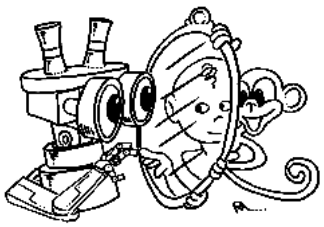


MIRROR
IST–2000-28159
Mirror Neurons based Object Recognition

Delivery Date: November, 2002
Classification: Internal
Responsible Person: Dr. Giorgio Metta, Prof. Giulio Sandini, Prof. Luciano Fadiga
Partners Contributed: ALL

Modeling the development of mirror neurons: initial considerations and future plans

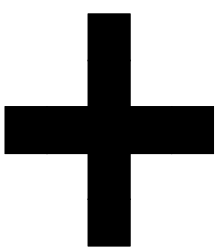
Annex 1
to
Deliverable Item 1.4
Periodic Progress Report N°: 1

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

Introduction

Vision and manipulation are inextricably intertwined in the primate brain. Neuroscientists are doing a very good job in elucidating the mixed structure of action and perception. We now know a great deal about this structure. By providing a plausible model of these same functions we can delve deeper into the whys: i.e. is this integration functionally important? If the answer is yes, how much is it important? A physical implementation, in the form of a robotic system, can shed new light into the linkage between acting and perceiving.

We argue that tracing chains of causality or cause-effect relations from the actor's own body to the environment leads to a natural developmental progression of visual and motor competences. Causality is intended as a descriptive tool and it is used to interpret aspects of the development of prospective control and learning. Eventually this procedure might lead to the developmental description of *mirror neurons*. The ability to form and interpret longer chains of causally-related events is seen as triggering the emergence of new functionality and/or a new set of behaviors.



a cross

0	0	1	0	0
0	0	1	0	0
1	1	1	1	1
0	0	1	0	0
0	0	1	0	0

a binary cross

4	12	17	4	15
9	21	3	10	25
5	23	11	37	13
8	18	7	42	6
27	46	31	32	50

?

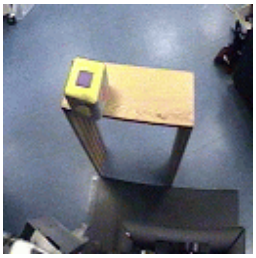


Figure 1 On the left are three examples of crosses. The human ability to segment objects is not general-purpose, and improves with experience. On the right is an image of a cube on a table, illustrating the ambiguities that plague machine vision. The edges of the table and cube happen to be aligned, the colors of the cube and table are not well separated, and the cube has a potentially confusing surface pattern.

Object or illusion?

[1]Following (Manzotti & Tagliasco, 2001), we can ask whether macroscopic objects exist completely in their own right, or instead owe something of their existence to their interaction with an observer. How the world is divided up, and what parts of it we grant status as objects, says as much about us as about the world around us (Hendriks-Jansen, 1996). For example, would a chair still be a chair if we had a completely different embodiment? Further, even if a part of the physical world could be separated out from the background in an objective manner, its function still depends on our body and skills – for example, a floppy disk is of little use to one who is computer illiterate, and perhaps can be just regarded as a clumsy frisbee or ugly drink coaster.

[1] A longer and somewhat similar discussion is presented in G. Metta and P. Fitzpatrick's *Early Integration of Vision and Manipulation* submitted to Adaptive Behavior.

Consider the example in Figure 1. It is clear that the cross on the left is a cross and does not seem to owe its existence to us as observers. The array in the middle for many of us is still a cross. This would still be the case even if we had not developed the concept of number or these particular graphic symbols to identify numbers. What can we say about the array on the right? On a first examination it looks like a random collection of numbers. But if we are told that the criterion is "prime numbers vs. non-prime" then a cross can still be identified.

On the very right of Figure 1 we show a cube sitting on a table. While humans are very good in analyzing scenes like this one, there are many features that can fool a computer vision system. The edges of the cube and table happen to be aligned, the color is poorly separated, and the surface pattern of the cube does not really tell much about the object itself. Is the internal dark square a different object lying on top of the cube? Another possibility is that the cube is extremely heavy or even part of the table and thus it is not manipulable or movable. Does it make sense then to speak about objects in images, as if there were a unique correspondence between the two? As early as 1734, Berkeley observed that:

> *...objects can only be known by touch. Vision is subject to illusions, which arise from the distance-size problem... (Berkeley, 1972)*

Vision is indeed subject to many illusions. But touch also can be fooled since it has been shown that vision and touch combine optimally with respect to a maximum likelihood criterion (Ernst & Banks, 2002). Which sensory modality dominates depends on the experimental conditions and apparently we shouldn't always "blindly" trust our senses. The key to resolving ambiguity is to take action, rather than remain a passive observer.

A brief survey

The example of the cross composed of prime numbers is a novel (albeit unlikely) type of segmentation in our experience as adult humans. We might imagine that when we were very young, we had to initially form a set of such criteria to solve the object identification/segmentation problem in more mundane circumstances. That such abilities develop and are not completely innate is suggested by many investigators. For example Kovacs (Kovacs, 2000) has shown that perceptual grouping is slow to develop and continues to improve well beyond early childhood (14 years). Long-range contour integration was tested and this work elucidated how this ability develops to enable extended spatial grouping.

A useful concept to understand how such capabilities could develop is the well-known theory of Ungerleider and Mishkin (Ungerleider & Mishkin, 1982) who first formulated the hypothesis that objects are represented differently during action than they are for a purely perceptual task. Briefly, they argue that the brain's visual pathways split into two main streams: the dorsal and the ventral (Milner & Goodale, 1995). The dorsal deals with the information required for action, while the ventral is important for more cognitive tasks such as maintaining an object's identity and constancy. Although the dorsal/ventral segregation was emphasized by many commentators, it is significant that there is actually a great deal of cross talk between the streams. Observation of agnosic patients (Jeannerod, 1997) shows a much more

complicated relationship than the simple dorsal/ventral dichotomy would suggest. For example, although some patients could not grasp generic objects (e.g. cylinders), they could correctly preshape the hand to grasp known objects (e.g. a lipstick): interpreted in terms of the two pathways, this implies that the ventral representation of the object can supply the dorsal stream with size information.

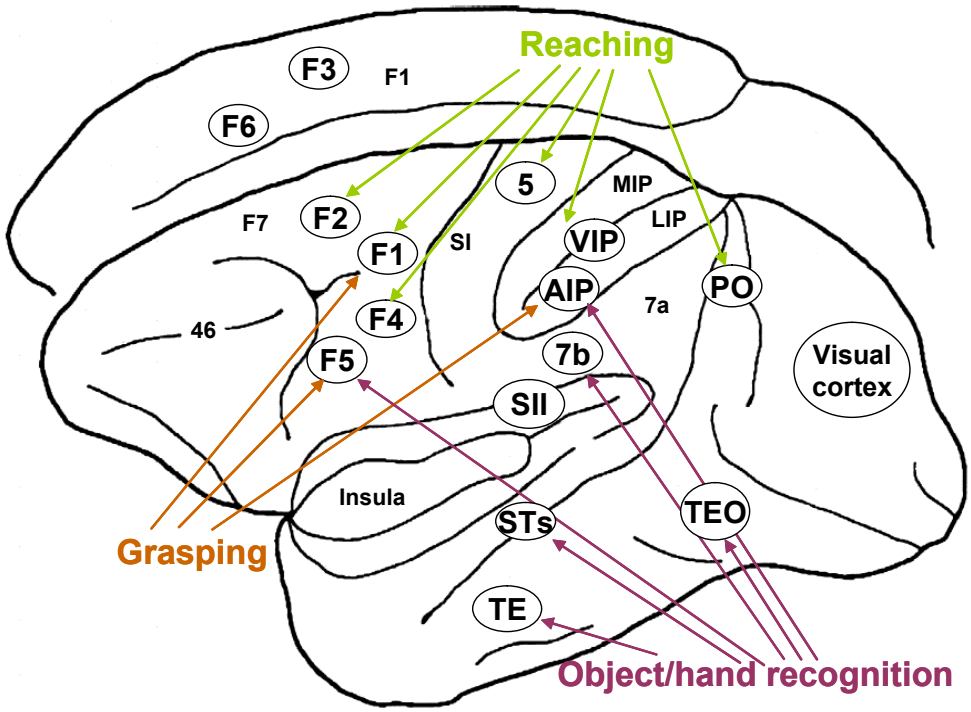


Figure 2 Monkey brain with indication of the main areas participating in object oriented actions (adapted from (Fagg & Arbib, 1998)). As described in the text, three main functions can be identified: object recognition, reaching, and grasping. These form three parallel yet connected streams of processing. The circuit connecting the visual cortex to the inferior parietal lobule VIP, F4 and F1 is thought to compute the visuomotor transformations required to control reaching. Some evidence also suggests a possible role in the organization of reaching played by the posterior parietal cortex PO and dorsal premotor area F2, reciprocally connected. AIP and F5 are responsible for grasping. Temporal areas (TE, TEO) and STs are correlated to the semantic of object recognition.

Grossly simplifying, the brain circuitry responsible for object oriented actions is thought to consist of at least four interacting regions (Figure 2), namely the primary motor cortex (F1), the premotor cortex (F2, F4, F5), the inferior parietal lobule (AIP, VIP), and the temporal cortex (TE, TEO) (see (Fadiga, Fogassi, Gallese, & Rizzolatti, 2000; Jeannerod, 1997; Rizzolatti, Fogassi, & Gallese, 1997) for a review). While this is a useful subdivision, it is worth bearing in mind that the connectivity of the brain is much more complex, that bidirectional connections are present, and that behavior is the result of a population activity of these areas. The example about the grasping of known objects in agnosic patients testifies the *abundance of anatomical connections* between different regions (Jeannerod, Arbib, Rizzolatti, & Sakata, 1995).

Another way of looking at the same connectivity is in terms of the main function of each area. For example F4, VIP, and 7b are involved in the control of reaching; F5 and AIP contain the majority of grasp related neurons, while TE and TEO are thought to subserve object recognition. These regions together form a network of parallel and yet interacting processes. In fact, at the behavioral level, it has been observed that reaching and grasping need to interact to correctly orient and preshape the hand (Jeannerod et al., 1995).

Neurons responsive to reaching are present in the inferior parietal lobule. For example, Jeannerod et al. reported that the temporary inactivation of the caudal part (VIP) of the intraparietal sulcus by injecting a GABA agonist disrupts reaching. Conversely, injection in the more rostral part (area AIP) interferes with the preshaping of the hand. Some of the VIP neurons have bimodal visual and somatic receptive fields (RF). About 30% of them have a RF which does not vary with movement of the head (Rizzolatti et al., 1997). The tactile and visual RFs often overlap (e.g. a central visual RF corresponds to a tactile RF in the nose or mouth). The parietal cortex also contains cells related to eye position/movements that appear to be involved in the visuo-motor transformation required for reaching. VIP projects to area F4 in the premotor cortex. Area F4 contains neurons that respond to objects and are related to the description of the peripersonal space with respect to reaching (Fogassi et al., 1996; Graziano, Hu, & Gross, 1997b). A subset of the F4 neurons has a somatosensory, visual, and motor receptive field. The visual receptive field extends in 3D from a given body part, such as the forearm. The somatosensory RF is usually in register with the visual one (as in VIP neurons). Motor information is integrated into the representation by maintaining the receptive field anchored to the correspondent body part (the forearm in this example) irrespective of the relative position of the head and arm.

Also, Graziano et al. (Graziano, Hu, & Gross, 1997a) described neurons that maintain a memory of the position of objects for the purpose of reaching. They found neurons that change their firing rate after an object is illuminated briefly within reaching distance. The neurons return to their baseline firing rate only after the monkey is shown that the object has been taken away or moved to a different position.

Sakata and coworkers (Sakata, Taira, Kusunoki, Murata, & Tanaka, 1997) investigated the response of neurons in the parietal cortex and in particular in area AIP (anterior intra-parietal). They found cells responsive to complex visual stimuli. Neurons in AIP responded during grasping/manipulative actions and when an object was presented to the monkey but no reaching was allowed. Neurons were classified as motor dominant, visual dominant or visuo-motor type depending on how they fired in the dark. Of the visual dominant neurons, some responded to the presentation of the object alone and often they were very specific to the size and orientation of the object, others to the type of object, while yet others responded indifferently to the presentation of a broad class of objects. Area AIP is interesting because it contains both motor and visually responsive cells intermixed in various proportions; it can be thought of as a visuo-motor vocabulary for controlling object directed actions. It is also interesting because projections from AIP terminate in the agranular frontal cortex. For many years, because of the paucity of data, this part of the cortex was considered a unitary motor control area. Recent studies (see (Fadiga et al., 2000; Jeannerod, 1997)) have demonstrated that this is not the case. Particularly surprising was the discovery of visual responsive neurons. A good proportion of them have both visual/sensory and motor responses. Area F5, one of the main targets of the projection

from AIP (to which it sends back recurrent connections), was thoroughly investigated by Rizzolatti and colleagues (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996).

F5 neurons can be subdivided in two groups: the purely motor neurons (80%) and those with visuomotor responses (20%). The visually responsive neurons can be then classified into canonical and mirror. Canonical and mirror neurons are indistinguishable from each other on the basis of their motor responses; their visual responses however are quite different. The canonical type is active in two situations: i) when grasping an object and ii) when fixating that same object. For example, a neuron active when grasping a ring also fires when the monkey simply looks at the ring. This could be thought of as a neural analogue of the "affordances" of Gibson (Gibson, 1977). However, given the heavy projection from AIP, it is not entirely true that the affordances are fully described/computed by F5 alone. A more conservative stance is that the system of AIP, F5, and other areas (such as TE) participate in the visual processing and motor matching required to compute the affordances of a given object.

The second type of neuron identified in F5, the mirror neuron, becomes active under either of two conditions: i) when manipulating an object (e.g. grasping it, as for canonical neurons), and ii) when watching someone else performing the same action on the same object. This is a more subtle representation of objects, which allows and supports, at least in theory, mimicry behaviors. In humans, area F5 is thought to correspond to Broca's area; there is an intriguing link between gesture understanding, language, imitation, and mirror neurons (Rizzolatti & Arbib, 1998).

The superior temporal sulcus region (STs) and parts of TE contain neurons that are similar in response to mirror neurons (Perrett, Mistlin, Harries, & Chitty, 1990). They respond to the sight of the hand; the main difference compared to F5 is that they lack the motor response. It is likely that they participate in the processing of the visual information and then communicate with F5 (Gallese et al., 1996) most likely via the parietal cortex.

Causation in a nutshell

Animals are actors in their environment, not simply passive observers. They have the opportunity to examine the world using causality, by performing probing actions and learning from the response. In other words animals can act and consequently observe the effects of their actions. Effects can be more or less direct, e.g. I feel my hand moving as the direct effect of sending a motor command, or they can be eventually ascribed to complicate chains of causally related events producing what we simply call "a chain of causality". For example, I see the object rolling as a result of my hand pushing it as a result of a motor command. Tracing chains of causality from motor action to perception (and back again) is important both to understand how the brain deals with sensorimotor coordination and to implement those same functions in an artificial system, such as a humanoid robot. We propose that such causal probing can be arranged in a developmental sequence leading along the way to a manipulation-driven representation of objects, to the perception/interpretation of manipulative actions, and to perceiving our own body. The same analysis could be used to explain why we observe certain developmental patterns or behaviors. Vice versa, by analyzing development we can probe deeper the structure of a particular function.

Table 1 shows three levels of causal complexity that we addressed in different forms. The simplest causal chain that an actor – whether robotic or biological – may

experience is the perception of its own actions. The temporal aspect is immediate: visual information is tightly synchronized to motor commands. Once this causal connection is established, we can go further and use it to actively explore the boundaries of objects. In this case, there is one more step in the causal chain, and the temporal nature of the response may be delayed since initiating a reaching movement does not immediately elicit consequences in the environment. Finally we argue that extending this causal chain further will allow the actor to make a connection between its own actions and the actions of another. This is clearly reminiscent of what has been observed in the response of the monkey's premotor cortex.

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

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

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

A working hypothesis

Taken together the results from neuroscience suggest a critical role for motor action in perception. Certainly vision and action are intertwined at a very basic level. While an experienced adult can interpret visual scenes perfectly well without acting upon them, linking action and perception seems crucial to the developmental process that leads to that competence. We can construct a working hypothesis: that action is required for object recognition in cases where an agent has to develop categorization autonomously. Further, the ability to act is also fundamental in interpreting actions performed by a conspecific. Of course if we were in standard supervised learning setting action would not be required since the trainer would do the job of pre-segmenting the data by hand. In an ecological context, some other mechanism has to be provided. Ultimately this mechanism is the body itself that through action (under some suitable developmental rule) generates informative percepts.

A possible developmental explanation of the acquisition of these functions can be framed in terms of tracing/interpreting chains of causally related events. Although it is

still speculative, this analysis predicts that i) development of functions roughly follows a dorsal to ventral temporal gradient (i.e. see reaching, grasping, recognition in Figure 2); ii) the ability to probe longer chains triggers the emergence of new functionality and/or a new set of behaviors.

We can distinguish three main conceptual functions (similar to the schema of Arbib et al. (Arbib, 1981)): reaching, grasping (manipulation), and object recognition. These functions correspond to the three levels of causal understanding introduced in Table 2. They form also an elegant progression of abilities which emerge out of very few initial assumptions. All that is required is the interaction between the actor and the environment, and a set of appropriate developmental rules specifying what information is retained during the interaction, the nature of the sensory processing, the range of motor primitives, etc.

The results outlined in the previous sections can be streamlined into a developmental sequence roughly following a dorsal to ventral gradient. Unfortunately this is a question which has not yet been investigated in detail by neuroscientists, and there is very little empirical support for this claim (beside the work of Kovacs et al. (Kovacs, 2000)).

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

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

The next step along this hypothetical developmental route is to acquire the F5 mirror representation. We might think of canonical neurons as an association table of grasp/manipulation (action) types with object (vision) types. Mirror neurons can then be thought of as a second-level associative map which links together the observation of a manipulative action performed by somebody else with the neural representation of one's own action. Mirror neurons bring us to an even higher level of causal

understanding. In this case the action execution has to be associated with a similar action executed by somebody else. The two events do not need to be temporally close to each other. Arbitrary time delays might occur.

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

The last subsystem to develop is object recognition. Object recognition can build on manipulation in finding the boundaries of objects and segmenting them from the background. More importantly, once the same object is manipulated many times the brain can start learning about the criteria to identify the object if it happens to see it again. These functions are carried out by the infero-temporal cortex (IT). The same considerations apply to the recognition of the manipulator (either one's own, or another's). In fact, the STs region is specialized for this task. Information about object identity is also sent to the parietal cortex and contributes to the formation of the affordances. However object recognition is performed, at a minimum all information (visual in this case) pertaining to a certain object needs to be grouped during development so that a model of the object can be constructed.

Nature of causation	*Main path*	*Function and/or behavior*
Direct causal chain	VC-VIP/LIP/7b-F4-F1	Reaching
One level of indirection	VC-AIP-F5-F1	Grasping
Complex causation involving multiple causal chains	VC-AIP-F5-F1+STs+IT	Mirror neurons, mimicry
Complex causation involving multiple instances of manipulative acts	STs+TE-TEO+F5-AIP(?)	Object recognition

Table 2 Degrees of causal indirection, localization and function in the brain.

A model

The model we hypothesize extends along two dimensions: first, we try to provide a description of the development of mirror neurons (the temporal dimension), and second, the localization of different sub-functions in the brain (the spatial dimension). To this end, development of the mirror system involves also the development of reaching, grasping, and eventually the observation and interpretation/representation of the action of others. A clear cut separation of functions is perhaps an extreme stance; rather, a distributed, intermingled structure is a more plausible description of the infant’s brain. The hypothesis we put forward, for logical consistency, is that reaching develops first to enable transport of the hand close to the object. An example of modeling the development of reaching can be found in Metta et al. (Metta, Sandini, & Konczak, 1999). A second step, roughly in between reaching and full blown manipulation, is that of orienting correctly the hand. This is an example of early

prospective control which is also somewhat simpler that the complete pre-shaping needed for grasping.

The "posting" task, inserting an envelope in a mailbox, was for example used in (Milner & Goodale, 1995) to dissociate ventral- from dorsal-like visual processing. A slightly modified version of it, reaching and grasping for a rotating rod, allows studying a similar ability in a more dynamic context. The aim of this experiment is to investigate when and how infants start to control hand posture in relation to the shape of the object to be grasped and to what degree it is linked to reaching. This ability is shown very clearly in adults by the pre-shaping of the hand during the transport phase of reaching/grasping. What pre-shaping does is, in fact, to prepare the hand to the "best" contact with the object to be grasped before the object is touched. Therefore it has, at least, two important components: one is based on the physical shape of the object determining which type of grasp is best, the other is related to dynamics of the grasping action and the need to anticipate the position and orientation of the object at the time of contact. In practice, it is very hard to precisely measure the posture of the hand of infants during grasping (e.g. no data-glove is available for infant-hand size) and therefore it is experimentally difficult to investigate the onset of pre-shaping abilities. For this reason it was decided to simplify the measure by assuming that the orientation of the hand with respect to a rod-like object can be studied as an example of pre-shaping ability. Preliminary results of this experiment (as mentioned in the summary of Deliverable 4.4 of the first year report) show that approaching and grasping an object are independent actions. Further, there is evidence of prospective control of grasping.

Next, we identified two stages of the development of the mirror system and area F5 proper. The first stage is in a one-to-one correspondence with the emergence of the F5's canonical neurons. Canonical neurons could develop autonomously (without an external teacher) simply by trial-and-error-type learning. They encode information about the object identity and the type of grasping. As mentioned before, we should also consider the connections between F5 and AIP that is thought to modulate the "canonical" response. Afterwards, the development of mirror neurons can be accounted for by imagining a process that reinforces the link between executed and observed action. The executed object-bound action is coded by the canonical system; this knowledge is factored in when learning or developing the mirror-like response. In one extreme view, the presence of the object already tells the actor what is the action most likely to be observed: e.g. a pen is likely to be grasped by using a precision grip, it would be awkward (or very inconvenient) to use a power grasp. Conversely, if an object with many different affordances such as a coffee mug[2] is involved, then the combination of hand- and object-related information is required to disambiguate the action type. Therefore, we believe that both object-goal and hand appearance information is used in the primate brain. Experimentally, we can try then to elucidate what is the contribution of the vision of the hand to the response of the mirror system (see report of Deliverable 4.3 included in the first year report).

One testable possibility is that, during the ontogenetic process of motor learning, different visual information coming from the observation of one's own hand performing repetitively the same action, are associated by the brain as "reference signals" sharing the same motor goal – this is where object-related information can be incorporated and it is conceptually analogue to the response of canonical neurons. The

[2] A coffee mug affords many different grasping types e.g. grasp from the handle, a power grasp of the cup, etc.

sensory to motor coupling at the basis of the "mirror" mechanism would be initially generated by the observation of one's own acting effector (e.g. the hand seen from different perspectives and, in particular in the first development phase, during several attempts to reach the target). This visuomotor transformation process, acting initially as a control system, becomes progressively capable of generalizing from the "visual hand" to the "motor hand". It could become therefore capable to extract motor invariants also during observation of actions made by others (this is one of the goals of the experiments with the set-up described in WP 3).

The monkey experiment we are currently setting up (workpackage 4) aims at investigating the role of visual feedback relative to hand self-observation during the execution of grasping. Grasping will be performed by the monkey in: a) full vision (both object and hand visible), b) without hand vision (only the object will be illuminated) and, c) with a manipulated visual feedback –object will be dimly illuminated and the position of fingertips will be shown to the monkey by means of LEDs glued on finger nails. Note that the presence of a dim light inside the object allows grasping in the dark condition without illuminating the hand when it reaches for the object. During the experiment, both mirror and F5 motor neurons will be recorded and submitted to the same experimental paradigm. The study of F5 motor neurons is further important in order to exclude that the expected modification of mirror discharge is due to difference in motor execution induced by the experimental manipulations.

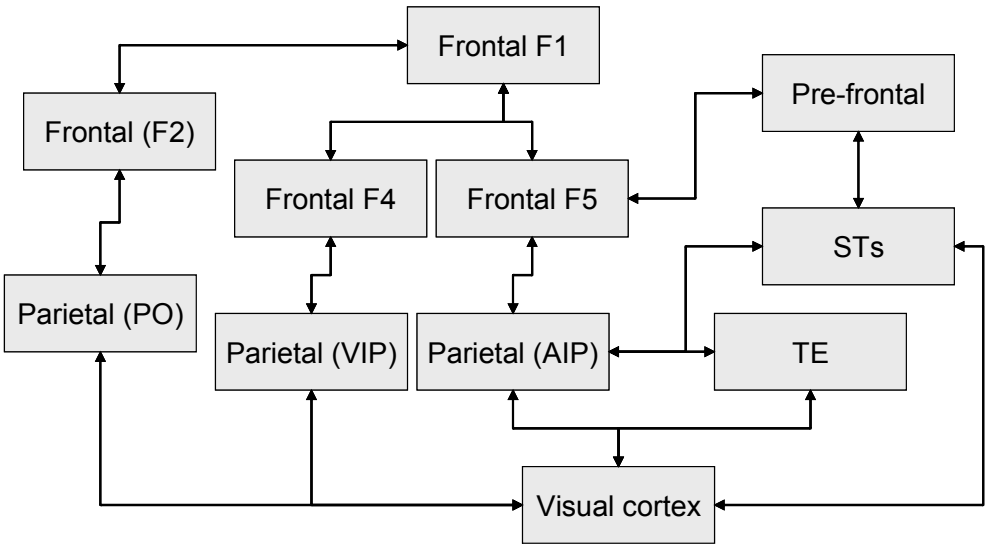


Figure 3 Block diagram of the proposed model (see text).

The block diagram in Figure 3 shows our model in some more details. The reason for producing a diagram like this one is to identify both the temporal and spatial dimension of the model in a single chart. Engineering-wise, it is also a clearer way to present things, and to see whether it is amenable of implementation.

We used artificial setups, although presently in a simplified form, to streamline results and hypotheses in a testable model. We identified two steps in the realization of a device that embeds the model of mirror neurons: simulating the artificial hand and arm by recording real human trajectories (workpackage 3), some tactile and visual data, and in a longer term by realizing a complete head-arm-hand system

(workpackage 2). The first setup has been now used to gather some initial data; however we did not yet follow a specific protocol. It is foreseeable that it could be used in the short time scale to build a realistic training set.

The role of motor information in defining a common motor goal is clear here. Actions observed from different points of view – showing a big variance – can be clustered using motor information as a learning signal – which shows a much smaller variance. In the "grasping of the cup" example, a few cluster, one for each possible grasp type, can be defined. For each cluster a much bigger set of visual information can be associated describing the visual appearance of the same action observed from different points of view. We started analyzing optic flow data as visual feature. We would like to describe the movement of the hand by means of "global features" without relying on the explicit reconstruction of the hand kinematics.

With the robotic setup (workpackage 2), we employed poking and prodding, as a precursor of grasping, in two different sets of experiments. Even if poking does not represent the whole range of complexity of grasping, we were nonetheless able to show how the complete model could be implemented in a real robotic system. From the philosophical stand point, the presence of manipulation operationally solves the problem of figure-ground segmentation, i.e. the robot is not fooled even if it encounters the yellow cube sitting on the yellow table as in Figure 1. Manipulation, even as simple as poking, allows gathering data to build models of the objects encountered during training. Needless to say that grasping can be even more powerful providing for free a simple form of object constancy: during grasping, the object remains the same unless it is dropped.

The same line of reasoning can be applied in identifying the manipulator. An operational definition of manipulator consequently calls for anything that gets in contact with an object and causes some measurable consequence. Next to the model of the object, a model of the manipulator can be formed. As a first step in this direction, we tried to visually identify the manipulator, and classify its visual appearance. Some preliminary results obtained by using a support vector machine classifier are encouraging. However, visual classification alone might seem misguided. It is intended here as the first stage of analysis. We believe that, in analogy with the superior temporal sulcus area (STs), visual information about the configuration of the hand contributes to the response of mirror neurons. Consequently, this analysis is necessary to devise suitable visual processing primitives and/or features to be employed in both artificial setups.

At a more theoretical level we analyzed the requirements to build a mirror-like representation. Briefly, two modules are required: 1) a goal matching criterion, and 2) a goal to motor transformation. The first module determines whether the observed action matches any of the possible consequences of the observer's motor repertoire. The latter, transforms the "identified" action into an actual execution. This view embeds many different possible mirror-like schemas: i.e. matching can be done in many different spaces and the transformations can be quite different. Biological plausibility calls for a matching criterion in a mixed visual-goal space: the hand's visual information, the visual description of the motor goal, and the object identity are all elements of the matching criterion (we globally call this the goal of the action). There are then two possible realization of the matching procedure. In one case all potential actions are first mapped into their motor descriptions and the comparison/match is executed in motor "coordinates". In a second case, actions are matched in the goal space and only one is then mapped into its motor description.

Again, for reason of biological plausibility, we favor the second option. Figure 4 shows graphically the two alternatives.

Experimental results so far

These are results beyond the construction of the experimental setups per se. The aim of this section is to show how the different activities fit in our conceptual (and biologically plausible) model of the development of mirror neurons.

1) G. Metta and P. Fitzpatrick: *Early Integration of Vision and Manipulation*. Submitted to Adaptive Behavior, special issue on Epigenetic Robotics. 2002. A global view of the model and a complete implementation is contained in this paper. A robotic hand wasn't available at the time. Experiments are based on poking and prodding rather than grasping.
2) Deliverable 4.4: preliminary results on the development of grasping of a rod. The document supports the view that approaching and grasping an object are two distinct processes and might develop separately.
3) Deliverable 2.3: A possible implementation of the conceptual schema shown in Figure 4 is presented. This represents a first step into the analysis of imitation by a mirror-like system. The system works in image space, while this is biologically implausible, we believe the paper shows clearly where and what are the issues in building a real world imitator.
4) L. Natale, S. Rao, G. Sandini. *Learning to act on objects*. In 2nd Workshop on Biologically Motivated Computer Vision (BMCV). Tübingen (Germany), November 22-24, 2002. This paper describes an experiment on pushing where the robot acquires a model of the behavior of objects. The repertoire of the robot consists of four different pushing actions. The robot uses this knowledge subsequently to solve a simple pushing task.

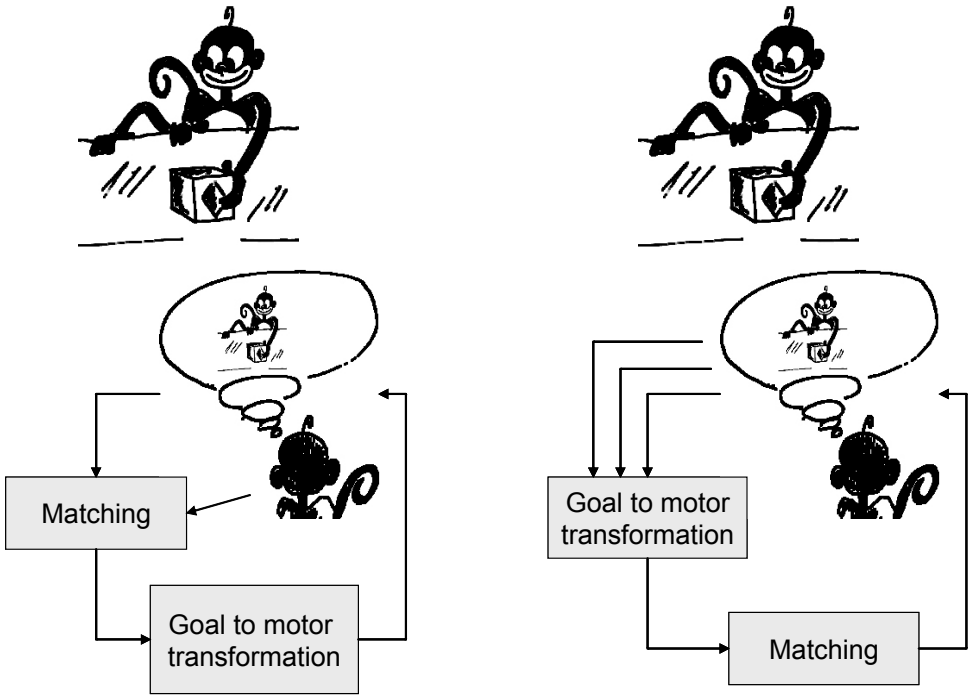


Figure 4 Two possible conceptual schemas. In panel (a) on the left, the matching criterion is applied in sensory space and then the interpretation in motor terms follows through the goal to motor transformation block. Panel (b) on the right, shows an alternate view where sensory information is first mapped into motor space and a number of potential actions is maintained. Successively, the matching criterion is applied in motor space.

Future directions

The proposed model automatically provides a view of the future directions. At least conceptually, the implementation of both artifacts is now described by the model. The next step is thus "simply" to implement and verify/test it in the real world. However, this might fulfill the goal only in part. The more difficult step is to compare the results of the modeling activity to the results from the "brain" experiments. This is a crucial and difficult step where we like to think the scientific "pot of gold" really lies.

Can we get a better understanding of the ontogenesis of the mirror system? Can we get a deeper understanding of why the brain evolved something like mirror neurons at all? Is it functionally the only one solution? These are the kind of questions we would like to answer soon.

References

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

Berkeley, G. (1972). *A new theory of vision and other writings*. London: Dent.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 425*, 429-433.

Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (2000). Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *Internation Journal of Psychophysiology, 35*(2-3), 165-177.

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

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

Fogassi, L., Gallese, V., Fadiga, L., Luppino, G., Matelli, M., & Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology*(76), 141-157.

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

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: toward an ecological psychology* (pp. 67-82). Hillsdale: Lawrence Erlbaum.

Graziano, M. S. A., Hu, X., & Gross, C. G. (1997a). Coding the Location of Objects in the Dark. *Science, 277*, 239-241.

Graziano, M. S. A., Hu, X., & Gross, C. G. (1997b). Visuo-spatial properties of ventral premotor cortex. *Journal of Neurophysiology*(77), 2268-2292.

Hendriks-Jansen, H. (1996). *Catching Ourselves in the Act*. Cambridge, MA: MIT Press.

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Cambrige, MA and Oxford UK: Blackwell Publishers Inc.

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.

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

Manzotti, R., & Tagliasco, V. (2001). *Coscienza e realta': una teoria della mente per costruttori di menti e cervelli*. -- City --: Il mulino.

Metta, G., Sandini, G., & Konczak, J. (1999). A Developmental Approach to Visually-Guided Reaching in Artificial Systems. *Neural Networks, 12*(10), 1413-1427.

Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action* (Vol. 27). Oxford: Oxford University Press.

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.

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

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences, 21*(5), 188-194.

Rizzolatti, G., Fogassi, L., & Gallese, V. (1997). Parietal cortex: from sight to action. *Current Opinion Neurobiology, 7*(4), 562-567.

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.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems, *Analysis of visual behavior* (pp. 549-586). Cambridge, MA: MIT Press.

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

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