area F5 of many factors including contextual and object related information. The presence of the object and of contextual information bias the activation of a specific motor plan among many potentially relevant plans stored in F5. The one which is most fit to the context is then enacted through the activation of a population of motor neurons. The motor plan specifies the goal of the motor system in motoric terms and, although not detailed here, we can imagine that it also includes temporal information. Contextual information is represented by the activation of F5's canonical neurons and by additional signals from parietal (AIP and PF for instance) and frontal areas as in other models of the mirror system (Fagg & Arbib, 1998; Oztop & Arbib, 2002).

None of these contributing neural activities (parietal, temporal, frontal, etc.) can bring, if considered in isolation, F5 over threshold and thus elicit action execution. Instead, activity in different brain areas represents separate unspecific components that become specific only when converging in F5. In this context, the activity of F5 canonical neurons should not be underestimated since it contributes to the definition of the goal of the action and without a goal there is no mirror neurons' response as pointed out in (Gallese et al., 1996).

We can then ask what two individuals do share when mutually interacting. Or similarly, what information can be shared in interacting with objects? Our claim is that it is exactly the goal of the action that is shared among individuals since it is independent of the viewpoint: that is, the final abstract consequences of a given action are, unlike its exact visual appearance, viewpoint independent. The fact that the goal of the action is shared among individuals allows two conspecifics to eventually develop mirror-like representations from the observation of each other's actions and from their own knowledge of actions. In fact, in this model, key to proper development of a mirror representation is the ability to recognize that a specific goal is approximately achieved by employing always the same action. Canonical neurons act as "filters" reducing the probability of generating implausible actions given the context and target object and, thus, actually throwing away irrelevant information. A similar role with respect to the specification of the hand posture is postulated for the hand responsive neurons of STS (Perrett, Mistlin, Harries, & Chitty, 1990).

With reference to Figure 5, our model hypothesize that the intention to grasp is initially "described" in the frontal areas of the brain in some internal reference frame and then transformed into the motor plan by an appropriate controller in premotor cortex (F5). The action plan unfolds mostly open loop. A form of feedback (closed loop) is required though to counteract disturbances and to learn from mistakes. This is obtained by relying on a forward or direct model that predicts the outcome of the action as it unfolds in real-time. The output of the forward model can be compared with another signal derived from sensory feedback, and differences accounted for (the cerebellum is believed to have a role in this). A delay module is included in the model to take into account the different propagation times of the neural pathways carrying the predicted and actual outcome of the action. Note that the forward model is relatively simple, predicting only the motor output in advance: since motor commands are generated internally it is easy to imagine a predictor for this signals. The inverse model (indicated with VMM for visuo-motor map), on the other hand, is much more complicated since it maps sensory feedback (vision mainly) back into motor terms. Visual feedback clearly includes both the hand-related information and the contextual information important for action recognition. Finally the predicted and the sensed signals arising from the motor act are compared and their difference (feedback error) sent back to the controller.

There are two ways of using the mismatch between the planned and actual action: i) compensate on the fly by means of a feedback controller, and ii) adjust over longer periods of time through learning (not explicitly indicated in the model).

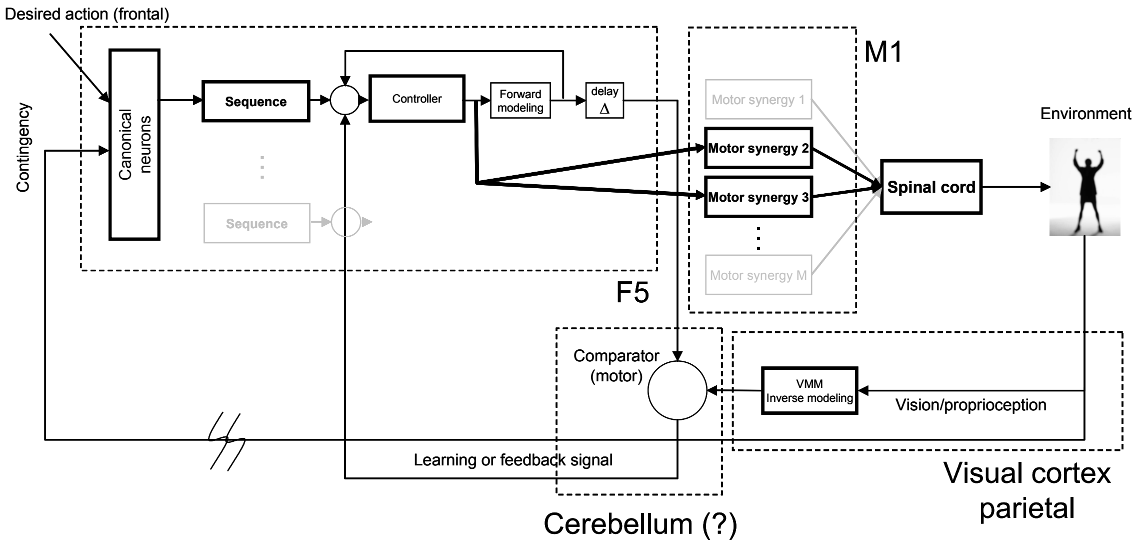

Figure 5: Model schematics of a forward-inverse model of F5 and mirror neurons. The response of the mirror system is seen as an adaptation of a feedback loop controlling the

execution of grasping. The model only contains details of some brain areas while it is known that many others participate to the specification and control of grasp (which are indicated generically in the diagram). Please, refer to the text for the detailed description of the model.

The output of area F5, finally activates the motor neurons in the spinal cord (directly or indirectly through motor synergies) to produce the desired action. This is indicated in the schematics by a connection to appropriate muscular synergies.

Learning of the direct and inverse models can be carried out during ontogenesis by a procedure of self-observation and exploration of the state space of the system: grossly speaking, simply by "detecting" the sensorial consequences of motor commands – examples of similar procedures are well known in the literature of computational motor control (Jordan & Rumelhart, 1992; Kawato, Furukawa, & Suzuki, 1987; Wolpert, 1997; Wolpert, Ghahramani, & Flanagan, 2001).

Learning of context specific information (e.g. the affordances of objects with respect to grasping) can also be achieved autonomously by a trial and error procedure, which explores the consequences of many different actions of the agent's motor repertoire (different grasp types) to different objects. This includes things such as discovering that small objects are optimally grasped by a pinch or precision grip, while big and heavy objects require a power grasp.

In addition, the model includes the concept of "motor vocabulary", since control of action is realized by a "graded" controller but the selection of which fingers to use (what grasp type to apply) is "discrete" and involves the activation of one of the "action" modules described above.

A slightly different activation pattern is hypothesized in the observer situation (see Figure 6). In this case clearly motor and proprioceptive information is not directly available. The only readily available information is vision. The central assumption of our model is that the structure of F5 could be co-opted in recognizing the observed actions by transforming visual cues into motor information as before. In practice, the inverse model is accessed by visual information and since the observer is not acting herself, visual information is directly reaching in parallel the sensori-motor primitives in F5. Only some of them are actually activated because of the "filtering" effect of the canonical neurons and other contextual information (possibly at a higher level, knowledge of the actor, etc.). A successive filtering is carried out by considering the actual visual evidence of the action being watched (implausible hand postures should be weighed less than plausible ones). This procedure could be used then to recognize the action by measuring the most active motor primitive (from the vocabulary). In probabilistic terms this is easily obtained by evaluating all evidence with its likelihood and looking for the maximum a-posteriori probability (Cabido Lopes & Santos-Victor, 2003a).

Comparison is theoretically done, in parallel, across all the active motor primitives (actions); the actual brain circuitry is likely to be different with visual information setting the various F5 populations to certain equilibrium states. The net effect can be imagined as that of many comparisons being performed in parallel and one motor primitive resulting predominantly activated.

Relying on motor information seems to facilitate the organization (clustering) of visual information: that is, the organizational principle of visual information becomes a motoric one. Clearly invariance from the point of view is much better achieved if the analysis of the action is done in motor terms (Cabido Lopes & Santos-Victor, 2003a).

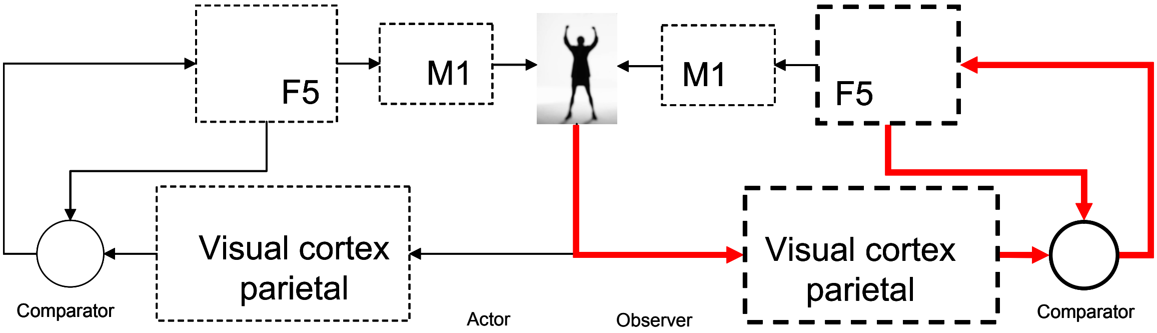

Figure 6: Action and observation activity. The model of Figure 5 is replicated to describe the observer's brain. The observation of a certain action activates the same feedback path used in executing that action (thick solid lines). Many details as shown in Figure 5 were dropped here for clarity of presentation.

The presence of a goal is fundamental to elicit mirror neuron responses (Gallese et al., 1996) and we believe it is also particularly important during the ontogenesis of the mirror system. Supporting evidence is described in the work of Woodward and colleagues (Woodward, 1998) who have shown that the identity of the target is specifically encoded during reaching and grasping movements: in particular, already at nine months of age, infants recognized as novel an action directed toward a novel object rather than an action with a different kinematics, thus showing that the goal is more fundamental than the enacted trajectory.

Developing mirror neurons thus might likely go through the initial maturation of the neural circuitry devoted to the understanding of the actor's goal (involving as we already mentioned F5 canonical neurons) and only afterward to the association of the observed action to one's internal representation of that same action. If this picture is consistent then we can construct an answer to the question of how mirror neurons originate: the actor first learns how to grasp objects and only subsequently associate its own representation of the action to the observed action in all those cases when the goal is the same. This view on mirror neurons has the advantage of not requiring an 'external teacher': that is, learning can proceed completely unsupervised and consistently with the model schematic of Figure 5.

1.3.1. A Bayesian model of canonical and mirror neurons

From the above discussion, two core elements of the prospective mirror neurons model emerge, namely, the use of motor information (or coding) also during the recognition of somebody else's actions and the use of object affordances (we provided support for the relevance of the presence of the target object during action execution).

In practice, many objects are grasped in very precise ways, since they allow the object to be used for some specific purpose or goal. A pen is usually grasped in a way that affords writing and a glass is held in such a way that we can use it to drink. Hence, if we *recognize* the object being manipulated, then recognition immediately provides some information about the most likely grasping possibilities (expectations) and hand appearance, simplifying the task of gesture recognition. The affordances of the object possess a filtering property in reducing the number of possible (or likely) events. Affordances provide expectancies that can be used to single out possible ambiguities. This has clearly to be a module of our overall system architecture.

The traditional approach to recognition involves comparing acquired visual features to data from a training set. Our approach is based on the use a *visuo-motor map* (VMM) to convert visual measurements to motor space first and, only subsequently, to perform the comparison in terms of motor representations (see Figure 5). The advantage of doing this final inference in motor space is two-fold: i) while visual features are to various extents always view-point dependent, motor information is intrinsically view-point independent, and ii) since motor information is directly exploited during this process, imitative behaviors could be trivially implemented given that all the relevant information is readily available. The VMM can be learnt during an initial phase of self-observation where the robot explores its action space (by trying different actions) and observes the visual consequences.

If, for a moment, we do not consider the aspect of the on-line control of the manipulative action, then gesture recognition can be modeled within a Bayesian framework. A Bayesian model allows combining naturally *prior* information and knowledge derived from observations (likelihood). The role played by canonical and mirror neurons will be interpreted within this setting.

Let us assume that we want to recognize (or imitate) a set of gestures, G_i, using a set of *observed* features, $\boldsymbol{F}$. For the time being, these features can either be represented in the motor space (as mirror neurons seem to do) or in the visual space (directly extracted from images). Let us also define a set of objects, O_k, that happen to be observed in the scene (not simultaneously) and which are the goals of a certain grasping actions.

Prior information is modeled as a probability density function, $p(G_i|O_k)$, describing the probability of each gesture given a certain object. The observation model is captured in the *likelihood function*, $p(F|G_i,O_k)$, describing the probability of observing a set of (motor or visual) features, conditioned to an instance of the pair of gesture and object. The *posterior* density can be directly obtained through Bayesian inference:

$$\begin{cases} p(G_i \mid F, O_k) = p(F \mid G_i, O_k) p(G_i \mid O_k) / p(F \mid O_k) \\ \hat{G}_{MAP} = \arg\max_{G_i} p(G_i \mid F, O_k) \end{cases} \tag{1.1}$$

where $p(F|O_k)$ is just a scaling factor that will not influence the classification. The *MAP* estimate, G_{MAP}, is the gesture that maximizes the posterior density in equation (1.1). In order to introduce some temporal filtering (since the information over time is available), features of several images can be considered:

$$p(G_i \mid F, O_k) = p(G_i \mid F_t, F_{t-1}, ..., F_{t-N}, O_k) \tag{1.2}$$

where F_j are the features corresponding to the image at time instant *j*. The posterior probability distribution can be estimated using a naive approach, assuming independence between the observations at different time instants. The justification for this assumption is that recognition does not necessarily require the accurate modeling of the density functions. We have:

$$p(G_i \mid F_t, ..., F_{t-N}, O_k) = \prod_{j=0}^{N} \frac{p(F_{t-j} \mid G_i, O_k) p(G_i \mid O_k)}{p(F_{t-j} \mid O_k)} \tag{1.3}$$

The role of canonical neurons in the overall classification system lies essentially in providing affordances, modeled here as the *prior* density function, $p(G_i|O_k)$ that, together with evidence from the observations, will shape the final decision. This density can be estimated by the relative frequency of gestures in the training set. In practice, if we were to work with a complete system estimation and learning of affordances would require a much more complex learning procedure. The ultimate goal would still be the estimation of prior probabilities (that could still be done by estimating the relative frequencies of actions) but acquiring the visuo-motor information autonomously is perhaps a feat in itself.

Canonical neurons are also somewhat involved in the computation of the likelihood function since they respond both on the *gesture* and *object* (and in the model $p(G_i|O_k)$ shows this relationship), thus implicitly defining another level of association. Computing the likelihood function, $p(F| G_i, O_k)$, might be difficult since the shape of the data clusters might be quite complicate. We modeled these clusters as mixtures of Gaussian and the Expectation-Maximization algorithm was used to determine both the number of the Gaussian terms and their coefficients.

Mirror neurons are clearly represented by the responses of the maximization procedure since both motor and visual information determine the activation of a particular unit (for real neurons) and the corresponding probability (for artificial neurons), as shown in Figure 7.

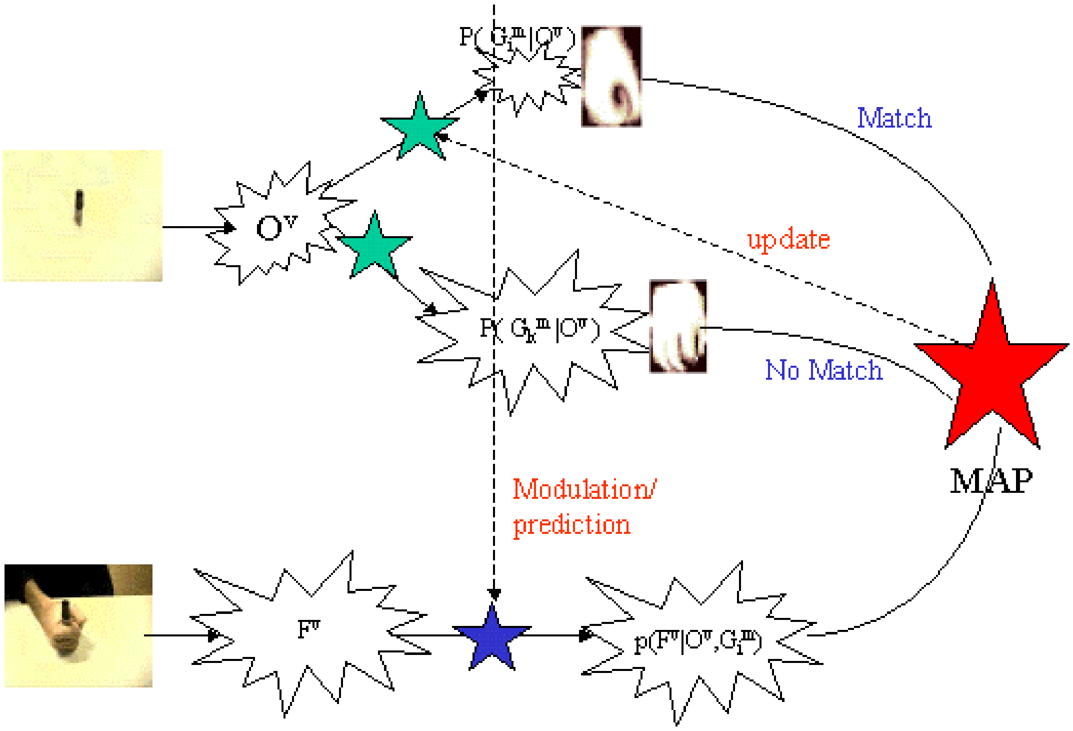

Figure 7: Bayesian model of canonical and mirror neurons exemplified for hand gesture recognition.

1.4. A Developmental Perspective

Although on a superficial reading it might seem that the Bayesian model encompasses all what it has to be said about mirror neurons, in fact it is substantially a supervised learning model. To relax the hypothesis of having to "supervise" the machine during training by indicating which action is which, we need to remind what the evidence on mirror neurons tells us. First of all, it is plausible that the 'canonical' representation is acquired by self exploration and manipulation of a large set of different objects. F5 canonical neurons represent an association between objects' physical properties and the action they afford: e.g. a small object affords a precision grip, or a coffee mug affords being grasped by the handle. This understanding of object properties and the goal of actions is what can be subsequently factored in while disambiguating visual information. There are at least two level of reasoning: i) certain actions are more likely to be applied to a particular object – that is, probabilities can be estimated linking each action to every object, and ii) objects are used to perform actions – e.g. the coffee mug is used to drink coffee. Clearly, we tend to use actions that proved to lead to certain results or, in other words, we trace backward the link between action and effects: to obtain the effects apply the same action that earlier led to those effects.

Bearing this is mind, when observing some other individual's actions; our understanding can be framed in terms of what we already know about actions. In short, if I see someone drinking from a coffee mug I can hypothesize a particular action (that I know already in motor terms) is used to obtain that particular effect (of drinking). This link between mirror neurons and the goal of the motor act is clearly present in the neural responses observed in the monkey. It is also the only possible way of autonomously learning a mirror representation. Technically speaking, the learning problem is still a supervised one but the information can now be collected autonomously. The association between the canonical response (object-action) and the mirror one (including vision of course) is made when the observed consequences (or goal) are recognized as similar in the two cases – self or other acting. Similarity can be evaluated following different criteria ranging from kinematic (e.g. the object moving along a certain trajectory) to very abstract (e.g. social consequences such as in speech).

1.5. Data acquisition setup and Robotic Artifacts

Experiments probing different aspects of the model were conducted on four different experimental platforms: i) the cyber glove data acquisition setup described next, ii) a humanoid robot developed at LIRA-Lab: Babybot, iii) a humanoid robot developed at IST: Baltazar, and iv) a humanoid robot at MIT during the collaboration between the LIRA-Lab and the AI-Lab. The reason for experimenting on different platforms resides in the fact that accurate manipulation is still beyond the state of the art of robotic systems – not to mention the difficulty of appropriately learning to grasp generic objects – while, on the other hand, the complete model, including learning, is better tested on a fully autonomous system.

1.5.1. A machine with hands

The rationale for the grasping data acquisition setup is to build a machine that embeds some of the principles of operation that we identified in the model to perform action recognition. Clearly, this requires accessing both motor and visual information in the operation of learning to recognize gestures. The simplest way to provide "motor awareness" to a machine is by recording grasping actions from multiple sources of information including joint angles, spatial position of the hand/fingers, vision, and touch. For this purpose we assembled a computerized system composed of a cyber glove (CyberGlove by Immersion), a pair of CCD cameras (Watek 202D), a magnetic tracker (Flock of bird, Ascension), and two touch sensors (FSR). Data was sampled at frame rate, synchronized, and stored to disk by a Pentium class PC. The cyber glove has 22 sensors and allows recording the kinematics of the hand at up to 112Hz. The tracker was mounted on the wrist and provides the position and the orientation of the hand in space with respect to a base frame. The two touch sensors were mounted on the thumb and index finger to detect the moment of contact with the object. Cameras were mounted at appropriate distance with respect to their focal length to acquire the execution of the whole grasping action with maximum possible resolution.

The glove is lightweight and does not limit anyhow the movement of the arm and hand as long as the subject is sitting not too far from the glove's interface. Data recording was carried out with the subject sitting comfortably in front of a table and enacting grasping actions naturally toward objects approximately at the center of the table. Data recording and storage were carried out through a custom-designed application; Matlab was employed for post processing.

Recording human movements for either teaching robots or animating robotic avatars is certainly not new (Mataric, 2000; Nakanishi, Morimoto, Endo, Schaal, & Kawato, 2003). Our setup though is not merely using this information for re-enacting precise trajectories or simply interpolating from exemplar movements as in (Rose, Cohen, & Bodenheimer, 1998). While the emphasis on previous work was on creating novel movements similar (according to certain criteria) to observed ones, it was our intendment from the beginning to use motor information as an aggregating principle to determine which visual features might be important and to actually select appropriate visual features for action recognition. Grossly speaking, in designing a classifier which uses visual information, it is crucial to choose a set of features (i.e. what to measure from images) that maximizes the distance between categories and minimizes the spread within each category. This guarantees large margins which are then related to generalization and potentially simplifies the task of the classifier (or simplifies the classifier itself, a bit along the line of the Statistical Learning Theory (Vapnik, 1998)). That this is the case is still to be shown by the ongoing experimental activity. In addition, since actions are coded by transforming visual information in motor terms we expect to obtain a much larger invariance to changes in the visual appearance of the action.

We simply want to point out here that even when acting is not strictly required, possessing a hand is not optional in at least two different ways: i) in humans where visual features have to develop autonomously; since there's no "engineer within the brain" deciding what is important for visual classification, and ii) in building machines; since the actual optimal features might be far from obvious and, simultaneously, the chance of selecting a sufficiently optimal set are exceedingly low (the space of possible visual features is large). In both situations what is important is the *unsupervised* (autonomous) acquisition of the visuo-motor representation. What we would like to characterize is the sequence of events that allows learning a visuo-motor representation starting from lesser elements and without assuming unreasonable pre-specification of the structures. Also, what is important in our model is to show how much the same learned representation can be subsequently co-opted in recognizing other individuals' actions (as for area F5). A more thorough discussion is given in the following sections.

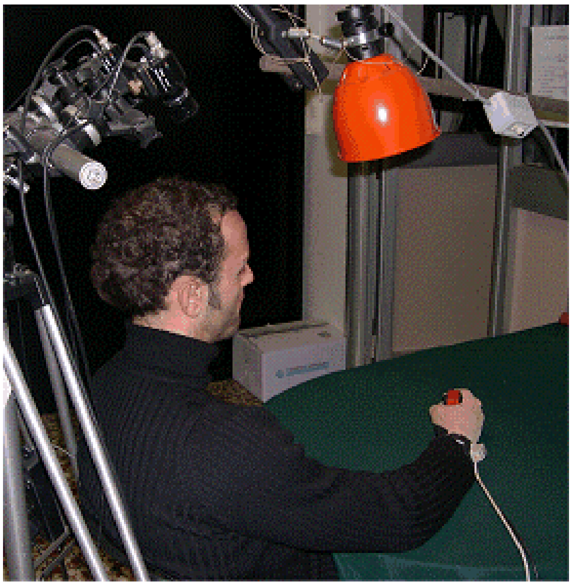

Figure 8: The recording setup. The user wears the cyber glove and reaches for an object. Cameras in this image are places behind the person and see a good portion of the table. The

visual environment for the experiments was well controlled (e.g. illumination) and the background was made uniform in color.

The selected grasping types approximately followed Napier's taxonomy (Napier, 1956) and for our purpose they were limited to only three types: power grasp (cylindrical), power grasp (spherical), and precision grip. Since the goal was to investigate how much invariance could be learned by relying on motor information for classification, the experiment included gathering data from a multiplicity of viewpoints. The database contains objects which afford several grasp types to assure that recognition cannot simply rely on exclusively extracting object features. Rather, according to our model, this is supposed to be a confluence of object recognition with hand visual analysis. Two exemplar grasp types are shown in Figure 9: on the left panel a precision grip using all fingers; on the right one a two-finger precision grip.

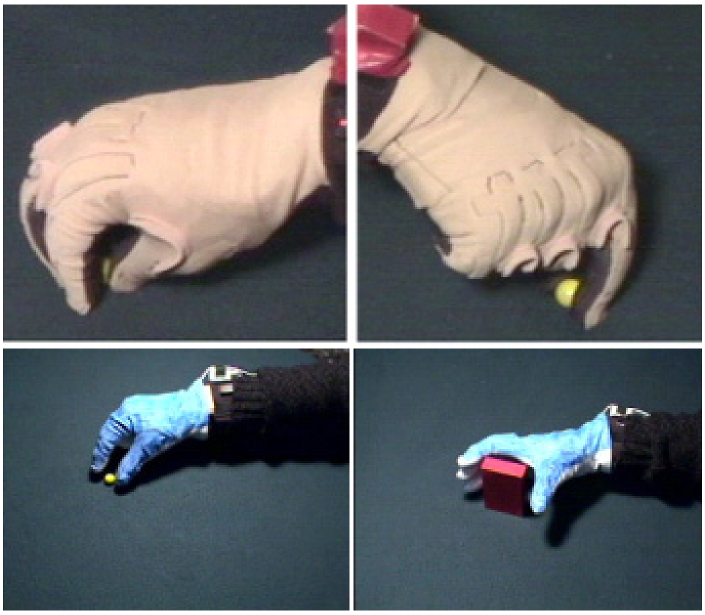

Figure 9: Exemplar grasp types as recorded during data collection. The topmost row shows two types of grasp applied to the same object (the small glass ball) from different points of view. The bottom images show two very different grasp types from the same point of view.

The objects were also three: a small glass ball, a parallelepiped which affords multiple grasps, and a large sphere requiring power grasp. Each grasping action was recorded from six different subjects (right handed, age 23-29, male/female equally distributed), and moving the cameras to 12 different locations around the subject including two different elevations with respect to the table top which amounts to 168 sequences per subject. Each sequence contains the vision of the scene from the two cameras synchronized with the cyber glove and the magnetic tracker data. This data set was used for building the Bayesian classifier in motor space (Cabido Lopes & Santos-Victor, 2003a).

1.5.2. Robots

Three robotic artifacts were used to conduct the experiments mainly in the second and third years of the project.

The first robotic setup was built in Genoa – the Babybot – and consists of an upper torso humanoid robot composed of a five degree of freedom (DOF) head, a six DOF arm and a five-fingered hand (Figure 10). Two cameras are mounted on the head; they can pan independently and tilt together on a common axis. Two additional degrees of freedom allow the head to pan and tilt on the neck. The arm is an industrial manipulator (Unimate Puma 260); it is mounted horizontally to closer mimic the kinematics of a human arm. The hand has seventeen joints distributed as follows: four joints articulate the thumb, whereas index, middle, ring and little fingers have three phalanges each. The fingers are underactuated to reduce the overall number of motors employed. Consequently only two motors allow the thumb to rotate and flex while two motors are connected to the index finger; finally the remaining fingers are linked and form a single virtual finger that is actuated by the two remaining motors. Intrinsic compliance in all joints allows passive adaptation of the hand to the object being grasped. Magnetic and optic encoders provide position feedback from all phalanges. As far as the sensory system is concerned, the robot is equipped with two cameras, two microphones, and a three axis gyroscope mounted on the head. Tactile feedback is available on the hand; a force sensor allows measuring force and torque at the wrist. Finally proprioceptive feedback is available from the motor encoders. More details about the robot can be found in (Giorgio Metta, 2000; Natale, 2004).

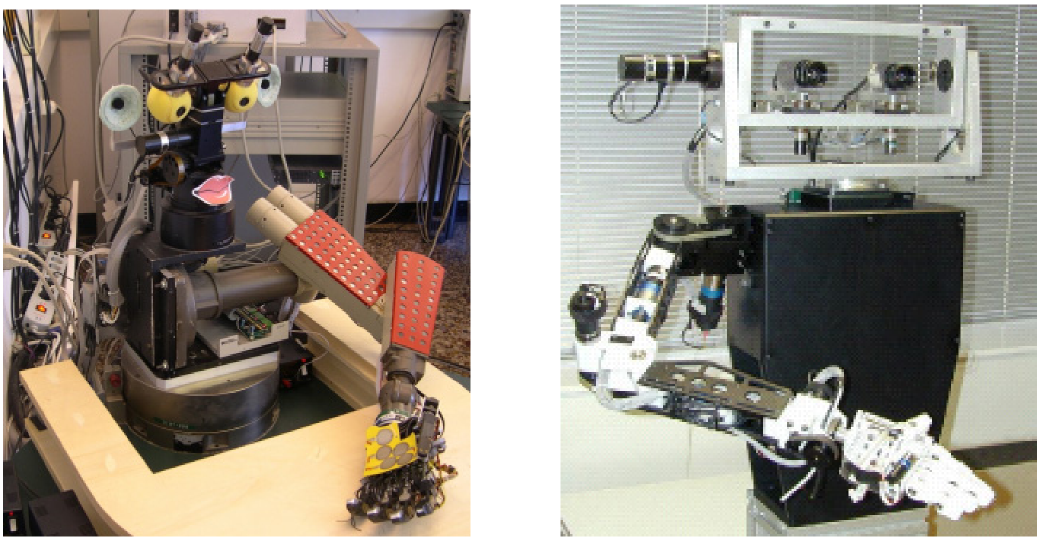

Figure 10: Two robotic setups used in the experiments. Left: the Babybot built in Genoa and used throughout the project. Right: the Baltazar humanoid built at IST and used for some of the experiments.

The second robotic setup used for the MIRROR experiments was built at IST during the last year of the project. It consists of an anthropomorphic arm with 6 degrees of freedom, equipped with a multi-fingered hand with 11 degrees of freedom (actuated by 4 motors only). The robot is shown in Figure 10 and details can be found in (Cabido Lopes, Beira, Praça, & Santos-Victor, 2004).

Finally, we would like to mention a third platform we accessed during the collaboration between the University of Genoa and the AI-Lab (now CSAIL) at MIT. This was an upper torso humanoid robot with head, two arms and torso for overall 22 degrees of freedom

(Brooks, Brezeal, Marjanovic, & Scassellati, 1999). Hands were not available and actions were thus limited to poking and prodding through the use of two simple flippers.

1.6. Experiments and overall status of the project

The following table gives a comprehensive view of the experiments developed during MIRROR and a tentative account of which part of the model they address. The outcome of this conceptual modeling activity is both a general outline of the development of motor skills in humans and a specific probabilistic model of the functioning of the mirror system. As we have already abundantly stressed, MIRROR had two somewhat interrelated goals: i) the clarification of some of the questions or gaps within the model or understanding of the functioning of the mirror system, and ii) implementation of the model in an autonomous learning/developing artifact.

Experiments cover aspects ranging from the development of reaching/hand localization towards understanding of the contribution/role of mirror neurons to communicative behaviors. In the following we describe in brief each experiment and their contribution to the global "picture". Experiments were conducted on one side on animals or human subjects and, on the other side on various robotic prototypes. Part of the activity of MIRROR (especially during the first year) consisted in preparing the experimental platforms – for example, one of the robots has been now equipped with a five-finger hand and new equipment for recording children's movements has been acquired.

Title	Type	Location in the model	Description
Hand localization	Robotics	Models to a certain extent the responses of neurons in area F4 (Graziano, 1999)	The robot uses vision and proprioception to learn the position of the hand/arm in space. This is reminiscent of the responses of neurons in F4 and VIP.
Reaching	Robotics	Models the pathway VIP-7b-F4-F1 that seems to be responsible for reaching in the monkey (Flanders et al., 1999)	The robot learns to combine gaze direction with the motor commands required to reach for a visually identified target.
Reaching	Biology		Reaching requires the appropriate timing of motor commands to move the hand near the object of interest. We investigated how children learn to predict the position in space where contact with a moving object occurs.
Tracking	Biology		An experimental paradigm has been established to investigate the development of infants' predictive tracking and reaching for moving objects. Also, situations where objects are

			subject to occlusions were investigated.
Affordances	Robotics	Models the AIP-F5 area with hints on the role of mirror neurons (Gallese et al., 1996)	In a set of experiments we explored the acquisition of object affordances in a simplified situation where the robotic artifact could only poke and prod objects.
Role of visual information	Biology	Area F5 and F1 are investigated	We recorder from area F5 and F1. The goal of this experiment was that of establishing the degree of modulation of the F5 response as a function of the amount of visual information available.
Building the VMM	Robotics	Models the transformation from visual cortex to parietal and finally F5.	Mapping visual space into motor space. Visual features (PCA components) were mapped into motor information. A Bayesian classifier was then trained on the motor space information.
Building the VMM	Biology		We asked whether stereoscopic vision is necessary to create the visuomotor map and how much finger occlusion during grasping influences action recognition. In practice we investigated the ability of predicting the outcome of grasping actions in various conditions with varying amount of visual feedback (2D vs. 3D).
Mirror system and communication	Biology	Premotor cortex, area 44 (Liberman & Mattingly, 1985)	In this TMS study, we investigated in humans to what extent the mirror system is involved in verbal communication. A second fMRI experiment was carried out more recently to assess the type of processing area 44 is contributing to: i.e. simple non-meaningful movements (phonemes) versus “whole word” stimuli.
Development of hand orientation	Biology		We investigated when and how infants start to control the hand posture in relation to the orientation of the object. Results showed that reaching and hand positioning are two somewhat different actions.
Development of	Biology		Children’s ability to adjust the

object affordances			orientation of objects to insert them into holes is studied. The study investigated at what age infants start to perceive properties such as shape and orientation relative to a goal.
Learning to recognize a grasping action	Robotics	F5, STS, PF, AIP	A Bayesian model of action recognition has been formulated including the VMM as described earlier. The model properly accounts for the activation of canonical and mirror neurons in action recognition.
Learning models of the hand and objects	Robotics		We developed an “active” approach to the problem of collecting visual information to build models of the objects the robot encounters. A Bayesian approach was developed based on visual features extracted by some low level attention processing.
Learning to grasp a visually identified object	Robotics		We integrated various components of the model into a single experiment where the robot learns to grasp a visually identified object. The robot first gets acquainted with objects, builds a model of them, and it is then able to search, gaze, and finally grasp them.
Development of motion processing	Biology	MT/MST	The development of area MT/MST is investigated using a new high-density EEG with a network of 128 sensors. The development of this area responsible for the perception of visual motion is related to the development of smooth pursuit.
Human recording	Biology	F5?	In collaboration with the Neurosurgery Unit of the Hospital in Udine – Italy, we were able to start recording in humans (see later for the ethical justification of this activity).
TMS experiment while viewing actions	Biology		We were able to assess the enhancement of the corticospinal system excitability as a consequence of training by either practice or observation only.

2. Experiments with the robotic artifacts

One of the major issues addressed during the last year of the project was the integration of the results into a developing artifact. Additional experiments were also conducted but the major drive was towards the integration of modules in the robotic platform. Most of this work was carried out using the humanoid robot available in Genoa. This setup includes a five-finger hand for experimentation with different action/grasp types in a reasonably natural environment. Additional work was developed using a new humanoid-type robotic platform developed at IST. All these experiments are briefly summarized in this section.

2.1. A developmental approach to grasping

If the first interaction with the environment happens through vision, then it is only by acting that we are able to discover certain properties about the entities populating the external world. For example by applying different actions on an object we can probe it for properties like weight, rigidity, softness and roughness, but also collect information about its shape. Furthermore we can carry out further exploration to learn how an object behaves when certain actions are applied to it or, in a similar way, how we can handle it to achieve a particular goal (affordances and tool use).

Besides, autonomous agents can exploit actions to actively guide exploration. For an artificial system like a robot this can be extremely useful to simplify learning. For instance the system can identify a novel object in a cluttered environment and grasp it, bring it closer to the cameras (so to increase the resolution), rotate it, squeeze it, and eventually drop it after enough information has been acquired. Exploration in this case is easier because it is initiated by the agent in a self-supervised way. This does not only mean that the agent has direct control on the exploration procedure, but simply, as we discussed in the introduction, that it can establish a causal link between its actions and the resulting sensory consequences. While holding and rotating an object, for example, its appearance, the tactile sensation coming from the hand, along with the forces exerted by gravity at the wrist, can be associated to the position of the fingers around the object and to its orientation. Similarly, affordances can be explored by trying to grasp the object in different ways and recording what actions are successfully applied to the object.

The ability to manipulate objects emerges relatively early in children during development; for instance at three months infants start reaching for objects to grasp them and bring them to their mouth; nine months-old babies are able to control the fingers to perform different grasp types (precision grip, full palm grasp, (von Hofsten, 1983)). Maybe it is not by chance that infants learn to grasp long before they can speak or walk. However, even the simplest form of grasping (with the full hand open as newborns do in the first months of their life) is not a trivial task. It involves at least the ability to control gaze, to move the arm to reach a particular position in space, pre-shape the hand and orient it according to the object’s size and orientation. In addition, the impact with the object must be predicted to correctly plan the pre-shaping of the hand (von Hofsten, Vishton, Spelke, Feng, & Rosander, 1998). In infants all these motor competences are not present at birth; rather they are acquired during development by exploiting an initial set of innate abilities which allow them to start the exploration of their body in the first place and subsequently of the external environment.

As per the project goal, we present a biologically plausible developmental path for a humanoid robot mimicking some aspects of infant development. In our artificial implementation we divided this process in three phases. The first phase concerns learning a body self-image; the robot explores the physical properties of its own body (e.g. the weight of the arm, the visual appearance of the hand) and basic motor abilities (e.g. how to control the head to visually explore the environment).

We call the second stage learning to interact with the environment: here the robot starts the active exploration of the external world and learns to perform goal-directed actions to objects (grasping).

Finally the third phase involves learning about objects and others; the robot's previous experience is used to create expectations on the behavior of other entities (objects as well as intentional agents).

It is important to stress that this classification is not meant to be strict and clear cut as presented here. These three stages, in fact, are not present in the robot as three separate entities; rather, all modules "grow" at the same time; the maturation of each subsystem allows the overall system to perform better and, at the same time, it increases the possibility of other parts to develop. Thus for instance the ability of the head to perform saccade allows the system to fixate objects and start reaching for them. Arm movements, although not accurate, in turn allow the system to initiate interaction and improve its performance based on its own mistakes.

The third phase is perhaps the most critical and challenging one as it leads to the development of advanced perceptual abilities. In previous work we have addressed at least some aspects related to this phase (Fitzpatrick, Metta, Natale, Rao, & Sandini, 2003; Natale, Rao, & Sandini, 2002). Here we order this developmental sequence so to show how it could lead to the development of something similar to the mirror system in the biological brain.

2.1.1. Learning a body-map

The physical interaction with the environment requires a few prerequisites. To grasp an object the robot must be able to direct gaze to fixate a particular region of the visual field, program a trajectory with the arm to bring it close to the object and eventually grasp it. Although reaching in humans is mostly ballistic, localization of the hand is required to perform fine adjustments at the end of the movement, or, in any case, during learning. We previously addressed the problem of controlling the head to perform smooth pursuit and saccades towards visual and auditory targets (Giorgio Metta, 2000; Natale, Metta, & Sandini, 2002; Panerai, Metta, & Sandini, 2002). Here we focus the discussion on the second aspect: that is learning to localize the arm end-point and to segment it out from the rest of the world.

It is known that in humans and primates the brain maintains an internal representation of the body, the relative positions of the body segments, their weight and size. This body-schema is used for planning but, maybe more importantly, also to predict the outcome of an ongoing action and anticipate its consequences. Prediction and anticipation are important aspects of cognition because they extend our ability to understand events by matching our perception and expectations.

Graziano and colleagues (Graziano, 1999; Graziano, Cooke, & Taylor, 2000) found neurons in the primate's motor cortex (area 5) which code the position of the hand in the visual field. Tested under different conditions these neurons had receptive fields coding the position of the hand in space; in particular some of them showed to be driven by visual information (that is they fired when the hand was visible), whereas others happened to rely on proprioceptive feedback only (they fired even in those cases when the hand was covered with a barrier).

In infants self-knowledge appears after a few months of development; for instance five-months-old infants are able to recognize their own leg movements on a mirror (Rochat & Striano, 2000). But what are the mechanisms used by the brain to build such representation? Pattern similarities between proprioceptive and other sensory feedbacks are cues that could be used to disambiguate between the external world and the body. Indeed, experimental results on infants corroborate the hypothesis that perception of intermodal form plays a dominant role in the development of self-recognition (Rochat & Striano, 2000).

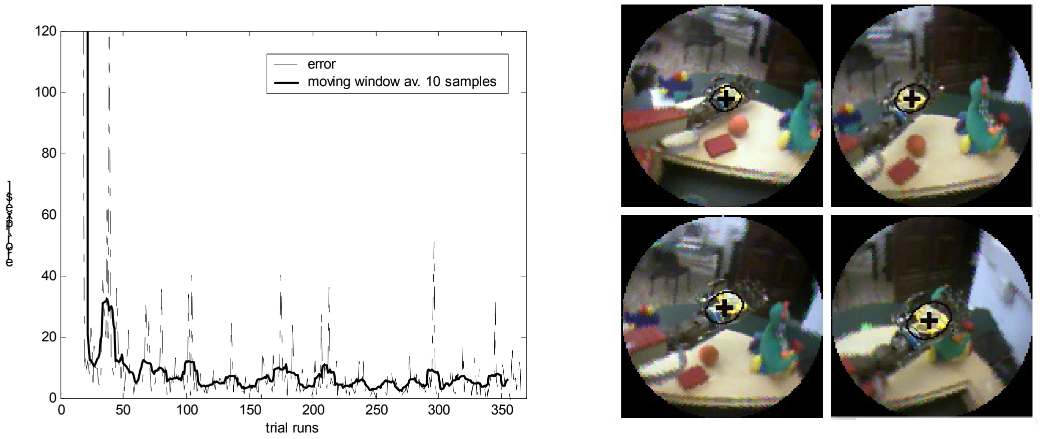

Figure 11: Learning the hand localization. Left: average error in pixels during learning. Right: result of the localization at the end of learning (robot's point of view, left eye).

The problem of learning a body-schema has been addressed in robotics as well. Yoshikawa et al. (Yoshikawa, Hosoda, & Asada, 2003) exploited the idea that the body is invariant with respect to the environment; in their work proprioceptive information is used to train a neural network to segment the arms of a mobile robot. In the case of Metta and Fitzpatrick (G. Metta & Fitzpatrick, 2003) the robot moved the arm in a repetitive way and optic flow was computed to estimate its motion in the visual field. Cross-correlation between visual and proprioceptive feedback was then used to identify those part of the image, which were more likely to be part of the arm end-point. Similarly, in our case the robot moves the wrist to produce small periodic movements of the hand. A simple motion detection algorithm (image difference with adaptive background estimation) is employed to compute motion in the visual field; a zero-crossing algorithm detects the period of oscillation for each pixel. The same periodic information is extracted from the proprioceptive feedback (motor encoders). Pixels which moved periodically and whose period was similar to the one computed in the motor feedback are selected as part of the hand. Instead, the algorithm segments out uncorrelated pixels (e.g. someone walking in the background). The segmentation is a sparse pixel map; a series of low-pass filters at different scales suffices in removing the outliers and produce a dense binary image (the segmentation mask).

By using this segmentation procedure the robot can learn to detect its own hand. In particular it builds three models: a color histogram and two forward models to compute the position and size of the hand in the visual field based on the current arm posture. The latter are two neural networks, which provide the expected position, shape and orientation of the hand given the arm proprioceptive feedback. The color histogram is independent (at least to a certain extent) of the orientation and position of the hand and can be easily computed out of a single trial. A better model of the hand is described in section 2.1.4. However by accumulating the result of successive experiments, it is possible to reduce the noise and increase the accuracy of the histogram. The forward models are trained as follow: the segmentation procedure is repeated several times thus randomly exploring different arm postures. For each trial the center of mass of the segmentation is extracted and used as a training sample for the first neural network. Additional shape information is extracted by fitting a parametric contour on the segmented regions; a good candidate for this purpose is an ellipse because it captures orientation and size of the hand. Accordingly, a second neural network is trained to compute the ellipse's parameters which fit the hand in the visual field given the current arm posture.

The color histogram gives a statistical description of the color of an object and can be used to spot regions of the image that are more likely to contain the hand. However, the histogram alone cannot distinguish objects that have similar colors. By putting together the contributions of the two neural networks it is possible to reduce the ambiguities and identify precisely the hand in the visual field. Figure 2 reports the result of learning and shows the result of the localization for a few different arm postures.

Overall the hand detection system can be employed in different ways. Since its output is expressed in a retinocentric reference frame, the x-y coordinate of the hand can be sent directly to the controller of the head which can then track the arm as it moves in space (see Figure 12). In the next section we will see how this coordinated behavior can be exploited to learn how to reach and touch visually identified objects. Another possibility is to make the robot look at its hand to explore an object that has grasped (later in Section 2.1.6). This feature may prove helpful especially in case the robot is endowed with foveated vision. Finally by querying the forward models with the desired joint angles (a “virtual” arm position), the robot can predict what will be the position of the hand for a given arm posture; in other words, the same map used for the localizing the hand can convert the hand trajectory from joint space to retinal coordinates.

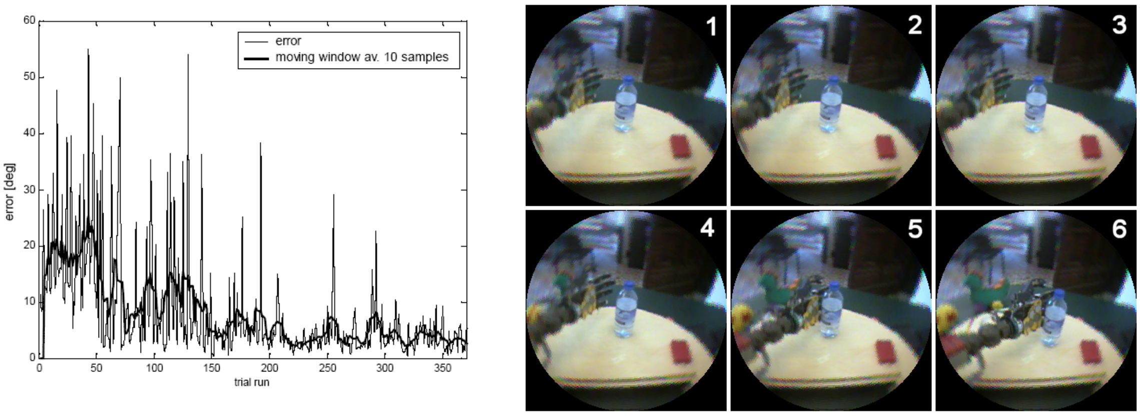

Figure 12: Learning to reach. Left: error during learning, joint angle (squared root of the sum squared error of each joint, in degrees). Right: an exemplar sequence after learning (robot’s point of view, left eye).

2.1.2. Learning to reach

Two problems need to be solved to successfully reach for a location in space: i) the estimation of the kinematic transformation between the target position and the corresponding arm posture, and ii) the generation of the motor commands required to achieve that particular posture (inverse dynamics and trajectory generation). In this section we focus on the first problem: i.e. how to learn the transformation required in computing the joint configuration to reach a specific point in space.

Let us assume that the robot is already fixating the target. In this case the fixation point implicitly defines the target for reaching; besides, if the correct vergence angle has been achieved, the posture of the head univocally defines any position in the three dimensional space (in polar form distance, azimuth and elevation). To solve the task the robot needs to have some knowledge of following functional relation:

$$\boldsymbol{q}_{arm} = f(\boldsymbol{q}_{head}) \qquad (2.1)$$

where $\boldsymbol{q}_{head}$ is a vector which represents the head posture (target point) and $\boldsymbol{q}_{arm}$ is the corresponding arm joint vector.

Thus reaching starts by first achieving fixation of the object; $\boldsymbol{q}_{head}$ is then used as input into the map of equation (2.1) and to recover the arm motor command $\boldsymbol{q}_{arm}$. Interestingly, the procedure to learn the reaching map is straightforward if we relay on the tracking behavior that was described in the previous section. At the beginning (before starting to reach) the robot explores the workspace by moving the arm randomly while tracking the hand; each pair of arm-head posture defines a training sample that can be used to learn ***f*** in equation (2.1) (the reaching map). After enough samples are collected, the robot can use equation (2.1) and start reaching for visual targets. However, exploration of the workspace and actual reaching do not need to be separate. If the map is properly initialized (for instance with a few initial values with a reasonable distribution) exploration can be achieved by adding noise to the output of the map and activating the hand tracking behavior to estimate the actual position of the arm. Proper initialization of the map is required to make sure that the control values sent to the arm are always meaningful (and safe); the noise component guarantees the exploration of the workspace (a sort of random walk exploration). As learning proceeds and new samples are collected the amount of noise (its variance) can be progressively reduced to zero to achieve precise reaching. In the experiment reported here, the two methods were interleaved. After reaching the robot performed a few random movements while tracking the hand (the noise in this case had a Gaussian distribution with zero mean and standard deviation of 5 degrees). This strategy is beneficial because it allows collecting more than a single training sample for each reaching trial; besides, in this way, the exploration is biased toward those regions of the workspace where reaching occurs more often (usually in the part of the workspace in front of the robot).

Once the final arm posture is retrieved from the map, it is still necessary to plan a trajectory to achieve it. For this purpose a linear interpolation is carried out between the current and final arm position; the arm command is thus applied in "small steps". The actual torque is computed using a PD controller employing gravity compensation (for details see (Natale, 2004)). The complete control schema is reported below (Figure 13).

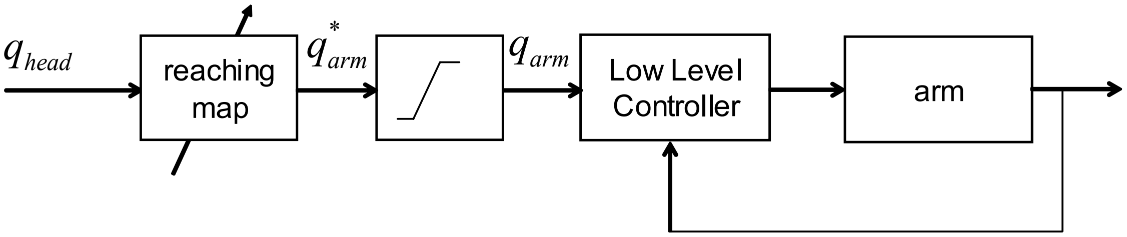

Figure 13: Reaching control schema.

2.1.3. Visual processing

One of the first steps of any visual system is that of locating suitable interest points in the scene, to detect events, and eventually to direct gaze toward these locations. It was already recognized years ago that the ability to move helps in solving computer vision problems (Aloimonos, Weiss, & Bandyopadhyay, 1988; Ballard & Brown, 1992). This paradigmatic shift was so important that led to the development of "active vision" as a modus operandi in a good part of the computer vision community. Further, it is now evident that the same "approach" is taken by the brain. 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 (e.g. our fovea). This leads to two sorts of 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. There are two traditional assumptions in the literature attempting to account for these facts. On the one hand, the space-based attention theory holds that attention is allocated to a region of space, with processing carried out only within a certain spatial window of attention. This theory sees attention as a spotlight, an internal eye or a sort of zoom lens. On the other hand, object-based attention theories argue that attention is directed to an object or a group of objects to process any properties of the selected object(s) rather than of regions of space.

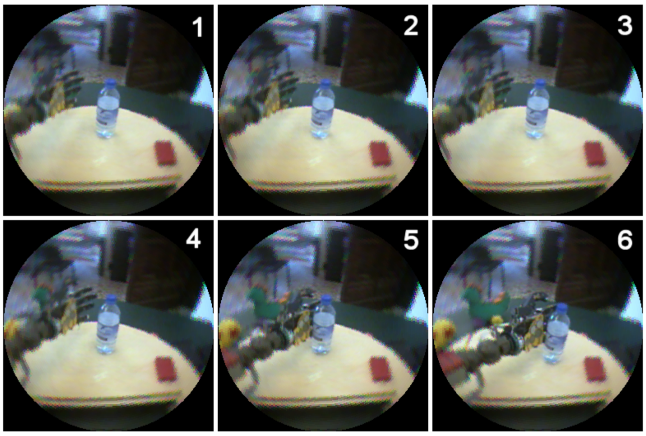

Figure 14: An experiment of reaching. The robot's attention is attracted by the bottle which is fixated. Once gaze is stable on the target a reaching action is initiated by querying the reaching map with the current gaze angles. The hand eventually touches the bottle after some hundreds of milliseconds (shown on frame 6).

Further, attention can be directed in two ways. One approach uses bottom-up information including basic features such as color, orientation, motion, depth, conjunctions of features, etc. A feature or a stimulus catches the attention of the system if it differs from its immediate surrounding in some dimensions and the surround is reasonably homogeneous in those dimensions. However, a bottom-up salience-based method cannot always capture attention if the system is already focused or directed elsewhere. For this reason, it is necessary to recognize the importance of how attention is also influenced by top-down information relevant to the particular ongoing task. Our approach is object based and includes both bottom-up and top-down information processing.

For the scope of this manuscript we only discuss the issue of how to determine where to look while we assume that once the decision is taken the robot can gaze efficiently to that location. Figure 15 shows the block diagram of the first stage of the visual processing of the robot. The input of the model is a sequence of color log-polar images (Sandini & Tagliasco, 1980). Each image is processed independently and as a first step the red, green and blue channels are separated. Successively they are combined to generate the so-called color opponent channels. Each of these channels typically indicated as (R+G-, G+R-, B+Y-) has center-surround receptive field (RF) with spectrally opponent color responses. That is, for example, a red input in the center of a particular RF increases the response of the channel R+G-, while a green one in the surrounding will decrease its response. The spatial response profile of the RF is expressed by a difference-of-Gaussians (DoG) over the two sub-regions of the RF, 'center' and 'surround'. For each pixel in the input image a response imagining a RF centered in the pixel is computed, generating an output image of the same size of the input. This computation, considering for example the *R+G-* channel is expressed by:

$$R^{+}G^{-}(\mathbf{x}) = a \cdot R(\mathbf{x}) \otimes \gamma_{c}(\mathbf{x},\sigma_{c}) - b \cdot G(\mathbf{x}) \otimes \gamma_{s}(\mathbf{x},\sigma_{s}) \qquad (2.2)$$

The two Gaussian functions are not balanced and the ratio ***b/a*** is 1.5 which means that the achromatic information is preserved. The achromatic information is not explicitly processed since it is implicitly coded in the three chromatic opponent channels similarly to what happens in the human retina.

Edges are then extracted on the three channels separately by employing a generalization of the Sobel filter due to (Li & Yuan, 2003). A single edge map is generated by a combination of the three filtered channels as follows:

$$E(\mathbf{x}) = \max\{abs(E_{RG}(\mathbf{x})), abs(E_{GR}(\mathbf{x})), abs(E_{BY}(\mathbf{x}))\} \qquad (2.3)$$

In Figure 15 it is shown an example of the edge map: it has to be noted that the log-polar transform has the side effect of sharpening the edge near the fovea due to the “cortical magnification factor”. This effect is compensated multiplying the edge image by an exponential function.

To fill the spaces between edges, the watershed algorithm (Smet & Pires, 2000; Vincent & Soille, 1991) is used. The watershed is chosen for efficiency and because under certain hypotheses it possesses some biological plausibility. The activation is spread from the center of the image (in the edge map) until it fills all the spaces between edges. As a result the image is segmented in blobs with either uniform color or uniform gradient of color. In the latter case, we are considering the condition where the gradient is not enough to determine edges and consequently more blobs.

It is known that a feature or stimulus is salient if it differs from its immediate surrounding in some dimensions (in a suitable feature space) and the surround is reasonably homogeneous along those same dimensions. The size of the spot or focus of attention is not fixed: it rather changes depending on the size of the objects in the scene. To account for this fact the greater part of the visual attention models in literature uses a multi-scale approach filtering with some type of “blob” detector (typically a difference of Gaussian filter) at various scales. We reasoned that this approach lacks of continuity in the choice of the size of the attention focus, so we propose instead to select dynamically the scale of interest based on the size of the blobs. That is the salience of each blob is calculated in relation to a neighborhood proportional to its size. In our implementation we consider a rectangular region 2.5 times the size of the bounding box of the blob.

The choice of a rectangular window is not incidental, rather it was chosen because filters over rectangular regions can be computed efficiently by employing the integral image as for example in (Viola & Jones, 2004). The bottom-up saliency is thus computed as:

$$S_{bottom-up} = \sqrt{(\langle R^{+}G^{-}\rangle_{center} - \langle R^{+}G^{-}\rangle_{surround})^{2} + (\langle G^{+}R^{-}\rangle_{center} - \langle G^{+}R^{-}\rangle_{surround})^{2} + (\langle B^{+}Y^{-}\rangle_{center} - \langle B^{+}Y^{-}\rangle_{surround})^{2}} \qquad (2.4)$$

Where the **<>** indicate the average of the image values over a certain area (indicated in the subscripts). In the formula above we simply compute the center-surround difference over the three channels with the size of the center and surround determined as a function of the size of the blob currently examined.

The top-down influence on attention is, for the moment, calculated in relation to the task of the visual search. In this situation a model of the object to search in the scene is given (see later) and thus this information is used to bias the saliency computation procedure. In practice, the top-down saliency map is computed simply as the distance between each blob's average color and the average color of the target.

$$S_{top-down} = \sqrt{((\langle R^+G^- \rangle_{blob} - \langle R^+G^- \rangle_{object})^2 + ((\langle G^+R^- \rangle_{blob} - \langle G^+R^- \rangle_{object})^2 + ((\langle B^+Y^- \rangle_{blob} - \langle B^+Y^- \rangle_{object})^2} \qquad (2.5)$$

With a notation similar to the one above. The total salience is simply estimated as the linear combination of the two terms above:

$$S = aS_{top-down} + bS_{bottom-up} \qquad (2.6)$$

The total salience map *S* is eventually normalized in the range 0-255 which as a consequence determines that the salience of each blob in the image is relative to the most salient one.

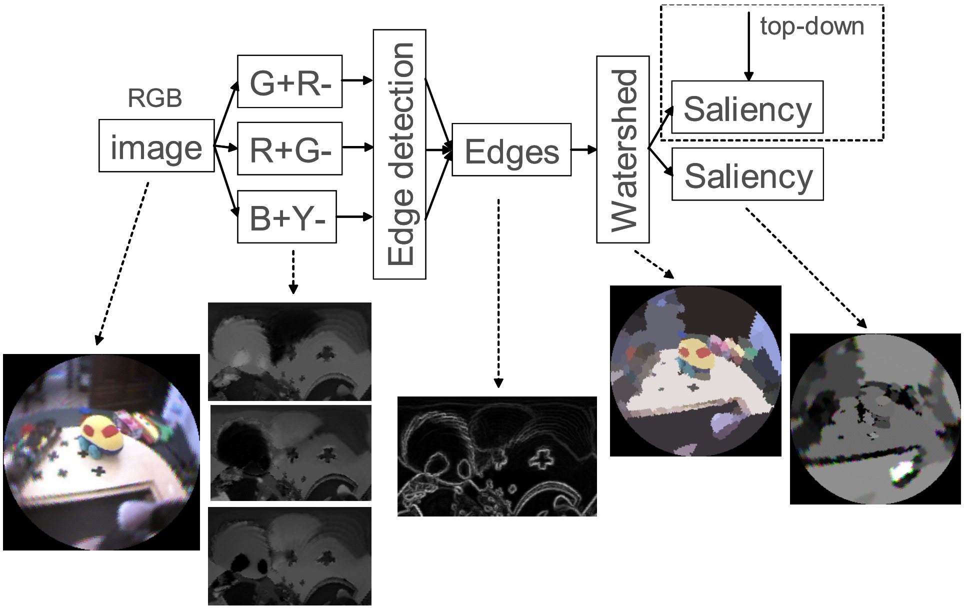

Figure 15: Model of the visual system that determines the saliency of areas of the visual scene.

Local inhibition is transiently activated in the salience map, in a variable area to prevent the focus of attention to be redirected immediately to a previously attended location. Such an "inhibition of return" (IOR) has been demonstrated in human visual psychophysics.

In particular, Posner and Cohen (Posner & Cohen, 1984) demonstrated that the IOR does not seem to function in retinal coordinates but it is instead attached to environmental locations. Together with Klein (Klein, 1988), they proposed that the IOR is required to allow an efficient visual search by discouraging shifting the attention toward locations that had already been inspected. Static scenes, however, are seldom encountered in real life: objects move so a "tagging system" that merely inhibited environmental locations would be almost useless in any real world situation. Tipper et al. (Tipper, 1991) were among the first to demonstrate that the IOR could be attached to moving objects, and this finding has been since then replicated and extended (Abrams & Dobkin, 1994; Gibson & Egeth, 1994; Tipper, 1994). It seems that humans work by attaching tags to objects and moving them as objects move, hence the IOR seems to be coded in an object-based frame of reference.

Our system implements a simple object-based IOR. The robot maintains circular list of the last five positions visited in a head-centered coordinate system together with the color information of the blobs at the marked positions. Even if the robot moves its gaze by moving the eyes or the head altogether, it can maintain memory of the visited objects. These locations are inhibited only if they show the same color seen earlier. In case the object moves, the color changes, and the location becomes available for fixation again.

2.1.4. Object models

Based on the visual processing described above and summarized in Figure 15, we decided to build the models of the objects as combination of blobs. In short, the idea is to represent objects as collections of blobs and their relative positions (neighboring relations). The model is created statistically by looking at the same object many times from different points of view. A histogram of the number of times a particular blob is seen is used to estimate the probability that the blob belongs to the object.

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:

$$\begin{aligned} P(O \mid M) &= \frac{P(M \mid O)P(O)}{P(M)} \\ O_{MAP} &= \arg\max_{O,\sim O}\{P(O \mid M), P(\sim O \mid M)\} \end{aligned} \tag{2.7}$$

where in our case: *P(O)* 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 in the fixated. In the following experiments we only carried out a detection experiment for a single object, there are consequently only two classes, one representing the object and another representing the background. *P(M)* is not considered since it does not affect the maximization. Also, *P(O)* or *P(~O)* is simply set to *0.5*.

The probabilities *P(M|O)* and *P(M|~O)* (object and background respectively) are estimated during an exploration phase using a Bayesian estimator (Papoulis & Pillai, 1991), so it is calculated as *(k+1)/(n+2)*, where *n* is the total number of frames used during the exploration phase and *k* is the number of occurrences of a specific blob in *n* frames.

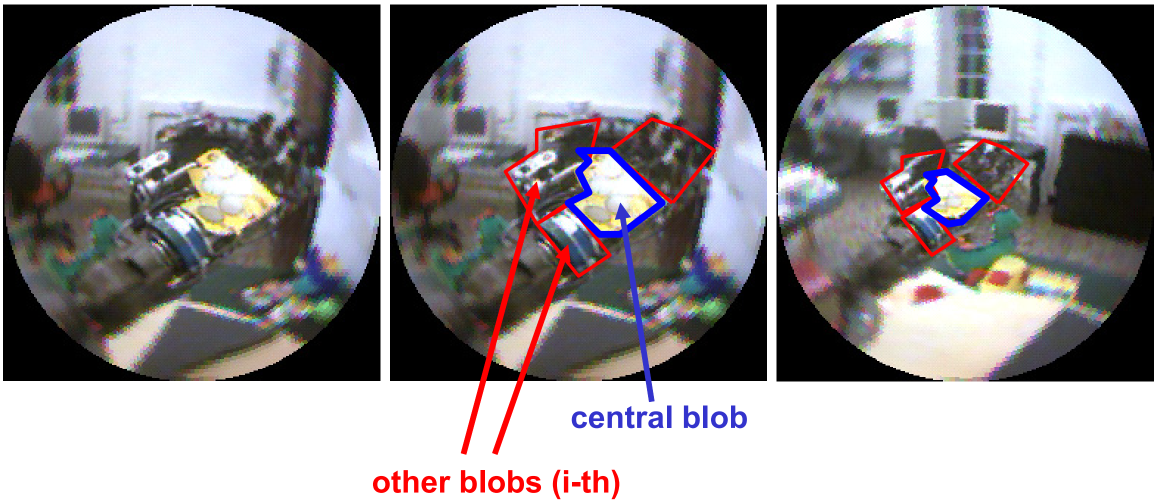

Figure 16: Example of the information collected in building the model of the hand. In this example the condition of fixating the center of the hand is easily obtained by relying on the body map described in section 2.1.1.

Since a single blob is not discriminative enough, we considered the probabilities of observing pairs of blobs instead (*M* represents pairs of blobs). In practice, we estimate the probability of all blobs adjacent to the central blob (taken as reference) of belonging to the object (see Figure 16). That is, we exploit the fact the robot is fixating the object and assume the central blob will be constant across fixations. If this condition is verified, we can then use the central blob as a reference and estimate the following probability:

$$P(B_i \in O \mid B_c) \qquad (2.8)$$

Where B_i is the i-th blob and B_c is the central blob. This procedure, although requiring the "active participation" of the robot (through gazing) is less computationally demanding compared with 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 were able to use in our experiments. Our requirement was that of building the object model on the fly with the shortest possible exploration procedure.

2.1.5. Object affordances

This section describes an experiment already reported at Y1 which forms a nice conceptual continuation of the acquisition of the object visual model. In this case, the robot acquires information about the behavior of objects and links it to the visual appearance of the object.

Since the robot did not have hands, it could not really grasp objects from the table. Nonetheless there are other actions that can be employed in exploring the physical extent of objects. Touching, poking, prodding, and sweeping form a nice class of actions that can be used for this purpose. The sequence of images acquired during reaching for the object, the moment of impact, and the effects of the action are measured following the approach of Fitzpatrick (Fitzpatrick, 2003a). An example of the quality of segmentation obtained is shown in Figure 17. Clearly, having identified the object boundaries allows measuring any visual feature about the object, such as color, shape, texture, etc.

Unfortunately, the interaction of the robot's flipper with objects does not result in a wide class of different affordances. In practice the only possibility was to employ objects that show a characteristic behavior depending on how they are approached. This possibility is offered by rolling affordances: in our experiments we used a toy car, an orange juice bottle, a ball, and a colored toy cube.

The robot's motor repertoire besides reaching consists of four different stereotyped approach movements covering a range of directions of about 180 degrees around the object.

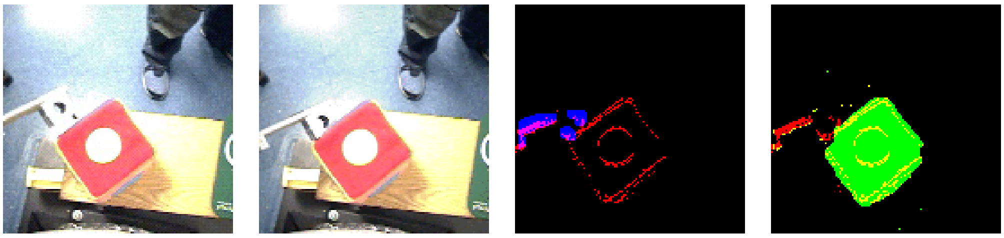

Figure 17: Example of segmentation obtained by reaching and poking an object sitting on a table in front of the robot; adapted from (Fitzpatrick & Metta, 2003). The first two pictures show the moment of impact with the object, the third picture is a color-coded version of the motion detection filter that shows the object motion and the robot flipper in different colors. The fourth image shows the segmented area obtained by further processing the motion information in the third picture.

The experiment consisted in presenting repetitively each of the four objects to the robot. During this stage also other objects were presented at random; the experiment run for several days and sometimes people walked by the robot and managed to make it poke (and segment) the most disparate objects. The robot "stored" for each successful trial the result of the segmentation, the object's principal axis which was selected as representative shape parameter, the action – initially selected randomly from the set of four approach directions –, and the movement of the center of mass of the object for some hundreds milliseconds after the impact was detected. We grouped (clustered) data belonging to the same object by employing a color based clustering techniques similar to Crowley et al. (Schiele & Crowley, 2000). In fact in our experiments the toy car was mostly yellow in color, the ball violet, the bottle orange, etc. In different situations the requirements for the visual clustering might change and more sophisticate algorithms could be used (Fitzpatrick, 2003b).

Figure 18 shows the results of the clustering, segmentation, and examination of the object behavior procedure. We plotted here an estimation of the probability of observing object motion relative to the object own principal axis. Intuitively, this gives information about the rolling property of the different objects: e.g. the car tends to roll along its principal axis, the bottle at right angle with respect to the axis. The training set for producing the graphs in Figure 18 consisted of about 100 poking actions per object. This "description" of objects is fine in visual terms but do not really bear any potential for action since it does not yet contain information about what action to take if it happens to see one of the objects.

For the purpose of generating actions a description of the geometry of poking is required. This can be easily obtained by collecting many samples of generic poking actions and estimating the average direction of displacement of the object. Figure 19 shows the histograms of the direction of movement averaged for each possible action. About 500 samples were used to produce the four plots. Note, for example, that the action labeled as "backslap" (moving the object outward from the robot) gives consistently a visual object motion upward in the image plane (corresponding to the peak at –100 degrees, 0 degrees being the direction parallel to the image *x* axis). A similar consideration applies to the other actions.

Having built this, the first interesting question is then whether this information (summarized collectively in Figure 18 and Figure 19) can be re-used when acting to generate anything useful showing exploitation of the object affordances. In fact, it is now possible to make the robot "optimally" poke an observed and known object. In practice the same color clustering procedure is used for localizing and recognizing the object, to determine its orientation on the table, its affordance, and finally to select the action that it is most likely to elicit the principal affordance (roll).

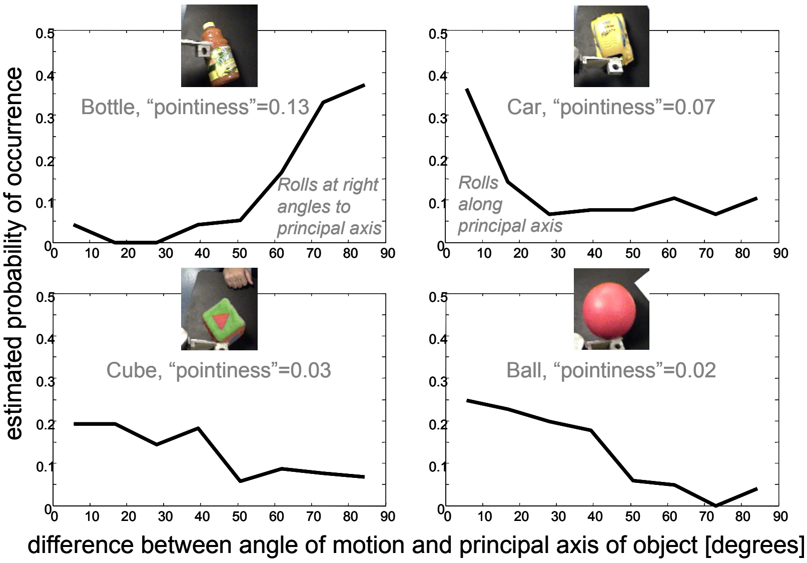

Figure 18: Probability of observing a roll along a particular direction with respect to the object principal axis. Abscissas are in degrees. Note as the toy car and the bottle show a very specific behavior: they possess a preferred rolling direction with respect to their principal axis.

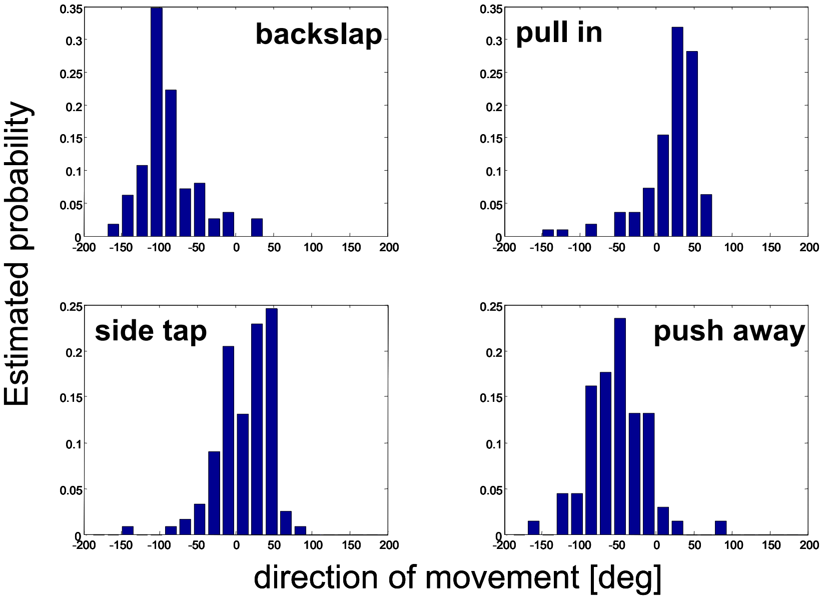

Figure 19: Histogram of the direction of movement of the object for each possible action. Abscissas are in degrees. This set of plots show that each action would generate on average a typical consequence on the approached object. The direction of motion expressed in angles is referred from the robot's point of view and it is relative to the image reference frame.

A simple qualitative test of the performance determined that out of 100 trials the robot made 15 mistakes. Further analysis showed that 12 of the 15 mistakes were due to poor control of reaching (e.g. the flipper touched the object too early bringing it outside the field of view), and only three to a wrong estimate of the orientation.

Although crude, this implementation shows that with little pre-existing structure the robot could acquire the crucial elements for building object knowledge in terms of their affordances. Given a sufficient level of abstraction, our implementation is close to the response of canonical neurons in F5 and their interaction with neurons observed in AIP that respond to object orientation (Sakata, Taira, Kusunoki, Murata, & Tanaka, 1997). Another interesting question is whether knowledge about object directed actions can be reused in interpreting observed actions performed perhaps by a human experimenter. It leads directly to the question of how mirror neurons can be developed from the interaction of canonical neurons and some additional processing.

To link with the concept of feedback from the action system, here, after the actual action has unfolded, the robot applied exactly the same procedure employed to learn the object affordances to measure the error between the planned and executed action. This feedback signal could then be exploited to incrementally update the internal model of the affordances. This feedback signal is fairly similar to the feedback signal identified in our conceptual model in section 1.3 (Figure 5).

2.1.6. Grasping an object on the table

This experiment shows the integration of some of the modules described in the previous sections. With reference to Figure 20, the experiment starts when an object is placed in the palm of the robot (frame 1). The pressure on the palm elicits a grasping action; the fingers flex toward the palm to close around the object. At this point the robot brings the object close to the eyes while maintaining fixation on the hand (frame 2 and 3). When the object is fixated a few frames are captured at the center of the cameras to train the object recognition algorithm described in section 2.1.4. After the probabilities are estimated – this requires moving the arm around at random locations – the object is released (frame 4). The rationale of this exploration phase is to gather several views of the object with possibly different backgrounds for the correct estimation of the probabilities of the various blobs.

The robot can now spot the object it has seen before in a cluttered background with possibly other objects, it can then fixate it and finally grasp it (frames 5-9). Haptic information is used to detect if the grasp was successful (the shape of the hand at the end of the grasp); if failure is detected the robot starts looking for the object again and performs another trial, otherwise it waits until another object is placed on the palm. The search for the object is carried out by the attention system described in section 2.1.3.

A few aspects need to be explained in greater detail. The hand motor commands are always preprogrammed; the robot uses three primitives to close the hand after pressure is detected on the palm, and to pre-shape and flex the fingers around the object during active grasping. The correct positioning of the fingers is achieved by exploiting passive adaptation and the intrinsic elasticity of the hand (see for instance (Natale, 2004)). The arm trajectory in also in part preprogrammed to approach the object from above, increasing the probability of success. This is obtained by including waypoints in joint space relative to the final arm posture (the latter is retrieved from the reaching map). No other knowledge is required by the robot to perform the task, as the object to be grasped is not known in advance.

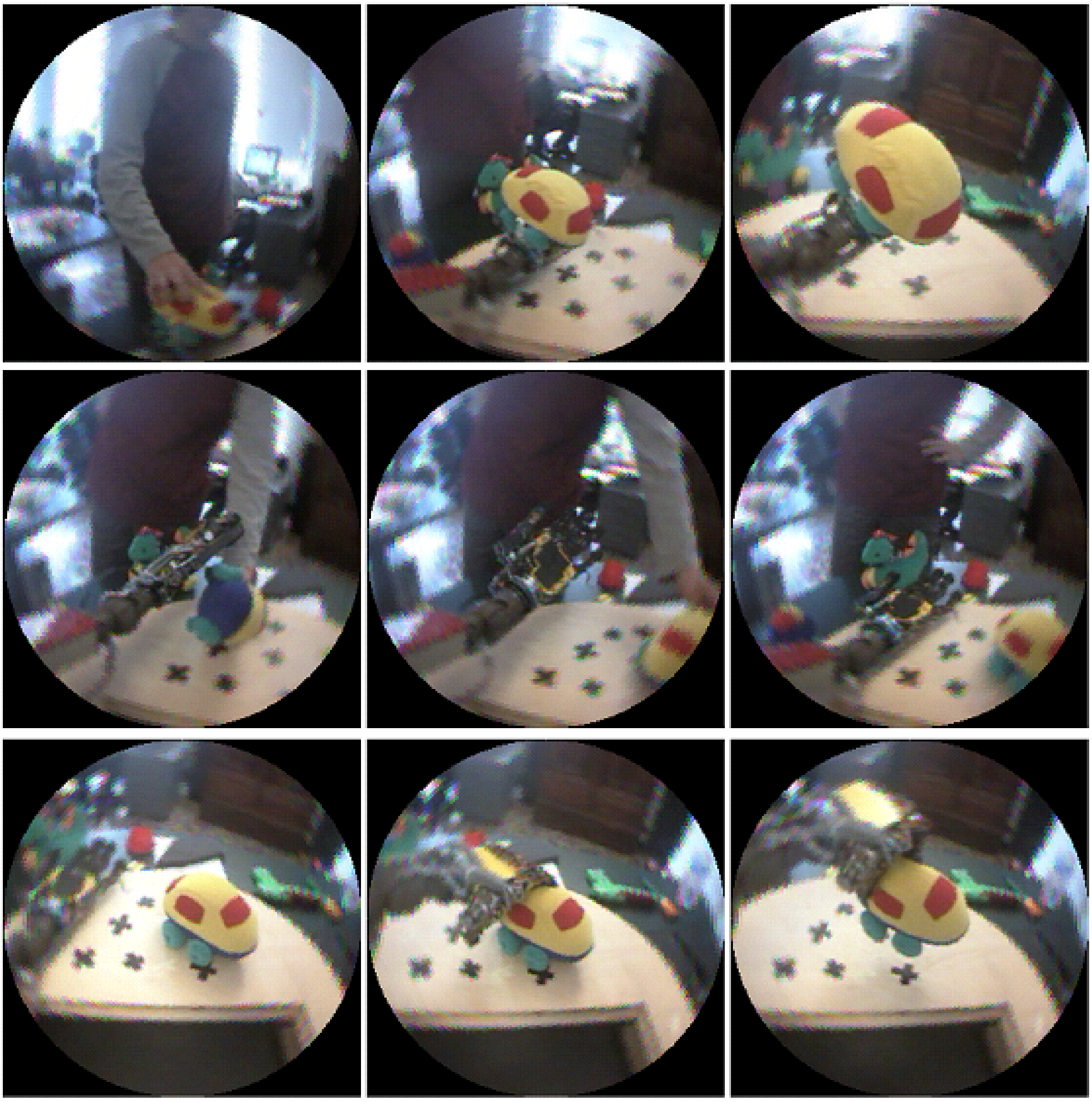

Figure 20: Grasping sequence (robot's point of view, left eye). At frame 1 a human places a toy in the robot's palm. Tactile feedback initiates a clutching action of the hand around the toy, while at the same time the robot begins moving the eyes to fixate the hand (frames 2 and 3). Once fixation has been achieved a few frames at the center of the cameras are captured to train the object recognition algorithm (frame 3); at frame 4 the toy is released. The robot then starts to search for the toy to grasp it. The object is hence localized, fixated and finally grasped (frames 6-9).

2.1.7. Reusing information for understanding actions

In answering the question of what is further required for interpreting observed actions, we could reason backward through the chain of causality employed in the previous sections. Whereas the robot identified the motion of the object because of a certain action applied to it, here it could backtrack and derive the type of action from the observed motion of the object. It can further explore what is causing motion and learn about the concept of manipulator in a more general setting (Fitzpatrick & Metta, 2003).

In fact, the same segmentation procedure cited in section 2.1.5 could visually interpret poking actions generated by a human as well as those generated by the robot. One might argue that observation could be exploited for learning about object affordances. This is possibly true to the extent passive vision is reliable and action is not required. Unfortunately passive observation could never learn (autonomously) the link to motor control as we showed in the affordance experiments. Also, in the active case, the robot can always tune/control the amount of information impinging on its visual sensors by, for instance, controlling the speed and type of action, which might be especially useful given the limitations of artificial perceptual systems.

Thus, observations can be converted into interpreted actions. The action whose effects are closest to the observed consequences on the object (which we might translate into the goal of the action) is selected as the most plausible interpretation given the observation. Most importantly, the interpretation reduces to the interpretation of the “simple” kinematics of the goal and consequences of the action rather than to understanding the “complex” kinematics of the human manipulator. The robot understands only to the extent it has learned to act.

One might note that a refined model should probably include visual cues from the appearance of the manipulator into the interpretation process. This is possibly true for the case of manipulation with real hands where the configuration of fingers might be important. Given our experimental setup the sole causal relationship was instantiated between the approach/poking direction and the object behavior; consequently there was not any apparent benefit in including additional visual cues.

The last question we propound to address is whether the robot can imitate the “goal” of a poking action. The step is indeed small since most of the work is actually in interpreting observations. Imitation was generated in the following by replicating the latest observed human movement with respect to the object and irrespective of its orientation. For example, in case the experimenter poked the toy car sideways, the robot imitated him/her by pushing the car sideways. Figure 21 shows an extended mimicry experiment with different situations originated by playing with a single object.

In humans there is now considerable evidence that a similar strict interaction of visual and motor information is at the basis of action understanding at many levels, and if exchanging vision for audition, it applies unchanged to speech (Fadiga et al., 2002). This implementation, besides serving as sanity check to our current understanding of the mirror system, provides hints that learning of mirror neurons can be carried out by a process of autonomous development.

However, these results have to be considered to the appropriate level of abstraction and comparing too closely to neural structure might even be misleading: simply this implementation was not intended to reproduce closely the neural substrate (the neural implementation) of imitation. Robotics, we believe, might serve as a reference point from which to investigate the biological solution to the same problem – although it cannot provide the answers, it can at least suggest useful questions.

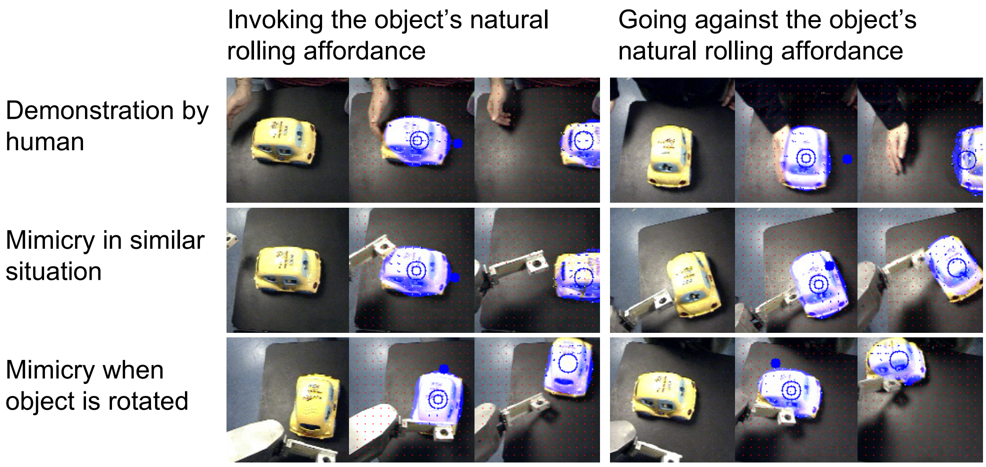
Invoking the object's natural rolling affordance
Going against the object's natural rolling affordance
Demonstration by human
Mimicry in similar situation
Mimicry when object is rotated

Figure 21: An extended imitation experiment. Here two different type of experiments are shown: acting according to an affordant behavior (left), or against the principal affordance (right). Further, for each situation we show the demonstrated action, the mimicry obtained when the object is oriented similarly to the demonstration, and when the object is oriented differently from the demonstration.

2.1.8. Action-level imitation

Although the behavior level imitation was addressed in the previous set of experiments, it is still interesting to investigate which are the requirements for imitating more sophisticate trajectories. As in the previous example, this can be sub-divided in several developmental phases (that could be possibly interleaved in the general picture presented in Section 1.4):

- Learning visuo-motor maps.
- Learning the View-Point Transformation (VPT).
- Imitating observed gestures.

As described in the Y2 progress report and in section 2.1.1, the first developmental phase corresponds to learning how to associate motor to visual information. During this phase, the robot performs arm movements and observes the visual consequences of those movements. Through this process, the robot can learn how to predict the visual appearance or the geometry of its own arm, when the arm is maintained in a certain configuration. These maps are learned through the use of sequential neural networks that estimate the relationship between some (or all) of the arm's degrees of freedom and its configuration in the image. This process, as described earlier, allows an initial level of eye-hand coordination as shown in Figure 22. Details of the method can be found in (Cabido Lopes et al., 2004).

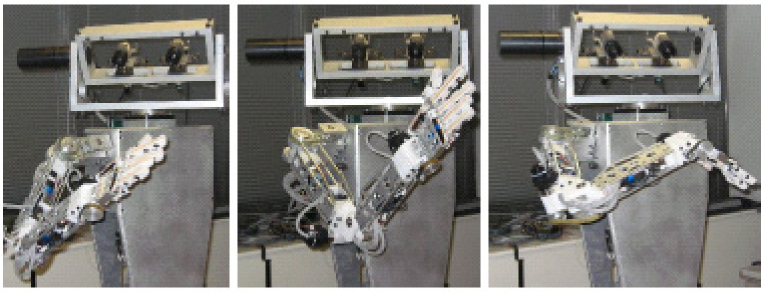

Figure 22: Head/hand Coordination example after learning hand-eye mapping.

In order to imitate observed gestures, the robot must be able to "recognize" those gestures or actions when seen from very different perspectives (associating gestures performed by the demonstrator to the subject's own gestures). For this reason, we have proposed a method that allows the system to "rotate" the view of the gestures enacted by the demonstrator (allo-image) into the corresponding image (ego-image) that represents the gesture as performed by the observer. We call this process the View-Point Transformation (VPT) (Cabido Lopes & Santos-Victor, 2003b) and it can be seen as part of the well-known "correspondence problem" in learning by imitation.

The VPT allows the robot to map observed gestures to a canonical point-of-view. The final step consists in transforming these mapped features to motor commands, which can be accomplished using the learned Visuo-motor Map (VMM).

The Visuo-Motor Map could be computed explicitly if the parameters of the arm-hand-eye configuration are known *a priori* but, more interestingly, it can be learned from observations of arm/hand movements, during an initial period of self-observation. Once the VMM has been estimated, the robot can observe a demonstrator, use the VPT to transform the image features to a canonical reference frame and map these features to motor commands through the VMM. The final result will be a posture similar to that observed.

2.1.9. Segmentation

In a first experiment we show how the robot mimics free arm movements presented by a human demonstrator. The robot needs to track the demonstrator's arm visually, estimate the position of the arm and simultaneously reproduce the gesture. In this section we present the necessary steps to do this.

In order to model the arm position we need three steps of segmentation: identification of the background, of the person and of the hand. The background is modeled by considering the intensity of each pixel, as a Gaussian random variable, during initialization. We need about 100 frames to be able to produce a good background model. After this process is completed, we can estimate the probability of each pixel being part of the background. In order to increase the robustness of segmentation to variation of the lighting conditions, we use an RGB color scheme normalized to the blue channel.

The position of the person is estimated by template matching. A simple template allows detecting the position of a person within an image. By scaling the template we can estimate the size of the person and a scale parameter of the camera model. In addition, if we need to detect whether the person is rotated with respect to the camera, we can scale the template independently in each direction, and estimate this rotation by the ratio between the head height and shoulder width. To detect the hand, we used a skin-color segmentation process.

Figure 23 shows a result of hand segmentation.

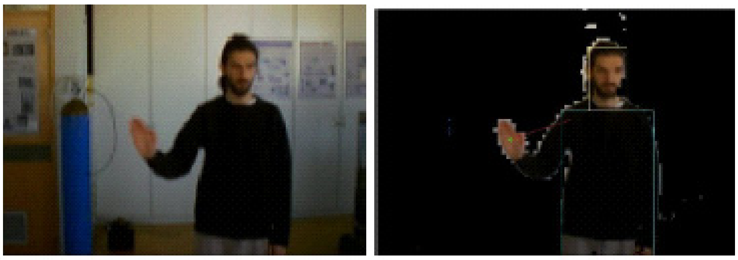

Figure 23: Vision system. Left: original image. Right: background segmentation with human (the frame corresponds to the template matching) and hand detection.

2.1.10. Imitation

If we assume that the movement of the hand is constrained to a plane or that the depth changes are small, we can use a simple View-Point Transformation to estimate the position of the person. The system is able to imitate the demonstrator in real-time. Results are shown in Figure 24. When approaching singularities, the arm might exhibit some irregular behavior, but techniques are available for correcting this issue if it would prove to be crucial. On the other hand, this problem occurs only when the arm is aligned with the first joint, after which the shoulder moves upwards.

This experiment shows, although summarily, results toward the imitation of generic gestures demonstrated in front of the robot. It is important to notice that a sequence of developmental phases is needed to achieve this goal with a similar motivation as the one introduced earlier in this manuscript.

Clearly, the imitation of the kinematics of gestures is not the ultimate goal of the autonomous system. As we described already, mirror neurons are related to goal of the gesture rather than to its exact geometric reproduction. Nonetheless, it is important to consider how certain aspects of the generation of proper trajectories would be included in a complete system. Whether this lower-level imitation is related to mirror neurons is another matter.

For the detailed analysis of the hand gestures, not considered here, we can resort to the approach introduced during Y2 of MIRROR, where we showed how the artificial system could learn how to recognize grasping by observing grasping sequences performed by a demonstrator and, more importantly, by relying on motor information.

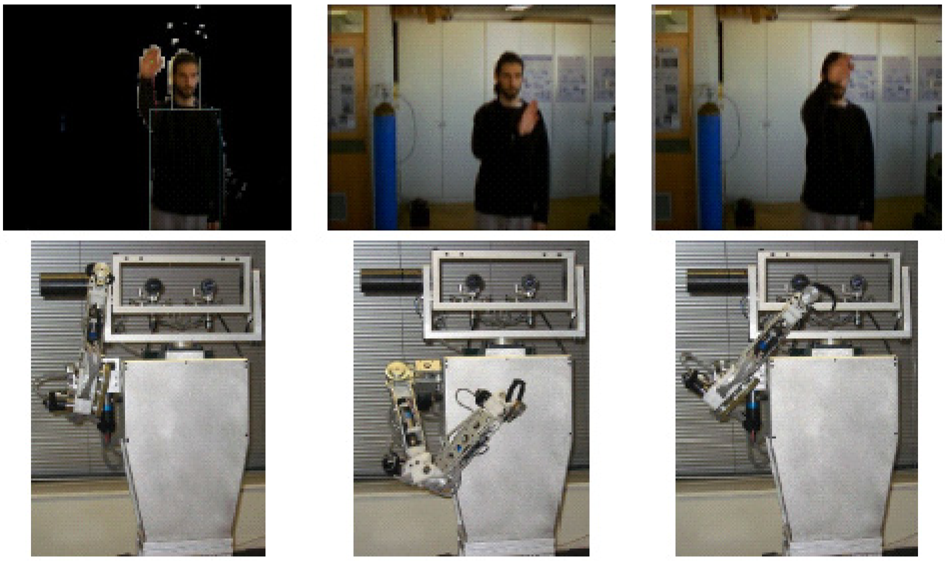

Figure 24: Imitation of a person's movements

2.1.11. Discussion

We have proposed a possible developmental path for a humanoid robot. The experiments described focus on the steps which allow the robot to learn to reach objects on a table. The knowledge initially provided to the robot consists of a set of stereotyped behaviors, basic perceptual abilities and learning rules. No prior knowledge about the objects to be grasped is assumed.

The robot starts by learning to control and recognize its own body (control of gaze and arm, hand localization); the first experiment showed how it is possible to build a model of the hand to allow the robot to detect and distinguish it in the environment. In the second experiment we describe how this knowledge can be used to learn to interact with the environment by reaching for objects. A few points are worth stressing. First, learning is fully online and it is not separated from the normal functioning of the robot. Second, all stages of development are required and equally important. Thus, reaching cannot start (and improve) if gaze is not controlled or the robot has not learnt to localize the hand.

Learning to act is an essential requirement to start interaction with the environment. By properly moving the arm in the workspace, in fact, the robot can try simple actions like pushing or pulling an object on a table. Even this simple form of interaction proves sufficient for developing more sophisticated perceptual abilities. This was shown in some of our previous works (Natale, Rao and Sandini 2002, Fitzpatrick et al. 2003, Metta and Fitzpatrick 2003) where we illustrated how a humanoid robot can learn to push/pull an object in different directions or even imitate pushing/pulling actions performed by another agent (a human). This stresses once more how important is the physical interaction between the agent and the world during ontogenesis; the motor repertoire of actions that is discovered and learnt by the agent in fact constitutes a reference system that can be used to map events that happen in the environment thus adding meaning to them. For this reason we believe that learning to act is at the basis of higher level functions like action/event recognition, interpretation and imitation.

Finally, in the last few experiments, we show how the motor and perceptual abilities developed in these initial stages can be integrated meaningfully. The resulting behavior allows the robot to autonomously acquire visual and haptic information about objects in the environment (another experiment is this direction is reported in (Natale, Metta and Sandini 2004)). Further, we showed how additional kinematic mimicry and the view-point transformation problem could be easily introduced into the robotic system. Although, unfortunately, not all the modules can run on the same platform, we believe these experiments contribute to a better understanding of the mirror system, and especially it can shed light on the whys of the existence of such a tight integration of motoric information into perception.

3. References

Abrams, R. A., & Dobkin, R. S. (1994). Inhibition of return: effects of attentional cuing on eye movement latencies. *Journal of Experimental Psychology: Human Perception and Performance, 20*(3), 467-477.

Aloimonos, J., Weiss, I., & Bandyopadhyay, A. (1988). Active Vision. *International Journal of Computer Vision, 1*(4), 333-356.

Arbib, M. A. (1981). Perceptual Structures and Distributed Motor Control. In V. B. Brooks (Ed.), *Handbook of Physiology* (Vol. II, Motor Control, pp. 1449-1480): American Physiological Society.

Ballard, D. H., & Brown, C. M. (1992). Principles of Animate Vision. *Computer Vision Graphics and Image Processing, 56*(1), 3-21.

Bertenthal, B., & von Hofsten, C. (1998). Eye, Head and Trunk Control: the Foundation for Manual Development. *Neuroscience and Behavioral Reviews, 22*(4), 515-520.

Brooks, R. A., Brezeal, C. L., Marjanovic, M., & Scassellati, B. (1999). The COG project: Building a Humanoid Robot. In *Lecture Notes in Computer Science* (Vol. 1562, pp. 52-87): Elsevier.

Cabido Lopes, M., Beira, R., Praça, M., & Santos-Victor, J. (2004, October 2004). *An anthropomorphic robot torso for imitation: design and experiments.* Paper presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan.

Cabido Lopes, M., & Santos-Victor, J. (2003a, October 31st). *Motor Representations for Hand Gesture Recognition and Imitation.* Paper presented at the IROS, Workshop on Robot Programming by Demonstration, Las Vegas, USA.

Cabido Lopes, M., & Santos-Victor, J. (2003b). *Visual transformations in gesture imitation: what you see is what you do.* Paper presented at the International Conference on Robotics and Automation, Taipei, Taiwan.

Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research, 91*(1), 176-180.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience, 15*(2), 399-402.

Fagg, A. H., & Arbib, M. A. (1998). Modeling parietal-premotor interaction in primate control of grasping. *Neural Networks, 11*(7-8), 1277-1303.

Fitzpatrick, P. (2003a, October 27-31). *First Contact: an active vision approach to segmentation.* Paper presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vagas, Nevada, USA.

Fitzpatrick, P. (2003b). *From First Contact to Close Encounters: A developmentally deep perceptual system for a humanoid robot.* Unpublished PhD thesis, MIT, Cambridge, MA.

Fitzpatrick, P., & Metta, G. (2003). Grounding vision through experimental manipulation. *Philosophical Transactions of the Royal Society: Mathematical, Physical, and Engineering Sciences, 361*(1811), 2165-2185.

Fitzpatrick, P., Metta, G., Natale, L., Rao, S., & Sandini, G. (2003). *Learning about objects through action: initial steps towards artificial cognition.* Paper presented at the International conference on Robotics and Automation, Taipei, Taiwan.

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

Fogassi, L., Gallese, V., Fadiga, L., & Rizzolatti, G. (1998). Neurons responding to the sight of goal directed hand/arm actions in the parietal area PF (7b) of the macaque monkey. *Society of Neuroscience Abstracts, 24*, 257.255.

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

Gallese, V., Fogassi, L., Fadiga, L., & Rizzolatti, G. (2002). *Action representation and the inferior parietal lobule.* Paper presented at the In Attention & Performance XIX. Common mechanisms in perception and action, Oxford.

Gibson, B. S., & Egeth, H. (1994). Inhibition of return to object-based and environment-based locations. *Perception & Psychophysics, 55*(3), 323-339.

Graziano, M. S. A. (1999). Where is my arm? The relative role of vision and proprioception in the neuronal representation of limb position. *Proceedings of the National Academy of Science, 96*, 10418-10421.

Graziano, M. S. A., Cooke, D. F., & Taylor, C. S. R. (2000). Coding the location of the arm by sight. *Sience, 290*, 1782-1786.

Hyvarinen, J. (1982). Posterior parietal lobe of the primate brain. *Physiology Reviews, 62*, 1060-1129.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*(3), 307-354.

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

Kerzel, D., & Bekkering, H. (2000). Motor activation from visible speech: Evidence from stimulus response compatibility. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 634-647.

Klein, R. M. (1988). Inhibitory tagging system facilitates visual search. *Nature, 334*, 430-431.

Kohler, E., Keysers, C., Umilta, M. A., Fogassi, L., Gallese, V., & Rizzolatti, G. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science, 297*, 846-848.

Kovacs, I. (2000). Human development of perceptual organization. *Vision Research, 40*, 1301-1310.

Leinonen, L., & Nyman, G. (1979). II. Functional properties of cells in anterolateral part of area 7 associative face area of awake monkeys. *Experimental Brain Research, 34*, 321-333.

Li, X., & Yuan, T. (2003). Adaptive color quantization based on perceptive edge protection. *Pattern Recognition Letters, 24*, 3165-3176.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431-461.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1-36.

Liberman, A. M., & Wahlen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Neuroscience, 4*(5), 187-196.

Mataric, M. J. (2000). Getting Humanoids to Move and Imitate. *IEEE Intelligent Systems,* 18-24.

Metta, G. (2000). *Babybot: a Study on Sensori-motor Development.* Unpublished Ph.D. Thesis, University of Genova, Genova.

Metta, G., & Fitzpatrick, P. (2003). Early Integration of Vision and Manipulation. *Adaptive Behavior, 11*(2), 109-128.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., & Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology*(78), 2226-2230.

Nakanishi, J., Morimoto, J., Endo, G., Schaal, S., & Kawato, M. (2003, October 31st). *Learning from demonstration and adaptation of biped locomotion with dynamical movement primitives.* Paper presented at the Workshop on Robot Learning by Demonstration, IEEE International Conference on Intelligent Robots and Systems, Las Vegas, USA.

Napier, J. (1956). The prehensile movement of the human hand. *Journal of Bone and Joint Surgery, 38B*(4), 902-913.

Natale, L. (2004). *Linking action to perception in a humanoid robot: a developmental approach to grasping.* Unpublished PhD, University of Genoa, Genoa.

Natale, L., Metta, G., & Sandini, G. (2002). Development of Auditory-evoked Reflexes: Visuo-acoustic Cues Integration in a Binocular Head. *Robotics and Autonomous Systems, 39*(2), 87-106.

Natale, L., Rao, S., & Sandini, G. (2002). *Learning to act on objects.* Paper presented at the Second International Workshop BMCV2002, Tuebingen, Germany.

Oztop, E., & Arbib, M. A. (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics, 87*, 116-140.

Panerai, F., Metta, G., & Sandini, G. (2002). Learning Stabilization Reflexes in Robots with Moving Eyes. *Neurocomputing, In press*.

Papoulis, A., & Pillai, S. U. (1991). *Probability, Random Variables and Stochastic Processes* (International Edition ed.). Singapore: McGraw-Hill.

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., et al. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146*, 87-113.

Perrett, D. I., Mistlin, A. J., Harries, M. H., & Chitty, A. J. (1990). Understanding the visual appearance and consequence of hand action. In M. A. Goodale (Ed.), *Vision and action: the control of grasping* (pp. 163-180). Norwood (NJ): Ablex.

Posner, M. I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X: Control of language processes* (pp. 531-556): Erlbaum.

Pouget, A., Ducom, J.-C., Torri, J., & Bavelier, D. (2002). Multisensory spatial representations in eye-centered coordinates for reaching. *Cognition, 83*, B1-B11.

Rizzolatti, G., Camarda, R., Fogassi, L., Gentilucci, M., Luppino, G., & Matelli, M. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. *Experimental Brain Research, 71*(3), 491-507.

Rizzolatti, G., & Fadiga, L. (1998). Grasping objects and grasping action meanings: the dual role of monkey rostroventral premotor cortex (area F5). In G. R. Bock & J. A. Goode (Eds.), *Sensory Guidance of Movement, Novartis Foundation Symposium* (pp. 81-103). Chichester: John Wiley and Sons.

Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*(2), 131-141.

Rizzolatti, G., & Luppino, G. (2001). The cortical motor system. *Neuron, 31*, 889-901.

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

Rose, C., Cohen, M. F., & Bodenheimer, B. (1998). Verbs and Adverbs: Multidimensional Motion Interpolation. *IEEE Computer Graphics & Applications, 18*(5), 32-40.

Sakata, H., Taira, M., Kusunoki, M., Murata, A., & Tanaka, Y. (1997). The TINS lecture - The parietal association cortex in depth perception and visual control of action. *Trends in Neurosciences, 20*(8), 350-358.

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

Schiele, B., & Crowley, J. L. (1996a). *Probabilistic object recognition using multidimensional receptive field histograms.* Paper presented at the The 13th International Conference on Pattern Recognition, Vienna, Austria.

Schiele, B., & Crowley, J. L. (1996b). *Where to look next and what to look for.* Paper presented at the IROS, Osaka, Japan.

Schiele, B., & Crowley, J. L. (2000). Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision, 36*(1), 31-50.

Smet, P. D., & Pires, R. (2000). *Implementation and analysis of an optimized rainfalling watershed algorithm.* Paper presented at the IS&T/SPIE's 12th Annual Symposium Electronic Imaging - Science and Technology, Conference: Image and Video Communications and Processing, San Jose, California, USA.

Tipper, S. P. (1991). Object-centred inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology, 43A*, 289-298.

Tipper, S. P. (1994). Object-based and environment-based inhibition of return of visual attention. *Journal of Experimental Psychology: Human Perception and Performance, 20*(3), 478-499.

Umilta, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., et al. (2001). I know what you are doing: a neurophysiological study. *Neuron, 31*, 155-165.

Vapnik, V. N. (1998). *Statistical Learning Theory*. New York: Wiley.

Vincent, L., & Soille, P. (1991). Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(6), 583-598.

Viola, P., & Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision, 57*(2), 137-154.

von Hofsten, C. (1983). Catching skills in infancy. *Experimental Psychology: Human Perception and Performance, 9*, 75-85.

von Hofsten, C., Vishton, P., Spelke, E. S., Feng, Q., & Rosander, K. (1998). Predictive action in infancy: tracking and reaching for moving objects. *Cognition, 67*(3), 255-285.

Wohlschlager, A., & Bekkering, H. (2002). Is human imitation based on a mirror-neurone system? Some behavioral evidence. *Experimental Brain Research, 143*, 335-341.

Wolpert, D. M. (1997). Computational approaches to motor control. *Trends in cognitive sciences, 1*(6), 209-216.

Wolpert, D. M., Ghahramani, Z., & Flanagan, R. J. (2001). Perspectives and problems in motor learning. *Trends in cognitive sciences, 5*(11), 487-494.

Woodward, A. L. (1998). Infant selectively encode the goal object of an actor's reach. *Cognition, 69*, 1-34.

Yoshikawa, Y., Hosoda, K., & Asada, M. (2003). *Does the invariance in multi-modalities represent the body scheme? A case study in vision and proprioception.* Paper presented at the Second intelligent symposium on adaptive motion of animals and machines, Kyoto, Japan.