Understanding mirror neurons: a bio-robotic approach

[1,*]Giorgio Metta, [1]Giulio Sandini, [1]Lorenzo Natale, [2]Laila Craighero, [2]Luciano Fadiga

[1]LIRA-Lab, DIST, University of Genoa

[2]Neurolab, DBS, University of Ferrara

*Contact author: Giorgio Metta, LIRA-Lab, Viale Causa, 13 – 16145 Genoa Italy

E-mail: pasa@liralab.it

Abstract

This paper reports about our investigation on action understanding in the brain. We are taking a three-pronged approach based on some recent results of the neurophysiology, on the modeling from recording of human movement, and on the implementation of the model on a robotic setup interacting in a natural environment.

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 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 are 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 word, as uttered by different subjects, possessed any common phonetic invariant. 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 are not sounds but rather articulatory gestures that have evolved exclusively at the service of language. Accordingly, a cognitive translation into phonology is not necessary because the articulatory gestures are phonologic in nature. This elegant idea was however strongly debated at the time mostly because it was difficult to test, verification through the implementation on a computer system was impossible, and in fact only recently the theory has 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 possibe alternative is for Liberman to be right and that speech perception and speech production use a common repertoire 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".

Perhaps it is the case that if the acoustic modality were replaced, for example, by vision this principle would still hold. In both cases, the brain requires a "resonant" system that matches the observed/listened actions onto 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 incipit of the action and the implicit knowledge of its evolution. Recent neurophysiological experiments show that such a motor resonant system indeed exists in the monkey's brain. 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.

In the next section we will describe the basic properties of this area. Then we will propose a biologically plausible model on how action recognition may be achieved by motor-resonant mechanisms similar to those observed in the monkey. Finally, somewhat inspired and fascinated by Liberman's idea, we started implementing part of the model in two different experimental platforms with the goal of validating the model and of understanding the variables at play in action recognition in the brain.

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.

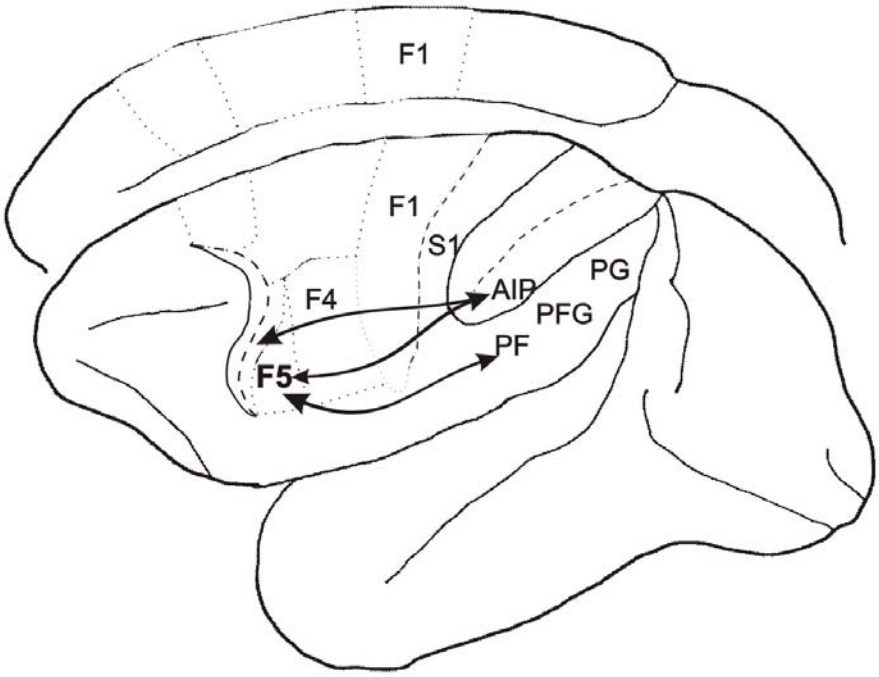

Figure 1: Lateral view of monkey right hemisphere. Area F5 is buried inside the *arcuate sulcus* (posterior bank) and emerges on the convexity immediately posterior to it. Area F5 is bidirectionally connected with the inferior parietal lobule (areas AIP, anterior intraparietal, PF and PFG). Areas F5 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.

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).

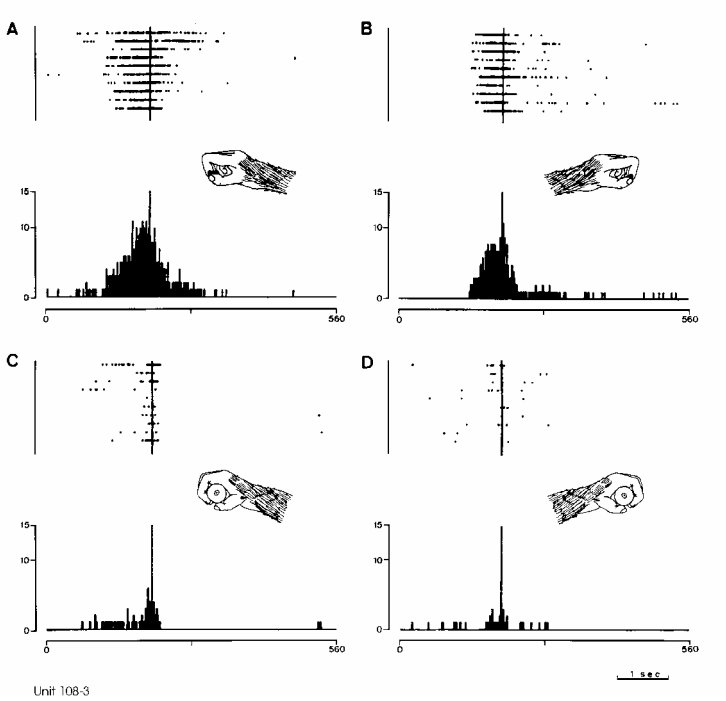

Figure 2: F5 grasping neurons. In the uppermost part of each panel eight successive trials are represented. Each dot represents an action potential. In the lowermost part the sum histogram is drawn. Trials are aligned with the moment at which the monkey touches the object (vertical lines across histograms).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

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).

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.

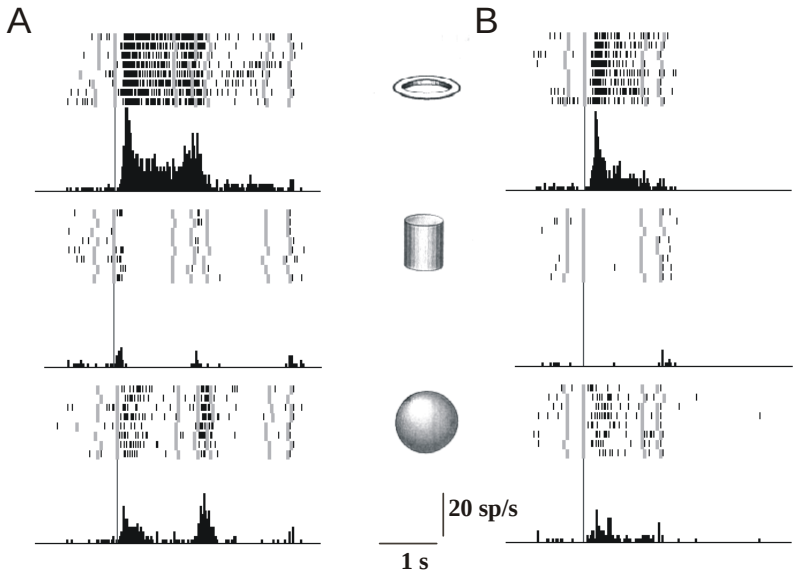

Figure 3: Responses of a visuomotor “canonical” neuron of area F5. Each panel shows the neuron activity recorded during the observation and grasping (A) or the mere observation (B) of three different three-dimensional objects. The alignment of the single trials coincides with the moment in which the 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).

2.4 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 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.

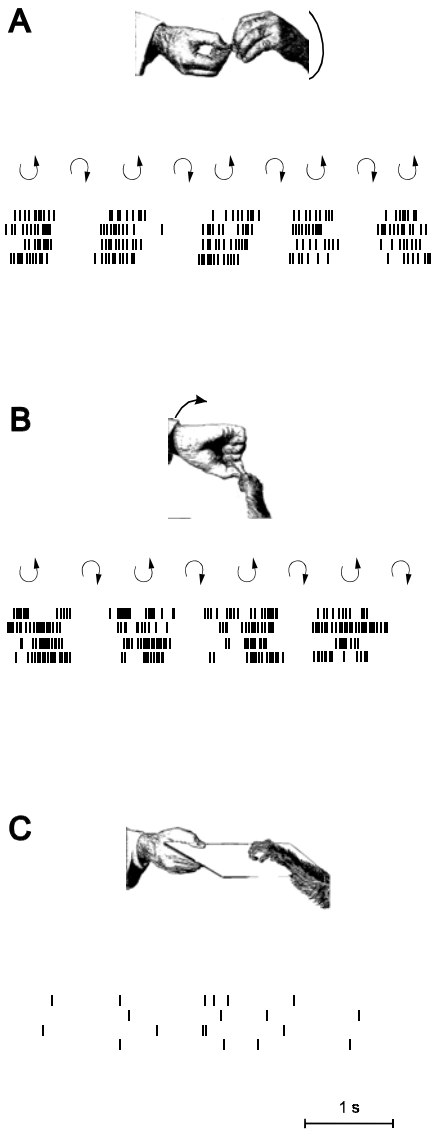

Figure 4: Highly congruent F5 visuomotor "mirror" neuron. Behavioral situations are schematically represented in the upper part of each panel above a series of consecutive rasters and relative response histograms. A, the monkey observes the experimenter who rotates his hands around a raisin alternating 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).

3 A model of area F5 and 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 are to be found in the 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 part of the activation of F5. For explanatory purpose, the description of our model of the mirror system can further 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 5): in her/his perspective, 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 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).

In fact, we can surely start asking what two individuals have in common: what do they share when mutually interacting? What information can be shared in interacting with objects? We claim 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 filtering out irrelevant information. A similar role with respect to the specification of the hand posture would be appropriate to 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 so 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).

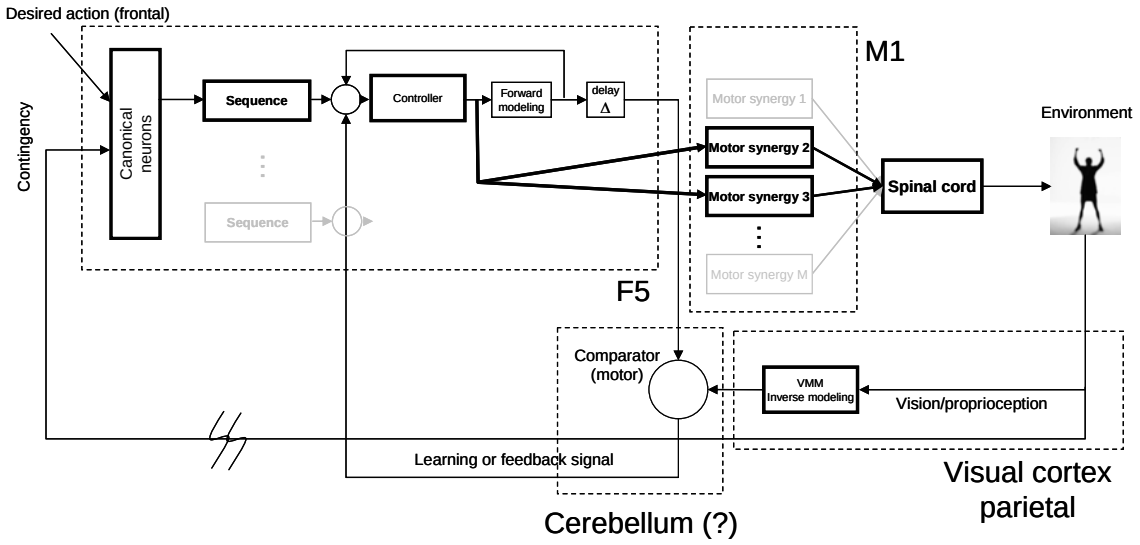

Figure 5: Model schematics of a forward-inverse model of F5 and mirror neurons. The response of the mirror system is seen as an adaptation of a feedback loop controlling the execution of grasping. The model only contains details of some brain areas while it is known that many others participate to the 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 in 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 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, 2003).

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, 2003).

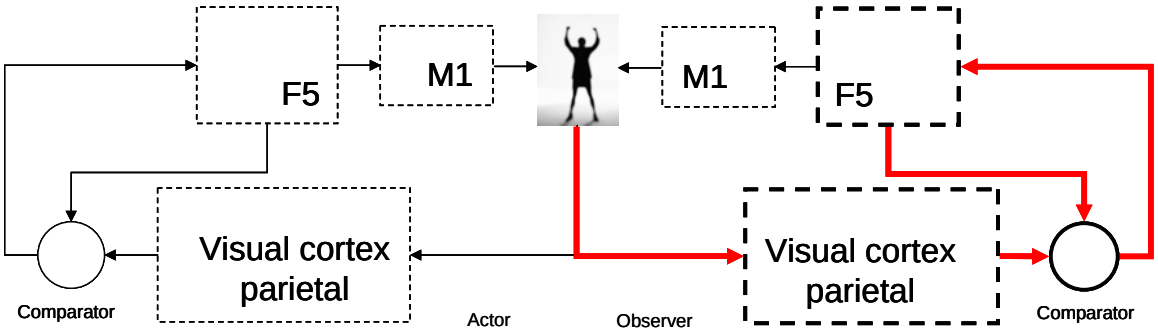

Figure 6: Action and observation activity. The model of Figure 5 is replicated to describe the observer's brain. The observation of a certain action activates the same feedback path used in executing that action (thick solid lines). Many details as shown in Figure 5 were dropped here for clarity of presentation.

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 where 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.

Experiments probing different aspects of the model were conducted on two different setups: i) the grasping data acquisition setup described next, and ii) a humanoid robot described later in section 5. The reason for experimenting on two 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.

4 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 to access 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 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 is our intendment to use motor information as an aggregating principle to determine which visual features are 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 section.

Figure 7: 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.

At this stage we have collected the data set for further off-line processing. 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 8: on the left panel a precision grip using all fingers; on the right one a two-finger precision grip.

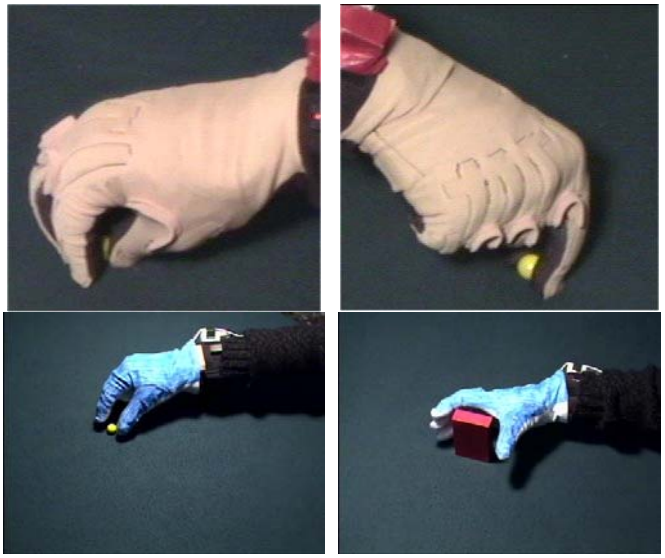

Figure 8: Exemplar grasp types as recorded during data collection. The topmost row shows two types of grasp applied to the same object (the small glass ball) from different points of view. The bottom images show two very different grasp types from the same point of view.

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 is the data set that is presently being used for building a Bayesian classifier in motor space (Cabido Lopes & Santos-Victor, 2003). An early analysis yielded promising results, especially in terms of generalization, and results that are, at the moment, subject of further investigation.

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, Brezeal, Marjanovic, & Scassellati, 1999). This is an upper-torso human shaped robot with 22 degrees of freedom distributed along the head, arms and torso. It lacks hands, it has instead simple flippers that could use to push and prod objects. It can't 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 (QNX) 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.

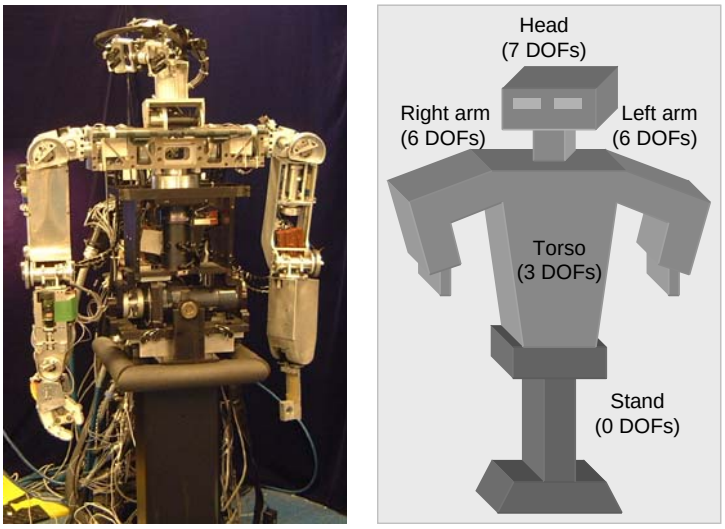

Figure 9: Experimental robotic setup. Cog, the humanoid robot platform developed at MIT, AI-Lab; for further details see (Brooks et al., 1999).

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 or evidence any lacuna in our model. 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 1 shows four level 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* (Lungarella, Metta, Pfeifer, & Sandini, 2003; Thelen & Smith, 1998); the identification of “developmental levels” here simplifies though our comprehension of the mechanisms of learning and development.

Table 1: degrees of causal indirection, brain areas and function in the brain.

Level	Nature of causation	Brain areas	Function and behavior	Time profile
1	Direct causal chain	VC-VIP/7b-F4-F1	Reaching	Strict synchrony
2	One level of indirection	VC-AIP-F5-F1	Poking, prodding, grasping	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

The first level in Table 1 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 contemporary 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 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 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 respect 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 exposition 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 any longer about step 4 which is relatively sideways with respect to this paper. The robot also possesses, and we are not going to enter much into the details here, some basic attentional 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 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 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 10. 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.

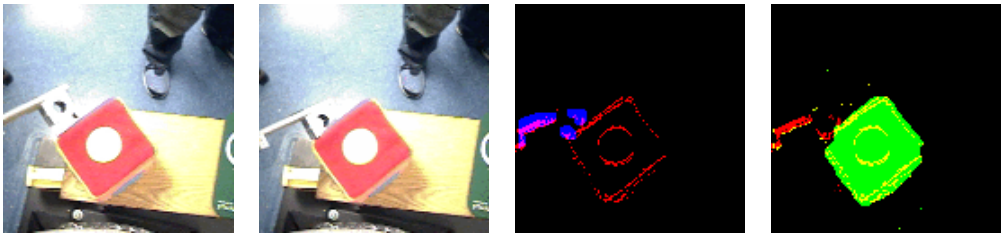

Figure 10: Example of segmentation obtained by reaching and poking an object sitting on a table in front of the robot; adapted from (Fitzpatrick & Metta, 2003). The first two pictures show the moment of impact with the object, the third picture is a color-coded version of the motion detection filter 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 11 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 11 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 12 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 11 and Figure 12) 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).

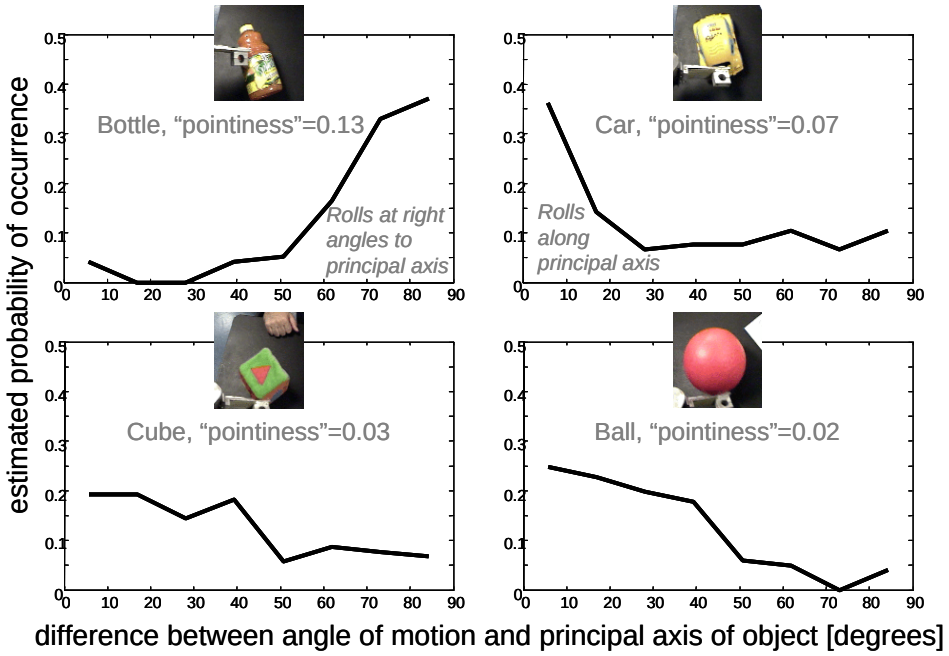

Figure 11: Probability of observing a roll along a particular direction with respect to the object principal axis. Abscissas are in degrees. Note as the toy car and the bottle show a very specific behavior: they possess a preferred rolling direction with respect to their principal axis.

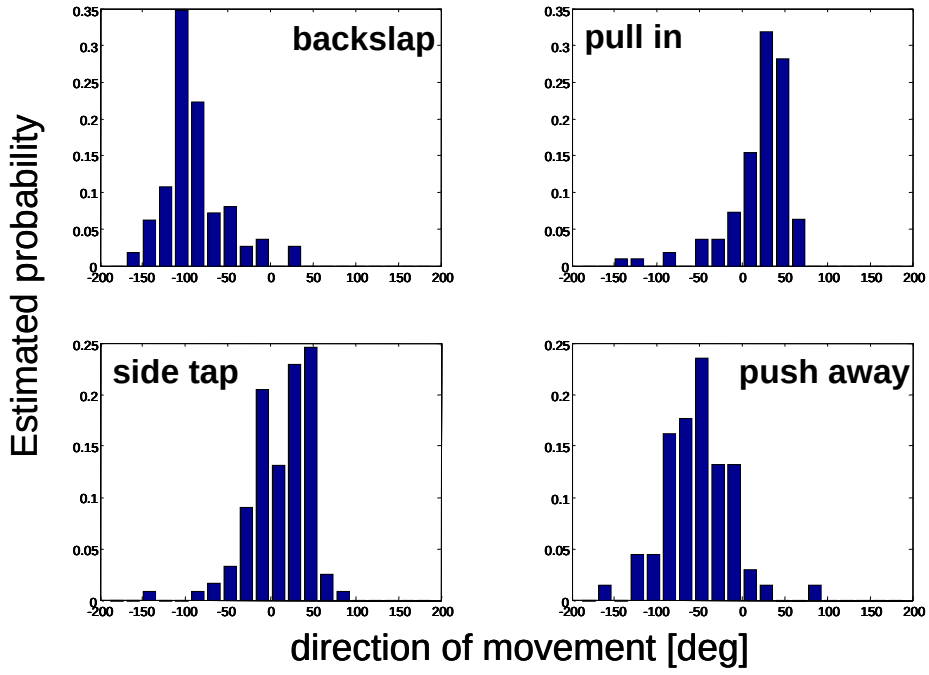

Figure 12: Histogram of the direction of movement of the object for each possible action. Abscissas are in degrees. This set of plots show that each action would generate on average a typical consequence on the approached object. The direction of motion expressed in angles is referred from the robot’s 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 3 (Figure 5).

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. 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 13 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.

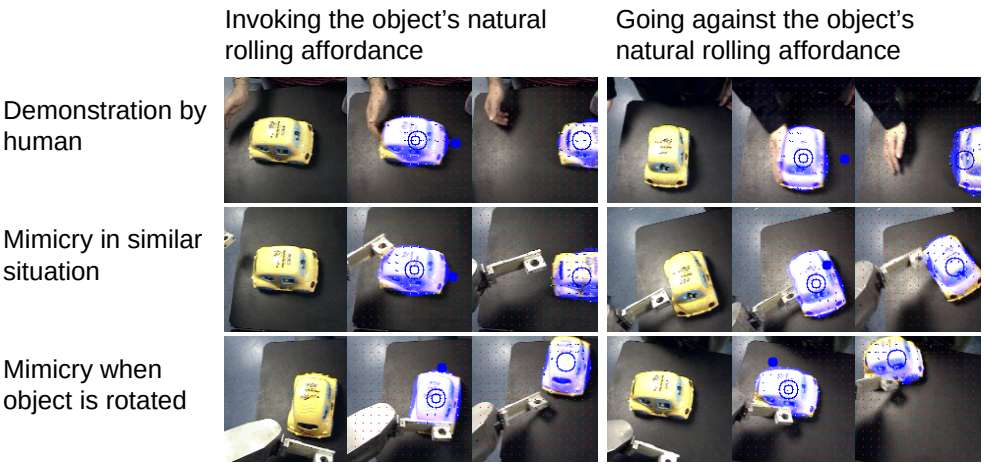

Figure 13: An extended imitation experiment. Here two different type of experiments are shown: acting according to an affordant behavior (left), or against the principal affordance (right). Further, for each situation we show the demonstrated action, the mimicry obtained when the object is oriented similarly to the demonstration, and when the object is oriented differently from the demonstration.

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 into 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.

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 effects 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 discussed in (Cabido Lopes & Santos-Victor, 2003).

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. 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.

9 Acknowledgment

The research described in this paper is supported by the EU project MIRROR (IST-2000-28159). The authors wish to thank Claes von Hofsten, Kerstin Rosander, Jose' Santos-Victor, Manuel Cabido Lopes, Alexandre Bernardino, Matteo Schenatti, and Paul Fitzpatrick for the stimulating discussion on the topic of this paper.

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

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.

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.

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

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

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental Robotics: A Survey. Connection Science. *Connection Science, 15*(4), 151-190.

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

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.

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.

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.

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

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

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

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.