Understanding mirror neurons

A bio-robotic approach

[1]Giorgio Metta, [1]Giulio Sandini, [1]Lorenzo Natale,
[2]Laila Craighero and [2]Luciano Fadiga
[1]LIRA-Lab, DIST, University of Genoa / [2]Neurolab, Section of Human Physiology, University of Ferrara

This paper reports about our investigation on action understanding in the brain. We review recent results of the neurophysiology of the mirror system in the monkey. Based on these observations we propose a model of this brain system which is responsible for action recognition. The link between object affordances and action understanding is considered. To support our hypothesis we describe two experiments where some aspects of the model have been implemented. In the first experiment an action recognition system is trained by using data recorded from human movements. In the second experiment, the model is partially implemented on a humanoid robot which learns to mimic simple actions performed by a human subject on different objects. These experiments show that motor information can have a significant role in action interpretation and that a mirror-like representation can be developed autonomously as a result of the interaction between an individual and the environment.

Keywords: mirror neurons, neurophysiology, robotics, action recognition

1. Introduction

Animals continuously act on objects, interact with other individuals, clean their fur or scratch their skin and, in fact, actions represent the only way they have to manifest their desires and goals. However, actions do not constitute a semantic category such as trees, objects, people or buildings: the best way to describe a complex act to someone else is to demonstrate it directly (Jeannerod, 1988). This is not true for objects such as trees or buildings that we describe by using size, weight, color, texture, etc. In other words we describe 'things' by using

Interaction Studies 7:2 (2006), **197–231**.
ISSN 1572–0373 / E-ISSN 1572–0381

visual categories and 'actions' by using motor categories. Actions are defined as 'actions' because they are external, physical expressions of our intentions. It is true that often actions are the response to external contingencies and/or stimuli but it is also certainly true that — at least in the case of human beings — actions can be generated on the basis of internal aims and goals; they are possibly symbolic and not related to immediate needs. Typical examples of this last category include most communicative actions.

Perhaps one of the first attempts of modeling perception and action as a whole was started decades ago by Alvin Liberman who initiated the construction of a 'speech understanding' machine (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Liberman & Wahlen, 2000). As one can easily imagine, the first effort of Liberman's team was directed at analyzing the acoustic characteristics of spoken words, to investigate whether the same *speech event*, as uttered by different subjects in different contexts, possessed any common *invariant phonetic percept*.[1] Soon Liberman and his colleagues realized that speech recognition on the basis of acoustic cues alone was beyond reach with the limited computational power available at that time. Somewhat stimulated by the negative result, they put forward the hypothesis that the ultimate constituents of speech (the events of speech) are not sounds but rather articulatory gestures that have evolved exclusively at the service of language. As Liberman states:

> *"A result in all cases is that there is not, first, a cognitive representation of the proximal pattern that is modality-general, followed by a translation to a particular distal property; rather, perception of the distal property is immediate, which is to say that the module*[2] *has done all the hard work (Liberman & Mattingly, 1985, page 7)".*

This elegant idea was however strongly debated at the time mostly because it was difficult to test, validation through the implementation on a computer system was impossible, and in fact only recently has the theory gained support from experimental evidence (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Kerzel & Bekkering, 2000).

Why is it that, normally, humans can visually recognize actions (or, acoustically, speech) with a recognition rate of about 99–100%? Why doesn't the inter-subject variability typical of motor behavior pose a problem for the brain while it is troublesome for machines? Sadly, if we had to rank speech recognition software by human standards, even our best computers would be regarded at the level of an aphasic patient. One possible alternative is for Liberman to be right and that speech perception and speech production use a common reper-

toire of motor primitives that during production are at the basis of the generation of articulatory gestures, and during perception are activated in the listener as the result of an acoustically-evoked motor "resonance" (Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Wilson, Saygin, Sereno, & Iacoboni, 2004).

Perhaps it is the case that if the acoustic modality were replaced, for example, by vision then this principle would still hold. In both cases, the brain requires a "resonant" system that matches the observed/heard actions to the observer/listener motor repertoire. It is interesting also to note that an animal equipped with an empathic system of this sort would be able to automatically "predict", to some extent, the future development of somebody else's action on the basis of the onset of the action and the implicit knowledge of its dynamics. Recent neurophysiological experiments show that such a motor resonant system indeed exists in the monkey's brain (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). Most interesting, this system is located in a premotor area where neurons not only discharge during action execution but to specific visual cues as well.

The remainder of the paper is organized to lead the reader from the basic understanding of the physiology of the mirror system, to a formulation of a model whose components are in agreement with what we know about the neural response of the rostroventral premotor area (area F5), and finally to the presentation of a set of two robotic experiments which elucidate several aspects of the model. The model presented in this paper is used to lay down knowledge about the mirror system rather than being the subject of direct neural network modelling. The two experiments cover different parts of the model but there is not a full implementation as such yet. The goal of this paper is that of presenting these results, mainly robotics, into a common context.

2. Physiological properties of monkey rostroventral premotor 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.

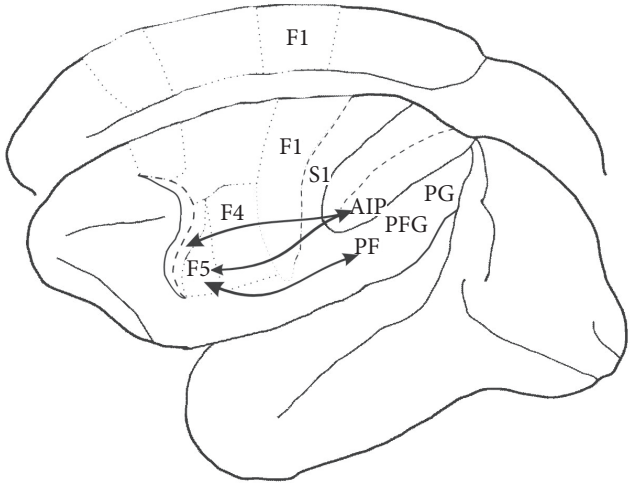

Figure 1.

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.

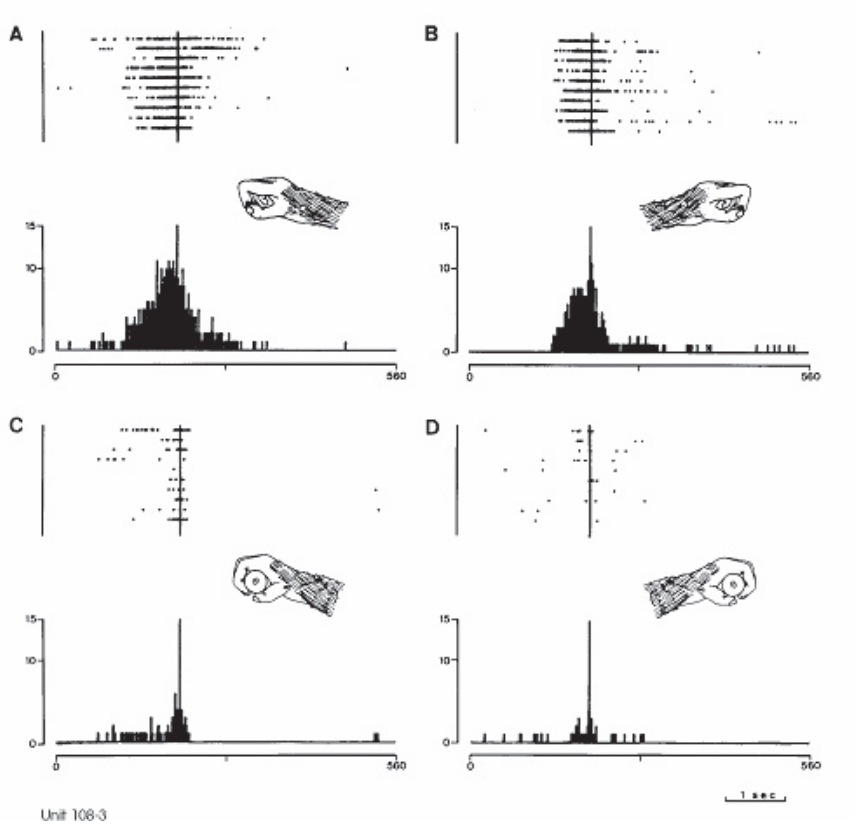

Figure 2.

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).

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, Fadiga, Fogassi, & Rizzolatti, 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).

2.2.1 *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 3A 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 Figure 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. In contrast, neurons selectively active when the monkey looks at large objects discharge also during actions directed towards large objects (e.g. whole hand prehension).

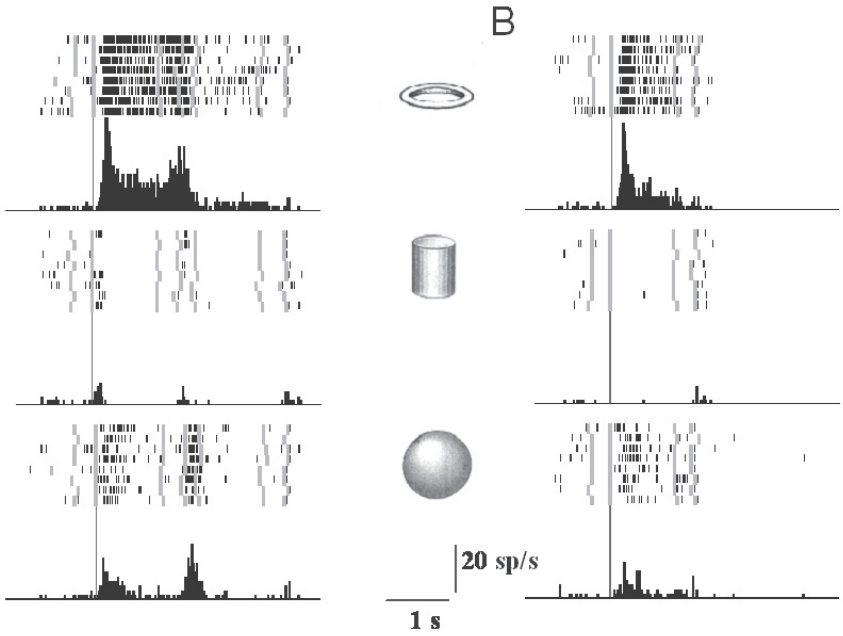

Figure 3.

2.2.2 *Mirror neurons*

Mirror neurons are F5 visuomotor neurons that activate 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, Fadiga, Fogassi, Gallese, & Rizzolatti, 1992; Gallese, Fadiga, Fogassi, & Rizzolatti, 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.

3. A model of area F5 and the mirror system

The results summarized in the previous sections tell us of the central role of F5 in the control and recognition of manipulative actions: the common interpretation proposed by (Luppino & Rizzolatti, 2000) and (Fagg & Arbib, 1998) considers F5 a part of a larger circuit comprising various areas in the parietal lobe (a large reciprocal connection with AIP), indirectly from STs (Perrett, Mistlin, Harries, & Chitty, 1990) and other premotor and frontal areas (Luppino & Rizzolatti, 2000). Certainly F5 is strongly involved in the generation and control of action indirectly through F1, and directly by projecting to motor and medullar interneurons in the spinal cord (an in-depth study is described in (Shimazu, Maier, Cerri, Kirkwood, & Lemon, 2004)). A good survey of the physiology of the frontal motor cortex is given in (Rizzolatti & Luppino, 2001).

A graphical representation of the F5 circuitry is shown in Figure 5. In particular, we can note how the response of F5 is constructed of various elements these being the elaboration of object affordances (canonical F5 neurons and AIP), of the visual appearance of the hand occurring in the Superior Temporal sulcus region (STs), and of the timing, synchronization of the action (Luppino & Rizzolatti, 2000). Parallel to F5-AIP, we find the circuit formed by F4-VIP that has been shown to correlate to the control of reaching. Further, a degree of coordination between these circuits is needed since manipulation requires both a transport and a grasping ability (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995).

In practice, the parieto-frontal circuit can be seen as the transformation of the visual information about objects into its motor counterpart, and with the addition of the Inferior Temporal (IT) and Superior Temporal Sulcus (STs) areas, this description is completed with the semantic/identity of objects and

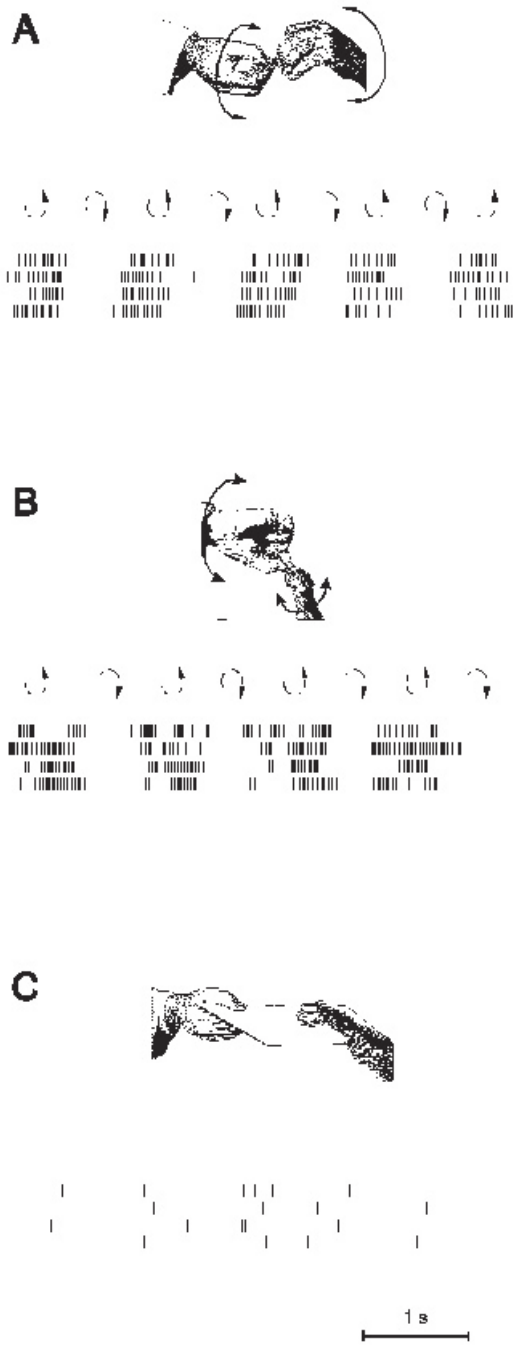

Figure 4.

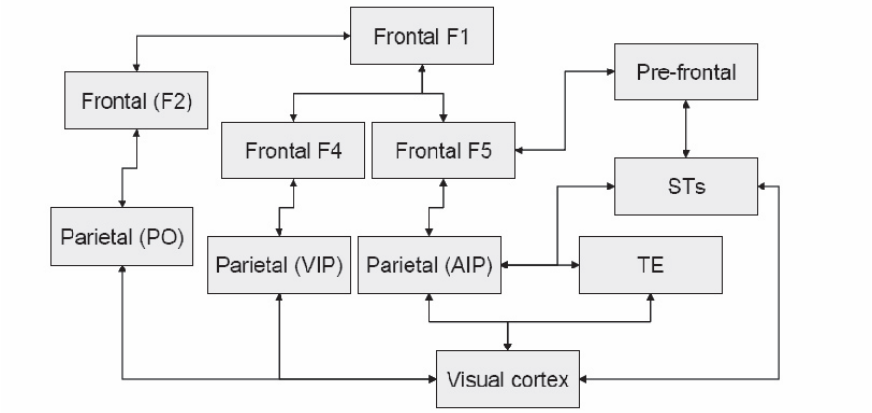

Figure 5.

the state of the hand. This description is also in good agreement with previous models such as the FARS (Fagg, Arbib, Rizzolatti, and Sakata model) and MN1 (Mirror Neuron model 1) (Fagg & Arbib, 1998). Our model is different in trying to explicitly identify a developmental route to mirror neurons by looking at how these different transformations could be learned without posing implausible constraints.

The model of area F5 we propose here revolves around two concepts that are likely related to the development of this unique area of the brain. The mirror system has a double role being both a controller (70% of the neurons of F5 are pure motor neurons) and a "classifier" system (being activated by the sight of specific grasping actions). If we then pose the problem in terms of understanding how such a neural system might actually autonomously develop (be shaped and learned by/through experience during ontogenesis), 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 part of the activation of F5.

3.1 Controller–predictor formulation

For explanatory purpose, the description of our model of the mirror system can be 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 6): in her/his perspective, the decision to undertake a particular grasping action is attained by the convergence in area F5 of many factors including context 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 for instance) and other frontal areas (mesial or dorsal area 6) as in other models of motor control systems (Fagg & Arbib, 1998; Haruno, Wolpert, & Kawato, 2001; Oztop & Arbib, 2002). For example, the contextual information required to switch from one model to another (determining eventually the grasp type) is represented exactly by the detection of affordances performed by AIP-F5.

With reference to Figure 6, our model hypothesizes 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 (i.e. without employing feedback). 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 the signals derived from sensory feedback, and differences accounted for (the cerebellum is believed to have a role in this (Miall, Weir, Wolpert, & Stein, 1993; Wolpert & Miall, 1996)). 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 these 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 (STs response) and the object information (AIP, IT, F5 canonical). Finally the predicted and the sensed signals arising from the motor act are compared and their difference (feedback error) sent back to the controller.

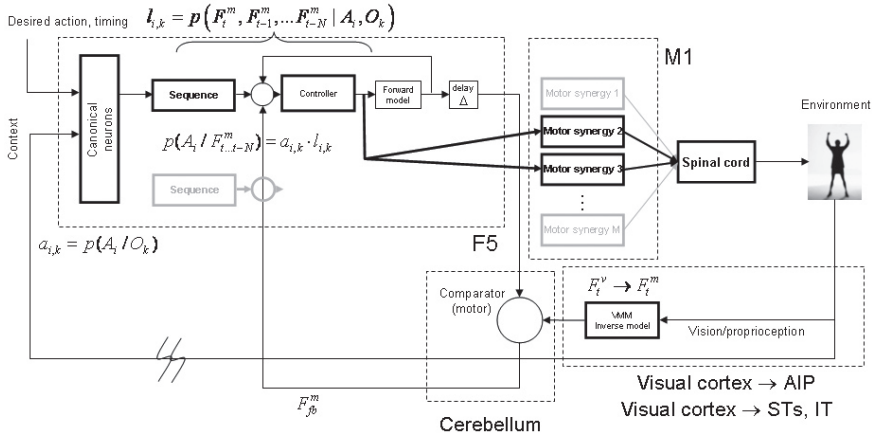

Figure 6.

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 (Kawato, Furukawa, & Suzuki, 1987).

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

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). Learning of 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 the observer situation (see Figure 7) motor and proprioceptive information is not directly available. The only readily available information is vision or sound. 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

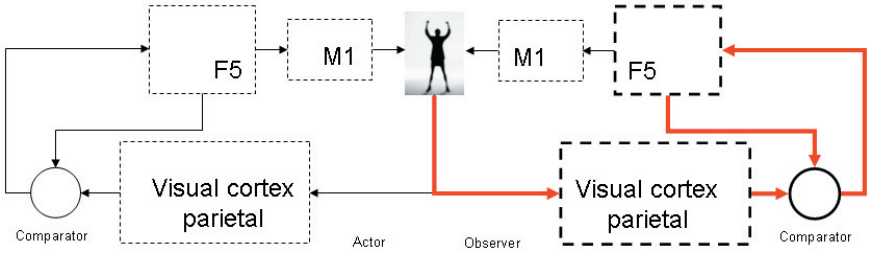

Figure 7.

is directly reaching in parallel the sensorimotor 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. It is important to note that canonical neurons are known to fire when the monkey is fixating an object irrespective of the actual context (tests were performed with the object behind a transparent screen or the monkey restrained to the chair).

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 (plausible implementations of this mechanism by means of a gating network is described in (Y. Demiris & Johnson, 2003; Haruno, Wolpert, & Kawato, 2001). While the predictor–controller formulation is somewhat well-established and several variants have been described in the literature (Wolpert, Ghahramani, & Flanagan, 2001), the evidence on the mirror system we reviewed earlier seems to support the idea that many factors, including the affordances of the target object, determine the recognition and interpretation of the observed action.

The action recognition system we have just outlined can be interpreted by following a Bayesian approach (Lopes & Santos-Victor, 2005). The details of the formulation are reported in the Appendix I.

3.2 Ontogenesis of mirror neurons

The presence of a goal is fundamental to elicit mirror neuron responses (Gallese, Fadiga, Fogassi, & Rizzolatti, 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.

The Bayesian interpretation we detail in Appendix I 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 ourselves of the significance of the evidence on mirror neurons. 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 actions 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 levels 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 then I can hypothesize that 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 a plausible way of autonomously learning a mirror representation. The learning problem is still a supervised[3] one but the information can now be collected autonomously through a procedure of exploration of the environment. The association between the canonical response (object-action) and the mirror one (including vision) is made when the observed consequences (or goal) are recognized as similar in the two cases — self or the other individual 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).

Finally, also the Visuo-Motor Map (VMM) as already mentioned can be learned through a procedure of self exploration. Motor commands and correlated visual information are two quantities that are readily available to the developing infant. It is easy to imagine a procedure that learns the inverse model on the basis of this information.

4. A machine with hands

The simplest way of confirming the hypothesis that motor gestures are the basis of action recognition, as amply discussed in the previous sections, is to equip a computer with means of "acting" on objects, collect visual and motor data and build a recognition system that embeds some of the principles of operation that we identified in our model (see Figure 8). In particular, the hypothesis we would like to test is whether the extra information available during learning (e.g. kinesthetic and tactile) can improve and simplify the recognition of the same actions when they are just observed: i.e. when only visual information is available. Given the current limitations of robotic systems the simplest way to provide "motor awareness" to a machine is by recording grasping actions of human subjects from multiple sources of information including joint angles, spatial position of the hand/fingers, vision, and touch. In this sense we speak of a machine (the recognition system) with hands.

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 birds, 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 in any way the movement of the arm and hand as long as the subject is sitting not too far from the glove interface. Data recording was carried out with the subject sitting comfortably in front of a table and performing 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.

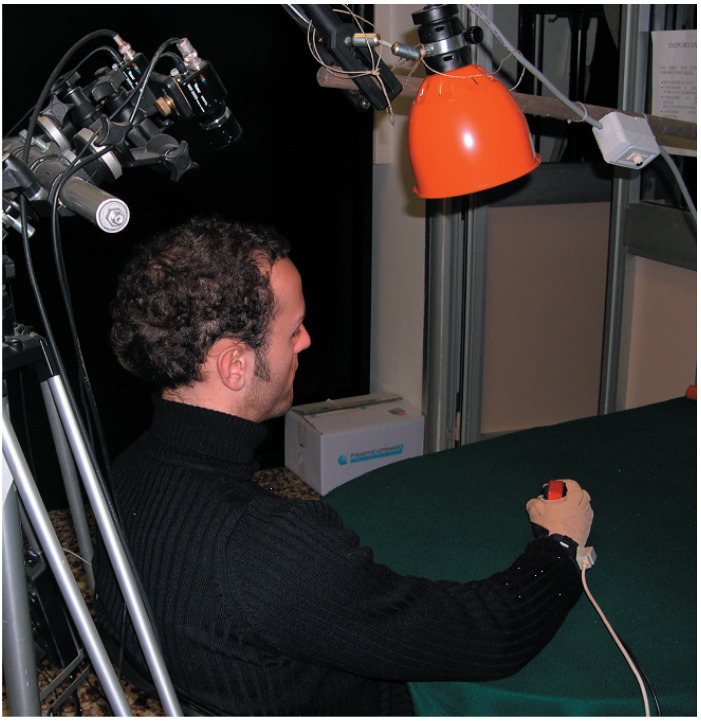

Figure 8.

We collected a large data set and processing was then performed off-line. 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 to what extent the system could learn invariances across different grasping types 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.

A set of three objects was employed in all our experiments: a small glass ball, a rectangular solid 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

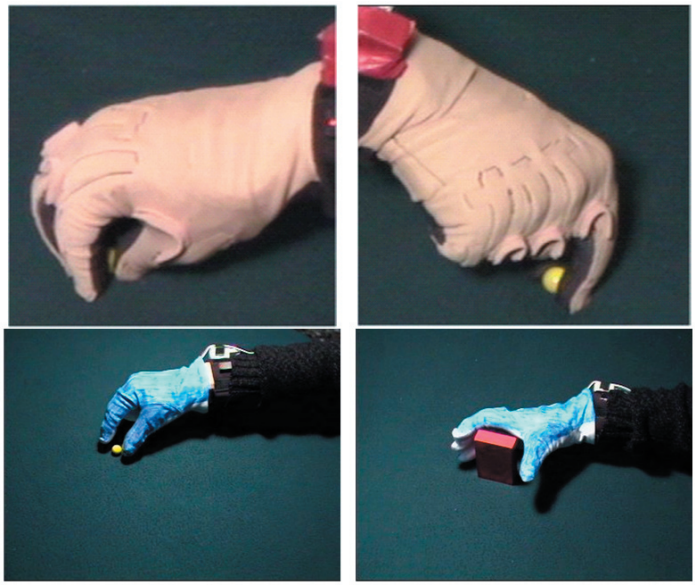

Figure 9.

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 images of the scene from the two cameras synchronized with the cyber glove and the magnetic tracker data. This is the data set that is used for building the Bayesian classifier outlined in the Appendix I (Lopes & Santos-Victor, 2005).

The visual features were extracted from pre-processed image data. The hand was segmented from the images through a simple color segmentation algorithm. The bounding box of the segmented region was then used as a reference to map the view of the hand to a standard reference size. The orientation of the color blob in the image was also used to rotate the hand to a standard orientation. This data set was then filtered through Principal Component Analysis (PCA) by maintaining only a limited set of eigenvectors corresponding to the first 2 to 15 largest eigenvalues.

One possibility to test the influence of motor information in learning action recognition is to contrast the situation where motor-kinesthetic information is available in addition to visual information with the control situation where only visual information is available.

The first experiment uses the output of the VMM as shown in Section 3.1 and thus employed motor features for classification. The VMM was approximated from data by using a simple backpropagation neural network with

Table 1. Summary of the results of the experiments. Left: motor information is available to the classifier. Right: only visual information is used by the classifier. Training was always performed with all available sequences after partitioning the data into equally-sized training and test set.

	Experiment (motor space)	**Control (visual space)**
	Training	
# of sequences	24 (+ VMM)	64
# of points of view	1	4
Classification rate (on the training set)	98%	97%
	Test	
# of sequences	96	32
# of points of view	4	4
Classification rate	97%	80%

sigmoidal units. The input of the VMM was the vector of the mapping of the images onto the space spanned by the first N PCA vectors; the output was the vector of joint angles acquired from the data glove.

As a control, a second experiment employed the same procedure but the VMM, and thus the classification was performed in visual space. The result of the two experiments is reported in the following Table 2.

The experiments were set up by dividing the database in two parts and training the classifier on one half and testing on the other. The number of training sequences is different since we chose to train the classifier with the maximum available data but from all points of view only in the control experiment. The clearest result of this experiment is that the classification in motor space is easier and thus the classifier performs better on the test set. Also, the number of Gaussians required by the EM algorithm to approximate the likelihood as per equation (1) within a given precision is smaller for the experiment than the control (1–2 vs. 5–7): that is, the distribution of data is more "regular" in motor space than in visual space. This is to be expected since the variation of the visual appearance of the hand is larger and depends strongly on the point of view, while the sequence of joint angles tends to be the same across repetitions of the same action. It is also clear that in the experiment the classifier is much less concerned with the variation of the data since this variation has been taken out by the VMM.

Overall, our interpretation of these results is that by mapping in motor space we are allowing the classifier to choose features that are much better suited for performing optimally, which in turn facilitates generalization. The same is not true in visual space.

5. Robotic experiment

Following the insight that it might be important to uncover the sequence of developmental events that moves either the machine or humans to a motoric representation of observed actions, we set forth to the implementation of a complete experiment on a humanoid robot called Cog (Brooks, Breazeal, Marjanović, & Scassellati, 1999). This is an upper-torso human shaped robot with 22 degrees of freedom distributed along the head, arms and torso (Figure 10). It lacks hands, it has instead simple flippers that could use to push and prod objects. It cannot move from its stand so that the objects it interacted with had to be presented to the robot by a human experimenter. The robot is controlled by a distributed parallel control system based on a real-time operating system and running on a set of Pentium based computers. The robot is equipped with cameras (for vision), gyroscopes simulating the human vestibular system, and joint sensors providing information about the position and torque exerted at each joint.

The aim of experimenting on the humanoid robot was that of showing that a mirror neuron-like representation could be acquired by simply relying on the information exchanged during the robot-environment interaction. This proof of concept can be used to analyze the gross features of our model or evidence

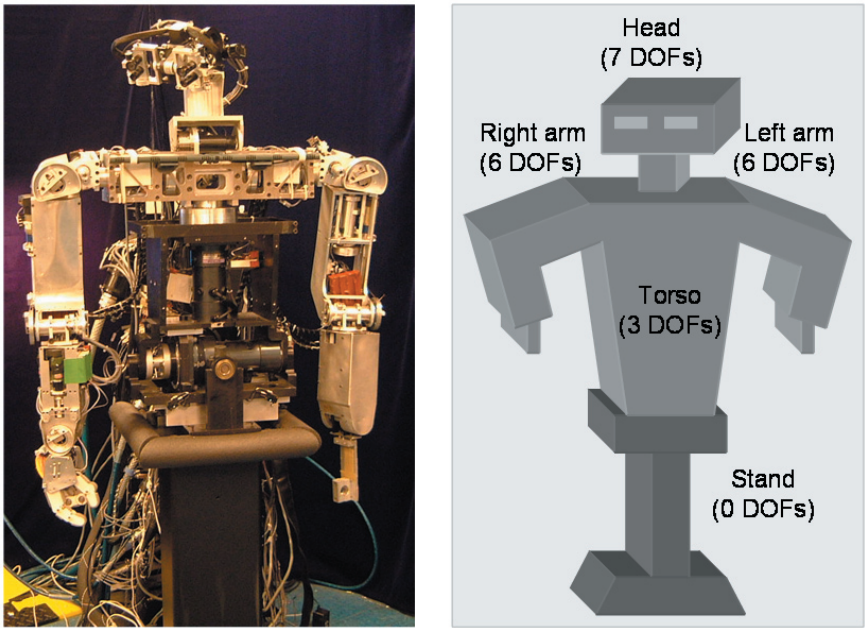

Figure 10.

any lacuna in it. We were especially interested in determining a plausible sequence that starting from minimal initial hypotheses steers the system toward the construction of units with responses similar to mirror neurons. The resulting developmental pathway should gently move the robot through probing different levels of the causal structure of the environment. Table 2 shows four levels of this causal structure and some intuition about the areas of the brain related to these functions. It is important to note as the complexity of causation evolves from strict synchrony to more delayed effects and thus it becomes more difficult to identify and learn anything from. Naturally, this is not to say that the brain develops following this step-like progression. Rather, brain development is thought to be fluidic, messy, and above all dynamic (Thelen & Smith, 1998); the identification of "developmental levels" here simplifies though our comprehension of the mechanisms of learning and development.

The first level in Table 2 suggests that learning to reach for externally identified objects requires the identification of a direct causal chain linking the generation of action to its immediate and direct visual consequences. Clearly, in humans the development of full-blown reaching requires also the simultaneous development of visual acuity, binocular vision and, as suggested by Bertenthal and von Hofsten (Bertenthal & von Hofsten, 1998), the proper support of the body freeing the hand and arm from its supporting role.

Only when reaching has developed then the interaction between the hand and the external world might start generating useful and reliable responses from touch and grasp. This new source of information requires simultaneously

Table 2. degrees of causal indirection, brain areas and function in the brain (this table has been compiled from the data in (Rizzolatti & Luppino, 2001).

Level	Nature of causation	Brain areas	Function and behavior	Time profile
1	Direct causal chain	VC-(VIP/7b)-F4-F1	Reaching	Strict synchrony
2	On level of indirection	VC-AIP-F5-F1	Object affordances, grasping (rolling in this experiment)	Fast onset upon contact, potential for delayed effects
3	Complex causation involving multiple causal chains	VC-AIP-F5-F1-STs-IT	Mirror neurons, mimicry	Arbitrarily delayed onset and effects
4	Complex causation involving multiple instances of manipulative acts	STs-TE-TEO-F5-AIP	Object recognition	Arbitrarily delayed onset and effects

new means of detecting causally connected events since the initiation of an action causes certain delayed effects. The payoff is particularly rich, since interaction with objects leads to the formation of a "well defined" concept of objecthood — this, in robotics, is a tricky concept as it has been discussed for example in (Metta & Fitzpatrick, 2003).

It is interesting to study subsequently whether this same knowledge about objects and the interaction between the hand and objects could be exploited in interpreting actions performed by others. It leads us to the next level of causal understanding where the delay between the acquisition of object knowledge and the exploitation of this knowledge when observing someone else might be very large. If any neural unit is active in these two situations (both when acting and observing) then it can be regarded in all respects as a "mirror" unit.

Finally we mention object recognition as belonging to an even higher level of causal understanding where object identity is constructed by repetitive exposure and manipulation of the same object. In the following experiments we concentrate on step 2 and 3 assuming step 1 is already functional. We shall not discuss about step 4 any longer since it is relatively more advanced with respect to the scope of this paper. The robot also possesses, and we are not going to enter much into the details here, some basic attention capabilities that allows selecting relevant objects in the environment and tracking them if they move, binocular disparity which is used to control vergence and estimate distances, and enough motor control abilities to reach for an object. In a typical experiment, the human operator waves an object in front of the robot which reacts by looking at it; if the object is dropped on the table, a reaching action is initiated, and the robot possibly makes a contact with the object. Vision is used during the reaching and touching movement for guiding the robot's flipper toward the object, to segment the hand from the object upon contact, and to collect information about the behavior of the object caused by the application of a certain action.

6. Learning object affordances

Since the robot does not have hands, it cannot really grasp objects from the table. Nonetheless there are other actions that can be employed in exploring the physical properties 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,

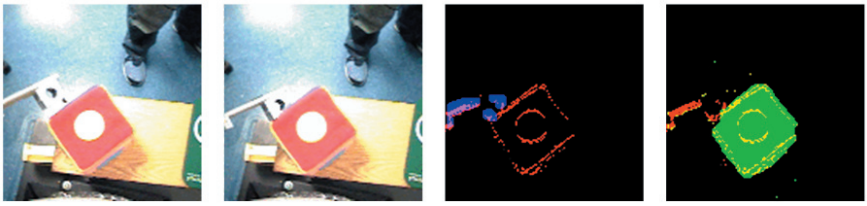

Figure 11.

2003a). An example of the quality of segmentation obtained is shown in Figure 11. 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 (the manipulator endpoint) 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.

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 ran 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 technique 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 sophisticated algorithms could be used (Fitzpatrick, 2003b).

Figure 12 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's own principal axis. Intuitively, this gives information about the rolling properties of the different objects: e.g. the car tends to roll along its principal axis, the bottle at

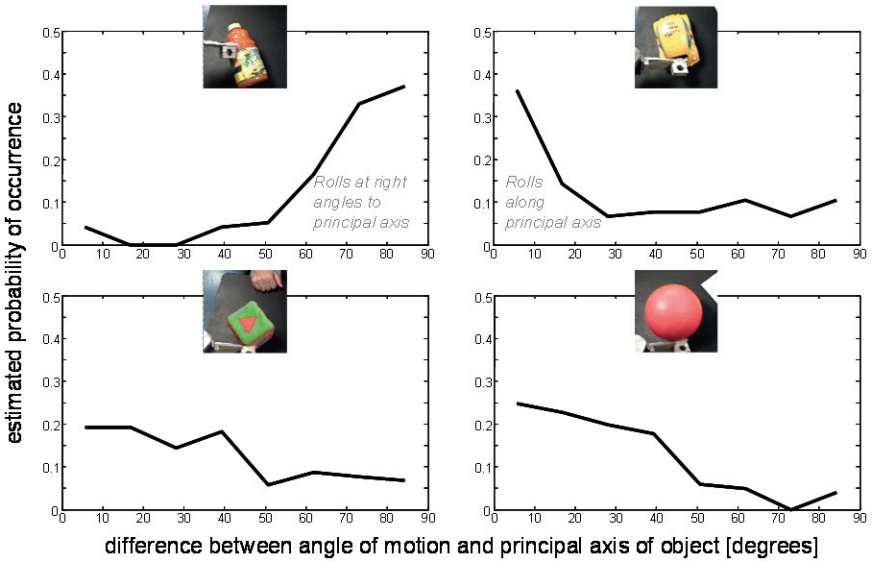

Figure 12.

right angle with respect to the axis. The training set for producing the graphs in Figure 12 consisted of about 100 poking actions per object. This "description" of objects is fine in visual terms but does not really bear any potential for action since it does not yet contain information about what action to take if the robot 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 13 shows the histograms of the direction of movement averaged for each possible action. About 700 samples were used to produce the four plots. Note, for example, that the action labeled as "backslap" (moving the object with the flipper 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; the y axis pointing downward). A similar consideration applies to the other actions.

Having built this, the first interesting question is then whether this information (summarized collectively in Figure 12 and Figure 13) 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 (i.e. selecting an action that causes maximum displacement) an observed and known object. In practice the same color clustering procedure is used for localizing

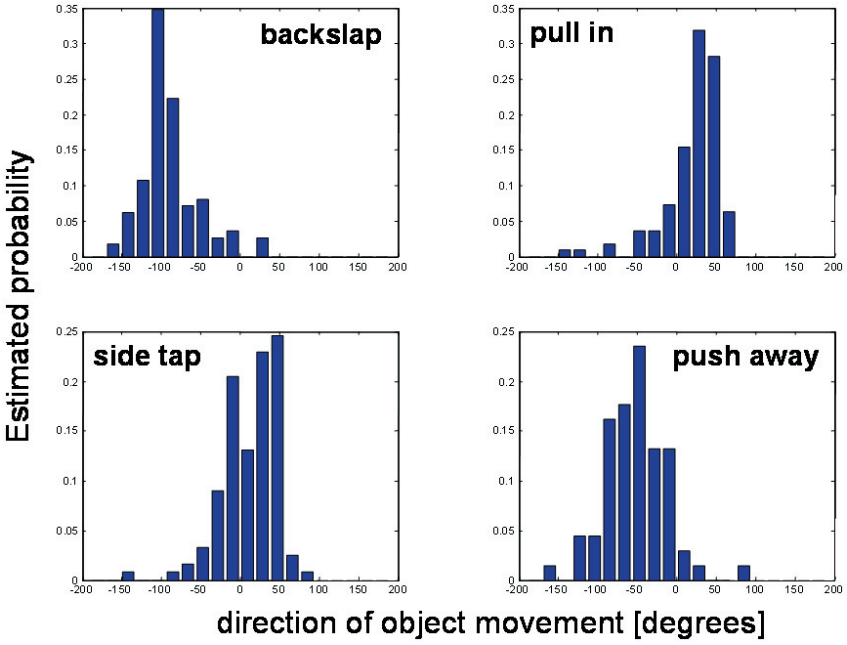

Figure 13.

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).

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 knowledge of objects 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 3 (Figure 6).

7. Developing mirror neurons

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 section. 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 6 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. In the architecture proposed by Demiris and Hayes (J. Demiris & Hayes, 2002), the authors make a distinction between active and passive imitation. In the former case the agent is already capable of imitating the observed action and recognizes it by simulating it (a set of inverse and forward models are simultaneously activated). In the latter case the agent passively observes the action and memorizes the set of postures that characterizes it. This view is not in contrast with our approach. In Demiris and Hayes, in fact, the system converts the available visual information and estimates the posture of the demonstrator (joint angles). Learning this transformation during ontogenesis is indeed an active process which requires access to both visual and proprioceptive information. The advantage of the active approach, at least for the robot, is that it allows controlling the amount of information impinging on the visual sensors by, for instance, controlling the speed and type of action. This strategy 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 14 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, Craighero, Buccino, & Rizzolatti, 2002). This implementation, besides serving as a 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 at the appropriate level of abstraction and comparing too closely to neural structure might even be misleading: simply this implementation was not meant 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 — and although it cannot provide the answers, it can at least suggest useful questions.

8. Conclusions

This paper put forward a model of the functioning of the mirror system which considers at each stage plausible unsupervised learning mechanisms. In addition, the results from our experiments seem to confirm two facts of the proposed model: first, that motor information plays a role in the recognition process — as would be following the hypothesis of the implication of feedback signals into recognition — and, second, that a mirror-like representation can be developed autonomously on the basis of the interaction between an individual and the environment.

Other authors have proposed biologically inspired architectures for imitation and learning in robotics (Y. Demiris & Johnson, 2003; Y. Demiris & Khadhouri, 2005; Haruno, Wolpert, & Kawato, 2001). In (J. Demiris & Hayes, 2002) a set of forward-inverse models coupled with a selection mechanism is employed to control a dynamical simulation of a humanoid robot which learns to perform and imitate gestures. A similar architecture is employed on a planar robot manipulator (Simmons & Demiris, 2004) and on a mobile robot with a gripper (Johnson & Demiris, 2005).

Although with some differences, in all these examples the forward-inverse model pairs are activated both when the robot is performing a motor task and when the robot is attending the demonstrator. During the demonstration of an action, the perceptual input from the scene is fed in parallel to all the inverse models. In turn, the output of the inverse models is sent to the forward models which act as simulators (at the same time the output of the inverse models is inhibited, so no action is actually executed by the robot). The idea is that the predictions of all the forward models are compared to the next state of the demonstrator and the resulting error taken as a confidence measure to select the appropriate action for imitation. In these respects, these various implementations are very similar to the one presented in this paper. Each inverse-forward model pair gives a hypothesis to recognize the observed action. The confidence measure that is accumulated during action observation can be interpreted as an estimation of the probability that that hypothesis is true, given the perceptual input. More recently the same architecture has been extended to cope with the obvious differences in human versus robot morphology, and applied to an object manipulation task (Johnson & Demiris, 2005). In the architecture, in this case, the set of coupled internal inverse and forward models are organized hierarchically, from the higher level, concerned with actions and goals in abstract terms, to the lower level, concerned with the motoric consequences of each action.

The MOSAIC model by Haruno, Wolpert and Kawato (Haruno, Wolpert, & Kawato, 2001) is also close to our model in the sense of explicitly considering the likelihood and priors in deciding which of the many forward-inverse pairs to activate for controlling the robot. In this case no explicit imitation schema is considered, but the context is properly accounted for by the selection mechanism (a Hidden Markov Model).

Our proposal of the functioning of the mirror system includes aspects of both solutions by considering the forward-inverse mapping and the contribution of the contextual information. In our model, and in contrast with Demiris' architecture, the goal of the action is explicitly taken into account. As we discussed in Section 2 the mirror system is activated only when a goal-directed action is generated or observed. Recognizing the goal is also the key aspect of the learning process and subsequently it works as a prior to bias recognition by filtering out actions that are not applicable or simply less likely to be executed, given the current context. This is related to what happens in the (Johnson & Demiris, 2005) paper where the architecture filters out the actions that cannot be executed. In our model, however, the affordances of a given object are specified in terms of the probabilities of all actions that can be performed on the object (estimated from experience). These probabilities do not only filter out impossible actions, but bias the recognition toward the action that is more plausible given the particular object. The actions considered in this paper are all transitive (i.e. directed towards objects), coherently with the neurophysiological view on the mirror system. In the same paper (Johnson & Demiris, 2005), moreover, the recognition of the action is performed by comparing the output of the forward models to the state of the demonstrator. In our model this is always performed in the motor space. In the experiments reported in this paper we show that this simplifies the classification problem and leads to better generalization.

The outcome from a first set of experiments using the data set collected with the cyber glove setup has shown that there are at least two advantages whether the action classification is performed in visual rather than motor space: i) simpler classifier, since the classification or clustering is much simpler in motor space, and ii) better generalization, since motor information is invariant to changes of the point of view. Some of these aspects are further discussed in (Lopes & Santos-Victor, 2005).

[Figure 14 about here]

The robotic experiment shows, on the other hand, that indeed only minimal initial skills are required in learning a mirror neuron representation. In practice, we only had to assume reaching to guarantee interaction with objects and a method to visually measure the results of this interaction. Surely, this is a gross simplification in many respects since, for example, aspects of the development of grasping per se were not considered at this stage and aspects of agency were neglected (the robot was not measuring the posture and behavior of the human manipulator). Though, this shows that, in principle, the acquisition of the mirror neuron structure is the almost natural outcome of the development of a control system for grasping. Also, we have put forward a plausible sequence of learning phases involving the interaction between canonical and mirror neurons. This, we believe, is well in accordance with the evidence gathered by neurophysiology. In conclusion, we have embarked in an investigation that is somewhat similar to the already cited Liberman's speech recognition attempts. Perhaps, also this time, the mutual rapprochement of neural and engineering sciences might lead to a better understanding of brain functions.

Acknowledgments

The research described in this paper is supported by the EU project MIRROR (IST-2000-28159) and ROBOTCUB (IST-2004-004370). The authors wish to thank Claes von Hofsten, Kerstin Rosander, José Santos-Victor, Manuel Cabido-Lopes, Alexandre Bernardino, Matteo Schenatti, and Paul Fitzpatrick for the stimulating discussion on the topic of this paper. The authors wish also to thank the anonymous reviewers for the comments on the early version of this manuscript.

Notes

1. (Liberman & Mattingly, 1985)

2. The module of language perception.

3. In supervised learning, training information is previously labeled by an external teacher or a set of correct examples are available.

References

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., Breazeal, C. L., Marjanović, M., & Scassellati, B. (1999). The COG project: Building a Humanoid Robot. In C. L. Nehaniv (Ed.), *Computation for Metaphor, Analogy and Agents* (Lecture Notes in Artificial Intelligence, Vol. 1562, pp. 52–87): Springer-Verlag.

Demiris, J., & Hayes, G. (2002). Imitation as a Dual-Route Process Featuring Predictive and Learning Components: A Biologically-Plausible Computational Model. In K. Dautenhahn & C. L. Nehaniv (Eds.), *Imitation in Animals and Artifacts* (pp. 327–361). Cambridge, MA, USA: MIT Press.

Demiris, Y., & Johnson, M. H. (2003). Distributed, predictive perception of actions: a biologically inspired robotics architecture for imitation and learning. *Connection Science, 15*(4), 231–243.

Demiris, Y., & Khadhouri, B. (2005). Hierarchical Attentive Multiple Models for Execution and Recognition (HAMMER). *Robotics and Autonomous Systems Journal*(In press).

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. *In proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2161–2166 Vol. 3, 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.

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.

Haruno, M., Wolpert, D. M., & Kawato, M. (2001). MOSAIC Model for sensorimotor learning and control. *Neural Computation, 13*, 2201–2220.

Jeannerod, M. (1988). *The Neural and Behavioural Organization of Goal-Directed Movements* (Vol. 15). Oxford: Clarendon Press.

Jeannerod, M., Arbib, M. A., Rizzolatti, G., & Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor transformation. *Trends in Neurosciences, 18*(7), 314–320.

Johnson, M., & Demiris, Y. (2005). Hierarchies of Coupled Inverse and Forward Models for Abstraction in Robot Action Planning, Recognition and Imitation. *In proceedings of the AISB 2005 Symposium on Imitation in Animals and Artifacts,* 69–76, Hertfordshire, UK.

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.

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., & Mattingly, I. G. (1985, page 7). 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.

Lopes, M., & Santos-Victor, J. (2005). Visual learning by imitation with motor representations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B Cybernetics, 35*(3), 438–449.

Luppino, G., & Rizzolatti, G. (2000). The organization of the frontal motor cortex. *News in Physiological Sciences, 15*, 219–224.

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

Miall, R. C., Weir, D. J., Wolpert, D. M., & Stein, J. F. (1993). Is the cerebellum a Smith predictor? *Journal of Motor Behavior, 25*(3), 203–216.

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.

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

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

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.

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.

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.

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

Shimazu, H., Maier, M. A., Cerri, G., Kirkwood, P. A., & Lemon, R. N. (2004). Macaque ventral premotor cortex exerts powerful facilitation of motor cortex outputs to limb motoneurons. *The Journal of Neuroscience, 24*(5), 1200–1211.

Simmons, G., & Demiris, Y. (2004). Imitation of Human Demonstration Using A Biologically Inspired Modular Optimal Control Scheme. *In proceedings of the IEEE-RAS/RSJ International Conference on Humanoid Robots*, 215–234, Los Angeles, USA.

Thelen, E., & Smith, L. B. (1998). *A Dynamic System Approach to the Development of Cognition and Action* (3rd ed.). Cambridge, MA: MIT Press.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701–702.

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.

Wolpert, D. M., & Miall, R. C. (1996). Forward models for physiological motor control. *Neural Networks, 9*(8), 1265–1279.

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

Authors' addresses

Giorgio Metta
LIRA-Lab, University of Genoa
Viale Causa, 13 — 16145 — Italy
Email: pasa@liralab.it

Giulio Sandini
LIRA-Lab, University of Genoa
Viale Causa, 13 — 16145 — Italy
Email: sandini@unige.it

Lorenzo Natale
LIRA-Lab, University of Genoa
Viale Causa, 13 — 16145 — Italy
Email: nat@liralab.it

Laila Craighero
Faculty of Medicine — D.S.B.T.A. Section of Human Physiology
Via Fossato di Mortara 17/19 — 44100 Ferrara — Italy
Email: crh@unife.it

Luciano Fadiga
Faculty of Medicine — D.S.B.T.A. Section of Human Physiology
Via Fossato di Mortara 17/19 — 44100 Ferrara — Italy
Email: fdl@unife.it

About the authors

Giorgio Metta: Assistant Professor at the University of Genoa where he teaches the courses of "Anthropomorphic Robotics and Operating Systems". Ph.D. in Electronic Engineering in 2000. Postdoctoral associate at MIT, AI-Lab from 2001 to 2002. Since 1993 at LIRA-Lab where he developed various robotic platforms with the aim of implementing bioinspired models of sensorimotor control. For more information: http://pasa.liralab.it and http://www.robotcub.org.

Giulio Sandini: Full Professor of bioengineering at the University of Genova where he teaches the course of "Natural and Artificial Intelligent Systems". Degree in Bioengineering at the University of Genova; Assistant Professor at the Scuola Normale Superiore in Pisa studying aspects of visual processing in cortical neurons and visual perception in human adults and children. Visiting Research Associate at Harvard Medical School in Boston working on electrical brain activity mapping. In 1990 he founded the LIRA-Lab (www.liralab.it) investigating the field of Computational and Cognitive Neuroscience and Robotics with the objective of understanding the neural mechanisms of human sensorimotor coordination and cognitive development by realizing anthropomorphic artificial systems such as humanoids.

Lorenzo Natale was born in Genoa, Italy, in 1975. He received the M.S. degree in Electronic Engineering in 2000 and Ph.D. in Robotics from the University of Genoa, Italy, in 2004. He is currently Postdoctoral Associate at MIT, Computer Science and Artificial Intelligence Laboratory. His research involves the study of sensorimotor development and motor control in artificial and natural systems.

Laila Craighero: Born in 1968, degree in Psychology, Ph.D. in Neuroscience. Assistant Professor of Human Physiology at the University of Ferrara. Scientific interests: neural mechanisms underlying orienting of visuospatial attention.
Experience in transcranial magnetic stimulation and electrophysiological investigation in monkeys (single neurons recordings). Among her contributions, in the frame of premotor theory of attention, it is the investigation of the role played by the motor preparation on attentional orienting. More recently, she collaborated to the study of motor cortex and spinal cord excitability during observation of actions made by other individuals ('mirror' and 'canonical' neurons). Laila Craighero is co-investigator in E.C. funded projects on action understanding and sensorimotor coordination.

Luciano Fadiga: Born in 1961. M.D., Ph.D. in Neuroscience, Full Professor of Human Physiology at the University of Ferrara. He has a consolidated experience in electrophysiological investigation in monkeys (single neurons recordings) and humans (transcranial magnetic stimulation, study of spinal excitability and brain imaging). Among his contributions is the discovery of monkey mirror neurons in collaboration with the team leaded by G. Rizzolatti at the University of Parma and the first demonstration by TMS that a mirror system exists also in humans. Other fields of his research concern visual attention and speech processing. He is responsible for a research unit funded by the European Commission for the study of the brain mechanisms at the basis of action understanding.

Appendix I

The role of the mirror system during action recognition can be framed in a Bayesian context (Lopes & Santos-Victor, 2005). In the Bayesian formulation the posterior probabilities can be written as:

$$Posterior = \frac{Likelihood \times Prior}{Normalization} \tag{1}$$

and the classification is performed by looking at the maximum of the posterior probabilities with respect to the set of available actions. The normalization factor can be neglected since it does not influence the determination of the maximum. We equated the prior probabilities with the object affordances and estimated them by counting the occurrences of the grasp types vs. the object type. In this interpretation, in fact, the affordances of an object identify the set of actions that are most likely to be executed. The likelihood was approximated by a mixture of Gaussians whose number and parameters were estimated with the Expectation Maximization (EM) algorithm.
More specifically the mirror activation of F5 can be thought as:

$$p(A_i \mid F, O_k) \tag{2}$$

where A_i is the i^{th} action from a motor repertoire of I actions, F are the features determining the activation of this particular action and O_k is the target object of the grasping action out of a set of K possible objects. This probability can be computed from Bayes rule as:

$$p(A_i \mid F, O_k) = p(F \mid A_i, O_k)\, p(A_i \mid O_k) \,/\, p(F \mid O_k) \tag{3}$$

and a classifier can be constructed by taking the maximum over the possible actions as follows:

$$\hat{A} = \max_i p(A_i \mid F, O_k) \tag{4}$$

The term $p(F \mid O_k)$ can be discarded since it does not influence the maximization. For the other terms we can give the following interpretation:

$p(A_i \mid F, O_k)$	Mirror neurons responses, obtained by a combination of the information as in equation 1.
$p(F \mid A_i, O_k)$	The activity of the F5 motor neurons generating certain motor patterns given the selected action and the target object
$p(A_i \mid O_k)$	Canonical neurons responses, that is the probability of invoking a certain action given an object

The feature vector F can also be considered over a certain time period and thus we should apply the following substitution to equations 1 to 3:

$$F = F_t, \ldots, F_{t-N} \tag{5}$$

which takes into account the temporal evolution of the action.

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 modelling of the probability density functions. We have:

$$p(A_i \mid F_i, \ldots F_{t-N}, O_k) = \prod_{j=0}^{N} \frac{p(F_{t-j} \mid A_i, O_k)\, p(A_i \mid O_k)}{p(F_{t-j} \mid O_k)} \tag{6}$$

This clearly does not have to be close to the actual brain responses but it was considered since it simplifies computation if compared to the full joint probabilities.