Grounding vision through experimental manipulation

By Paul Fitzpatrick[1] and Giorgio Metta[1,2]

[1] MIT AI Laboratory, Cambridge, MA 02139, USA
[2] LIRA-Lab, University of Genova, 16145 Genova, Italy

Experimentation is crucial to human progress at all scales, from society as a whole to a young infant in its cradle. It allows us to elicit learning episodes suited to our own needs and limitations. This paper develops active strategies for a robot to acquire visual experience through simple experimental manipulation. The experiments are oriented towards determining what parts of the environment are physically coherent – that is, which parts will move together, and which are more or less independent. We argue that following causal chains of events out from the robot's body into the environment allows for a very natural developmental progression of visual competence, and relate this idea to results in neuroscience.

Keywords: active vision, mirror neuron, humanoid robot, segmentation

1. Motivation

A truly autonomous robot needs to be able to explore and learn from its environment, since it cannot rely on receiving all the information it needs passively (Whaite and Ferrie, 1997). It is telling that some of the earliest autonomous robots ever built, the tortoises of W. Grey Walter, were given the mock-Linnean designation "Machina Speculatrix" by their creator to emphasise their exploratory behaviour, described as: "it explores its environment actively, persistently, systematically as most animals do" (Walter, 1950). These robots had very simple control circuitry, and their behaviour depended greatly on the morphology and dynamics of their own bodies. This observation of the utility of a robot's body has recurred over the years, perhaps most notably in the work of Brooks (Brooks et al., 1998). It has also played a role in active approaches to machine vision, where sensors are embedded in a robotic platform and moved in a manner that simplifies visual processing (Ballard, 1991). Since correctly perceiving the world comes so naturally to humans, and appears so free of effort, the motivation for this work can be difficult to grasp at an intuitive level. We begin our paper by seeking to clarify the difficulties a robot faces in perceiving the world, and how its body can come to the rescue.

2. The elusive object

Sensory information is intrinsically ambiguous, and very distant from the world of well-defined objects in which humans believe they live. What criterion should be applied to distinguish one object from another? How can perception support such a remarkable phenomenon as figure-ground segmentation? Consider the example

in Figure 1. It is immediately clear that the drawing on the left is a cross. The
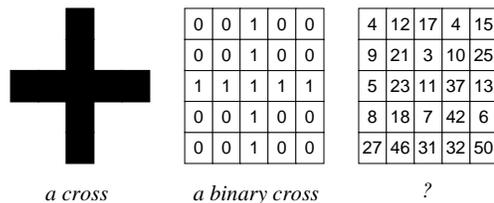

Figure 1. Three examples of crosses, following (Manzotti and Tagliasco, 2001). The human ability to segment objects is not general-purpose, and improves with experience.

intensity difference between the black cross and the white background is a powerful cue for segmentation. It is slightly less clear that the zeros and ones on the middle panel are still a cross. What can we say about the array on the right? If we are not told otherwise we might think this is just a random collection of numbers, since there is no obvious criterion to perform the figure-ground segmentation. But if we are told that the criterion is in fact "prime numbers vs. non-prime" then a cross can still be identified.

While we have to be inventive to come up with a segmentation problem that tests a human, we don't have to try hard at all to find something that baffles our robots. Figure 2 shows a robot's-eye view of a cube sitting on a table. At first glance this seems simple enough, but many rules of thumb used in machine vision for automatic object segmentation fail in this particular case. And even an experienced human observer, diagnosing the cube as a separate object based perhaps on its shadow and subtle differences in the surface texture of the cube and table, could in fact be mistaken – perhaps some malicious researcher has glued the cube tight to the table. The only way to find out for sure is to take action, and start poking and prodding. 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)

In this paper, we provide support for a more nuanced proposition: that while it is true that vision is full of illusions, these illusions evaporate when the robot can reach out and come into contact with objects – even if it has no sense of touch! While certainly touch is an important sense, we show that simply involving objects in a causal chain of events initiated by the robot itself is enough to wipe away much of the ambiguity that will plague a passive observer.

3. Grounding vision in action

Much of computer vision is passive in nature, with the emphasis on watching the world but not participating in it. There are advantages to moving beyond this to exploit dynamic regularities of the environment (Ballard, 1991). A robot has the potential to examine its world using causality, by performing probing actions and learning from the response. 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
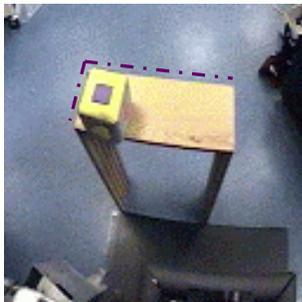
Figure 2. A cube on a table. The edges of the table and cube happen to be aligned (dashed line), the colours of the cube and table are not well separated, and the cube has a potentially confusing surface pattern.

system, such as a humanoid robot. And, as a practical matter, the ability to perform "controlled experiments", such as tapping an object lightly, is crucial to getting to grips with an otherwise complex and uncertain world.

Figure 3 illustrates three levels of causal complexity we would like our robot to probe. The simplest causal chain that the robot can experience is the perception of its own actions. The temporal aspect is immediate: visual information is tightly synchronised to motor commands. We use this strong correlation to identify parts of the robot body – specifically, the end-point of the arm. 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 doesn't immediately elicit consequences in the environment.

In this paper, we propose that such causal probing can be arranged in a developmental sequence leading to a manipulation-driven representation of objects. We present results for some important steps along such a sequence, and describe how we plan to proceed. We argue that following this causal chain outwards will allow us to approach the representational power of "mirror neurons" (Gallese et al., 1996), where a connection is made between our own actions and the actions of another.

4. Objects and actions in the human brain

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 in our infancy, we had to initially form a set of criteria to solve the object identification/segmentation problem in more mundane circumstances. We ask the question of whether we can discover these criteria during ontogenesis.

Humans and a small number of other primates are unique in their ability to manipulate their environment using tools. Our capacities are mirrored in the brain by the size of the cortex controlling them. Neuroscience has shown that our brains possess large cortical areas devoted to the control of manipulation – not surprising, given that encephalisation is believed to have evolved for the purpose of adaptively controlling action (Maturana and Varela, 1998).

A useful conceptual schema holds that visual information follows two distinct pathways in the brain, namely, the dorsal and the ventral (Ungerleider and Mishkin,
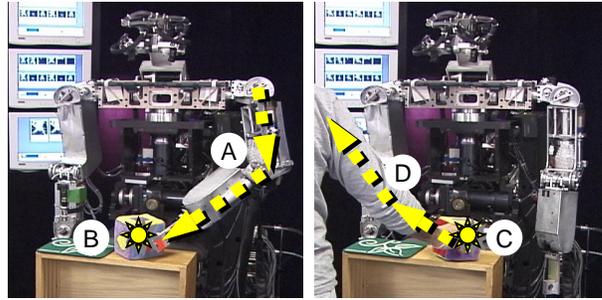
Figure 3. On the left, the robot establishes a causal connection between commanded motion and its own manipulator (A), and then probes its manipulator's effect on an object (B). The object then serves as a literal "point of contact" (C) to link robot manipulation with human manipulation (on the right, D), as is required for a mirror-neuron-like representation.

1982; Milner and Goodale, 1995). The dorsal pathway controls action directly and pragmatically; conversely, the ventral takes care of more conceptual skills such as object recognition. Of course it is important to remember, when making this dichotomy, that the two pathways are not completely segregated but rather complement each other and interact in different ways (Jeannerod, 1997).

Objects are thought to maintain a double "identity" depending on whether they are used in perceptual or in motor tasks. The concept of size, for example, might be represented multiple times in different brain areas. Observation of agnosic patients (Jeannerod, 1997) shows an even more complicated relationship than the simple dorsal/ventral dichotomy would suggest. 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-pathway system, this implies that the ventral representation of the object can supply the dorsal system with size information. What we consciously perceive as "size" is rather a collection of different percepts interacting in a complicated way, and under pathological circumstances they can be separated from each other. One of the "identities" of objects is thus connected to motor performance.

That such pathways develop and are not completely innate is suggested by the results of (Kovacs, 2000). She 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. These results further suggest that the development of action might precede that of categorisation: it is well established that by 4 months of age infants can process complex motion stimuli, depth, and colour. Roughly at the same age reaching becomes more consistent. That is, action comes first supported by the pragmatic use of diverse sensory modalities; perception conversely is a long developing process. More studies are needed though to ascertain how the dorsal pathway (action) influences the ventral (perception) both in situations like those already mentioned, and during ontogenesis.

Drawing more from the neural science literature, the results of Fogassi et al. (Fogassi et al., 1996) and Graziano et al. (Graziano et al., 1997) have shown the existence of neurons that respond to objects and are related to the description

of the peripersonal space with respect to reaching (area F4 and VIP). A subset of the F4 neurons have 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 receptive field is usually in register with the visual one. Motor information is integrated into the representation by maintaining the receptive field anchored to the corresponding body part (the forearm in this example) irrespective of the relative position of the head and arm. F4, together with areas in the parietal lobe, is thought to participate in the visual to motor transformations required to control reaching.

While F4 is concerned with the proximal muscles (i.e. reaching), F5 controls more distal muscles (i.e. the hand). Areas in the parietal lobe, such as AIP, also project to F5 in the premotor cortex. For many years the premotor cortex was considered just another area related solely to motor control. New studies (see (Jeannerod, 1997) for a review) have demonstrated that this is not the case. We have already described the properties of the neurons in F4; similarly, researchers have identified neurons in the area F5 of the frontal cortex (Fadiga et al., 2000) that are activated in two situations: *i)* when the host is acting upon an object (e.g. grasping), and *ii)* when looking at the same object (visual response). The corresponding firing patterns are quite specific, building a link between the size and shape of the object and the applied grasp type (e.g. a small object requires a precision grip). These neurons are called canonical. At the time, this was quite an astonishing discovery because area F5 was believed to be only a motor area. A possible interpretation is that the brain stores a representation of objects in motor terms, and uses these representations to generate an appropriate response to objects. Arbib and colleagues (Fagg and Arbib, 1998) interpreted these responses as the neural analogue of the affordances of Gibson (Gibson, 1977). In Gibson's theory an affordance is a visual characteristic of an object which can elicit an action without necessarily involving an object recognition stage. It seems that areas AIP and F5 are active in such a way as to provide the individual with a mechanism to detect affordances. F5 projects to the primary motor cortex and can therefore control behaviour.

The gap from object manipulation to hand gesture production and recognition is small. Rizzolatti and coworkers (Gallese et al., 1996) extensively probed area F5. They found another class of grasping neurons that also respond during observation of somebody else's action. A typical cell of this class (called mirror neuron) indeed responds in two situations: *i)* when executing a manipulative gesture, and *ii)* when observing somebody else executing the same action. These neurons provide a link between the observation of somebody else's and our own actions. The activation of F5 is consistent with the idea that the brain internally reproduces/simulates the observed actions. Along with their use for the recognition of manipulative actions, mirror neurons are thought to support imitative behaviours. An intriguing theory proposed by Rizzolatti and Arbib (Rizzolatti and Arbib, 1998) associates mirror neurons to language. In (Wohlscläger and Bekkering, 2002) the role of objects during an imitative task was tested. Also, Woodward (Woodward, 1998) investigated the role of objects in the understanding of action performed by others. In a series of experiments she elucidated the contribution that seeing an object makes for 5, 6, and 9 month old infants. Taken together, these results provided evidence that the object and the goal-directeness of the action represent an important component in the understanding of the intentions of others.

5. Objects and actions in robotic systems

Certainly, vision and action are intertwined at a very basic level in humans. 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. While not focusing on development, many researchers in machine vision have adopted the view that vision and action need to be tightly integrated for functional reasons. Their work is loosely termed "active vision". A vision system is said to be *active* if it is embedded within a physical platform that can act to improve perceptual performance. For example, a robot's cameras might servo a rapidly moving target in order to stabilise the image and keep the target in view. In fact, active vision is often equated with moving cameras, although in this paper we use it in a broader sense of any controllable resource recruited to serve vision (including, in our case, arm motion).

Historically, a number of logically distinct ideas are often associated with active vision. The first is that vision should be approached within the the context of an overall task or purpose (Aloimonos et al., 1987). If an observer can engage in controlled motion, it can integrate data from frame to frame to solve problems that are ill-posed statically. Well-chosen motion can simplify the computation required for widely studied vision problems, such as stereo matching (Bajcsy, 1988; Ballard, 1991). These interwoven ideas about active vision are teased apart in (Tarr and Black, 1994).

In our work, we show that the entire body can usefully be recruited to cooperate with the vision system, and we need not limit ourselves to just the head. In particular, we show that probing arm movements can be very revealing, and allow us to tackle long-standing problems in machine vision in an innovative way. We demonstrate that simple poking gestures (prodding, tapping, swiping, batting, etc.) are rich enough to evoke object affordances such as rolling and to provide the kind of training data needed to bootstrap perception.
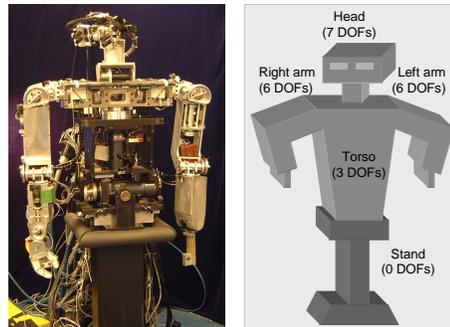

Figure 4. Degrees of freedom (DOFs) of the robot Cog. The arms terminate either in a primitive "flipper" or a four-fingered hand. The head, torso, and arms together contain 22 degrees of freedom.

6. Our experimental platform

This work is implemented on the robot Cog, an upper torso humanoid (Brooks et al., 1999). Cog has two arms, each of which has six degrees of freedom. The joints are driven by series elastic actuators (Williamson, 1999). The arm is not designed to enact trajectories with high fidelity. For that a very stiff arm is preferable. Rather, it is designed to perform well when interacting with a poorly characterised environment, where collisions are frequent and informative events. Cog runs an attentional system consisting of a set of pre-attentive filters sensitive to motion, colour, and binocular disparity. The different filters generate information on the likelihood that something interesting is happening in a certain region of the image. A voting mechanism is used to "decide" what to attend and track next. The pre-attentive filters are implemented on a space-variant imaging system, which mimics the distribution of photoreceptors in the human retina as in (Sandini and Tagliasco, 1980). The attentional system uses vision and non-visual sensors (e.g. inertial) to generate a range of oculomotor behaviours. Examples are saccades, smooth pursuit, vergence, and the vestibulo-ocular reflex (VOR).

7. Perceiving the body in action

Motion of the arm may generate optic flow directly through the changing projection of the arm itself, or indirectly through an object that the arm is in contact with. While the relationship between the optic flow and the physical motion is likely to be complex, the correlation in time of the two events should be exceedingly precise. This time-correlation can be used as a "signature" to identify parts of the scene that are being influenced by the robot's motion, even in the presence of other distracting motion sources. In this section, we show how this tight correlation can be used to localise the arm in the image without any prior information about visual appearance.

Reaching out

The first step towards manipulation is to reach objects within the workspace. If we assume targets are chosen visually, then ideally we need to also locate the end-effector visually to generate an error signal for closed-loop control. Some element of open-loop control is necessary since the end-point may not always be in the field of view (for example, when it is in its resting position), and the overall reaching operation can be made faster with a feed-forward contribution to the control.

The simplest possible open loop control would map directly from a fixation point to the arm motor commands needed to reach that point (Metta et al., 1999) using a stereotyped trajectory, perhaps using postural primitives (Mussa-Ivaldi and Giszter, 1992). If we can fixate the end-effector, then it is possible to learn this map by exploring different combinations of direction of gaze vs. arm position (Marjanović et al., 1996; Metta et al., 1999). So locating the end-effector visually is key both to closed-loop control, and to training up a feed-forward path. We shall demonstrate that this localisation can be performed without knowledge of the arm's appearance, and without assuming that the arm is the only moving object in the scene.
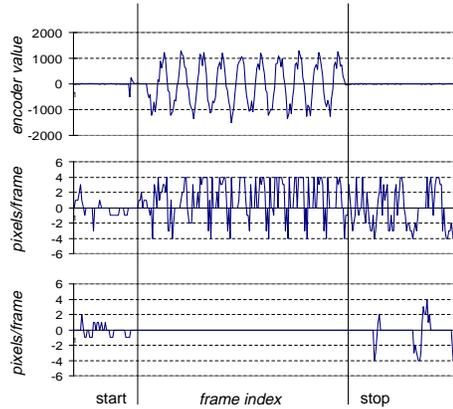
Figure 5. An example of the correlation between optic flow and arm movement. The traces show the movement of the wrist joint (upper plot) and optic flow sampled on the arm (middle plot) and away from it (lower plot). As the arm generates a repetitive movement, the oscillation is clearly visible in the middle plot and absent in the lower. Before and after the movement the head is free to saccade, generating the other spikes seen in the optic flow.

Localising the arm visually

The robot is not a passive observer of its arm, but rather the initiator of its movement. This can be used to distinguish the arm from parts of the environment that are more weakly affected by the robot. The arm of a robot was detected in (Marjanović et al., 1996) by simply waving it and assuming it was the only moving object in the scene. We take a similar approach here, but use a more stringent test of looking for optic flow that is correlated with the motor commands to the arm. This allows unrelated movement to be ignored. Even if a capricious engineer were to replace the robot's arm with one of a very different appearance, and then stand around waving the old arm, this detection method will not be fooled.

The actual relationship between arm movements and the optic flow they generate is complex. Since the robot is in control of the arm, it can choose to move it in a way that bypasses this complexity. In particular, if the arm rapidly reverses direction, the optic flow at that instant will change in sign, giving a tight, clean temporal correlation. Since our optic flow processing is coarse (a 16×16 grid over a 128×128 image at 15 Hz), we simply repeat this reversal a number of times to get a strong correlation signal during training. With each reversal the probability of correlating with unrelated motion in the environment goes down.

Figure 5 shows an example of this procedure in operation, comparing the velocity of the arm's wrist with the optic flow at two positions in the image plane. A trace taken from a position away from the arm shows no correlation, while conversely the flow at a position on the wrist is strongly different from zero over the same period of time. Figure 6 shows examples of detection of the arm and rejection of a distractor.
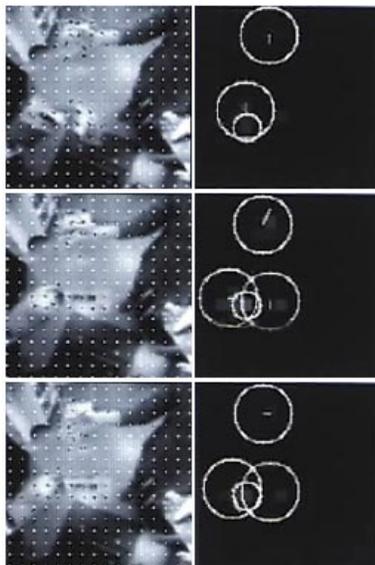
Figure 6. Detecting the arm/gripper through motion correlation. The robot's point of view and the optic flow generated are shown on the left. On the right are the results of correlation. Large circles represent the results of applying a region growing procedure to the optic flow. Here the flow corresponds to the robot's arm and the experimenter's hand in the background. The small circle marks the point of maximum correlation, identifying the regions that correspond to the robot's own arm.
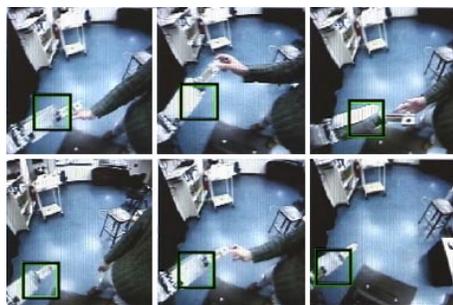

Figure 7. Predicting the location of the arm in the image as the head and arm change position. The rectangle represents the predicted position of the arm using the map learned during a twenty-minute training run. The predicted position just needs to be sufficiently accurate to initialise a visual search for the exact position of the end-effector.

Localising the arm using proprioception

The localisation method for the arm described so far relies on a relatively long "signature" movement that would slow down reaching. This can be overcome by training up a function to estimate the location of the arm in the image plane from proprioceptive information (joint angles) during an exploratory phase, and using that to constrain arm localisation during actual operation. Figure 7 shows the resulting behaviour after about twenty minutes of real-time learning.

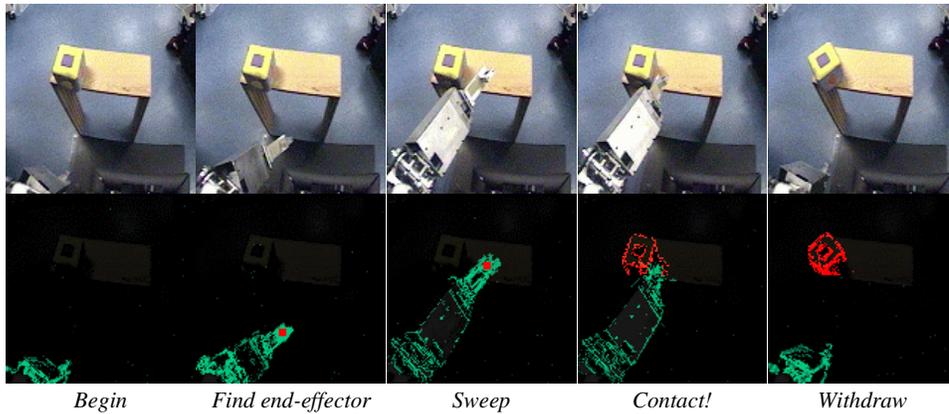| Begin | Find end-effector | Sweep | Contact! | Withdraw |

Figure 8. The upper sequence shows an arm extending into a workspace, tapping an object, and retracting. This is an exploratory mechanism for finding the boundaries of objects, and essentially requires the arm to collide with objects under normal operation, rather than as an occasional accident. The lower sequence shows the shape identified from the tap using simple image differencing and flipper tracking.

8. Perceiving actions on objects

Now that the robot knows something about its arm, it can start to use it to explore its environment. When the arm enters into contact with an object, one of several outcomes are possible. If the object is large, heavy, or otherwise unyielding, motion of the arm may simply be resisted without any visible effect. Such objects are of little interest, except in their role as obstacles, since the robot will not be able to manipulate them. But if the object is smaller, it is likely to move somewhat in response to the nudge of the arm. This movement will be temporally correlated with the time of impact, and will be connected spatially to the end-effector – constraints that are not available in passive scenarios (Birchfield, 1999). If the object is reasonably rigid, and the movement has some component in parallel to the image plane, the result is likely to be a flow field whose extent reflects the physical boundaries of the object. This visible response to the robot's action can be used to refine its model of the object's extent, which may be inaccurate. For the example scene in Figure 2 (a cube sitting on a table), the small inner square on the cube's surface pattern might be selected as a target. The robot can certainly reach towards this target, but grasping it would prove difficult without a correct estimate of the object's physical extent. In this section we show how the robot can experimentally determine an object's extent using the same idea of correlated motion used earlier to detect its own arm.

Making an impact

Figure 8 shows how a "poking" movement can be used to refine a target. During this operation, the arm begins by extending outwards from the resting position. The end-effector (or "flipper") is localised as the arm sweeps rapidly outwards, using the heuristic that it lies at the highest point of the region of optic flow swept out by the arm in the image (the head orientation and reaching trajectory are controlled so

that this is true). The arm is driven outward into the neighbourhood of the target which we wish to define, stopping if an unexpected obstruction is reached. If no obstruction is met, the flipper makes a gentle sweep of the area around the target. This minimises the opportunity for the motion of the arm itself to cause confusion; the motion of the flipper is bounded around the endpoint whose location we know from tracking during the extension phase, and can be subtracted easily. Flow not connected to the end-effector can be ignored as a distractor.

The sequence shown in Figure 8 is about the simplest case possible for segmenting the motion of the object, since most of the arm is stationary when contact occurs. In practise, we would rather have less constraints on the motion of the arm, so we can approach the object from any convenient direction. We found that it was possible to attain this flexibility without losing the simplicity of object segmentation that poking brings, by exploiting the unique visual opportunity afforded by the moment of impact.

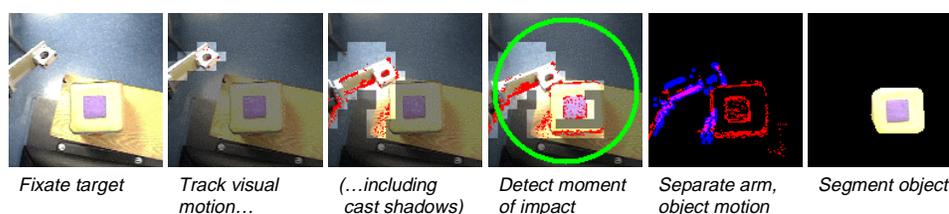| Fixate target | Track visual motion… | (…including cast shadows) | Detect moment of impact | Separate arm, object motion | Segment object |

Figure 9. The moment of impact is detected visually by the sudden expansion of motion away from the arm. Motion before and after contact is compared to gather information for segmentation.

The moment of (ground) truth

How can we detect when the arm collides with an object? One natural possibility would be to use proprioceptive or tactile information from the arm itself. Another possibility is to detect the collision visually. This is the method we use since, as Section 9 will discuss, it allows collision detection to be applied to human arm motion, a situation where the robot does not have access to any privileged information about the motion. When the robot is attempting to poke a target, it keeps the target fixated, so that the image processing does not need to compensate for egomotion. Under these conditions, it is possible to detect motion using image differencing. This is a very simple technique for detecting motion by simply subtracting successive frames coming from a camera and looking for pixel-level differences. A moving object that has some contrast with the background it is moving over will generate such differences. Of course, pixel differences can also be generated by changes in illumination, cast shadows, the refresh rate of computer monitors, movement of the camera itself, etc. A related technique called background modelling tries to estimate the appearance of the fixed, stationary background of a scene, and then subtract the current view from the reference to detect new foreground. Cog uses such a technique to detect motion while it is fixating a target.

Figure 9 shows the sequence of processing steps taken as the arm approaches and comes into contact with a target. As the arm approaches, its motion is tracked
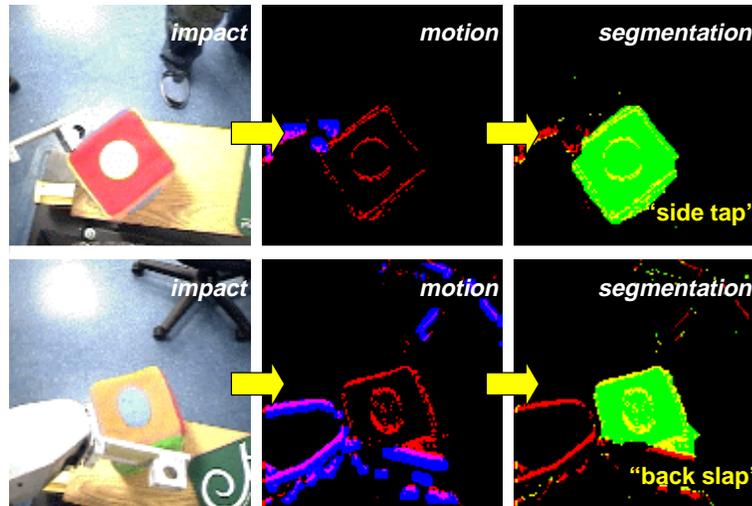
Figure 10. Cog batting a cube around. The top row shows the flipper poking an object from the side, turning it slightly. The second row shows Cog batting an object away. The images in the first column are frames prior to a collision. The second column shows the actual impact. The third column shows the motion signal at the point of contact. The bright regions in the images in the final column show the segmentations produced for the object.

very coarsely in real-time, and areas it passes through are marked as "clear" of the object. An impact event is detected through a signature explosion of movement connected with the arm, but spread across a much wider distance than the arm could possibly have moved in the time available. Once this impact is detected, we start to process at high resolution (and drop briefly out of real-time operation for a few seconds). The raw motion signature generated by the collision is computed. The translational component of the arm motion at the point of contact is also computed, so that motion present in previous frames can be aligned with the collision frame, and motion associated with the arm can be isolated from motion due to the target object. Since the impact may occur just before a frame is sampled (every 30 milliseconds) and so generate a relatively weak motion signature, motion information from one frame after collision is projected back and pooled with motion information in the collision frame. In the absence of strong texture there may be little apparent motion in the interior of the object, so we recruit a maximum-flow algorithm due to (Boykov and Kolmogorov, 2001) to fill in such regions efficiently. Figure 10 shows examples of segmentations generated by very different poking operations – one a gentle tap from the side, the other a violent "back-slap" striking the object away from the robot.

An operational definition of objecthood

The poking operation gives clear results for a rigid object that is free to move. What happens for non-rigid objects and objects that are attached to other objects? Here the results of poking are likely to be more complicated to interpret – but in a sense this is a good sign, since it is in just such cases that the idea of an

object becomes less well-defined. Poking has the potential to offer an operational theory of "objecthood" that is more tractable than a vision-only approach might give, and which cleaves better to the true nature of physical assemblages. The idea of a physical object is rarely completely coherent, since it depends on where you draw its boundary and that may well be task-dependent. Poking allows us to determine the boundary around a mass that moves together when disturbed, which is exactly what we need to know for manipulation. As an operational definition of object, this has the attractive property of breaking down into ambiguity in the right circumstances – such as for large interconnected messes, floppy formless ones, liquids, and so on. Poking also gives the robot the opportunity to collect many views of a single object, and so we can hope to deal with recognising objects like the cube shown in Figure 11, which look different from every side.
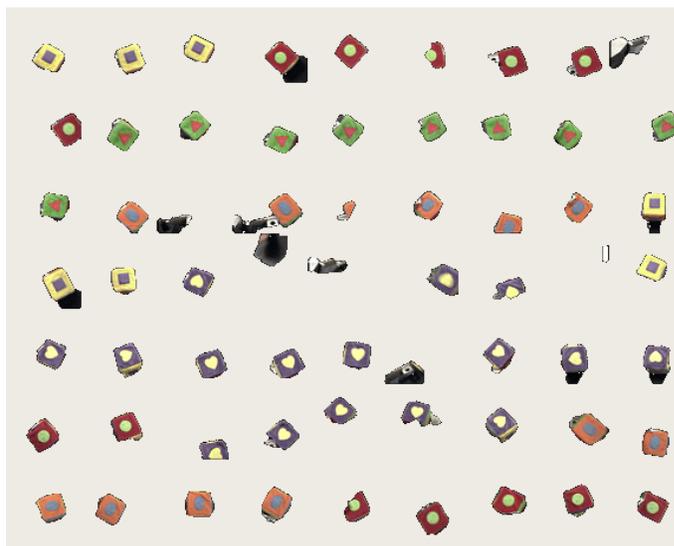

Figure 11. Results of a training session, where a toy cube was repeatedly offered to the robot for poking. Each image of the cube corresponds to the segmentation found for it during a single poke. The most common failure mode is inclusion of the robot arm in the segmentation.

9. Perceiving actors on objects

Previous sections have shown how the robot can, without any prior knowledge of their appearance, familiarise itself with objects in its environment and its own arm. It is possible to go still further, and find human arms and hands in the environment, again without any prior knowledge of their appearance. We could imagine segmenting any moving objects in the scene, and relying on the heuristic that hands are often the fastest moving objects around. A more principled approach is to adopt an operational definition of what an arm is in terms that the robot can perceive, just as we defined objects earlier by their physical coherence. Once it has poked around, the robot will be familiar with a set of objects in its environment that are the right scale for it to manipulate. A reasonable definition for an arm, then, is something

Figure 12. The segmentation algorithm will work for human poking operations, if the robot is fixating the object the human pokes. The robot's attention system makes this a simple condition to satisfy.

that acts upon these familiar objects. That includes the robot's own arm, but may well include the arm of humans around the robot. The ideal situation to detect arms is when they make contact with a familiar object that the robot is fixating, since all the visual processing already developed carries over without modification (see Figure 12). Of course the robot can easily distinguish segmentations of its own arm from that of others simply by whether its arm was near the target at the time! Our experience shows that clean segmentations of the arm (human or robot) can be extracted as it approaches collision with a known object, when we don't need to worry about the full complexity of human movement (see Figure 13). But this work is still at an early stage.
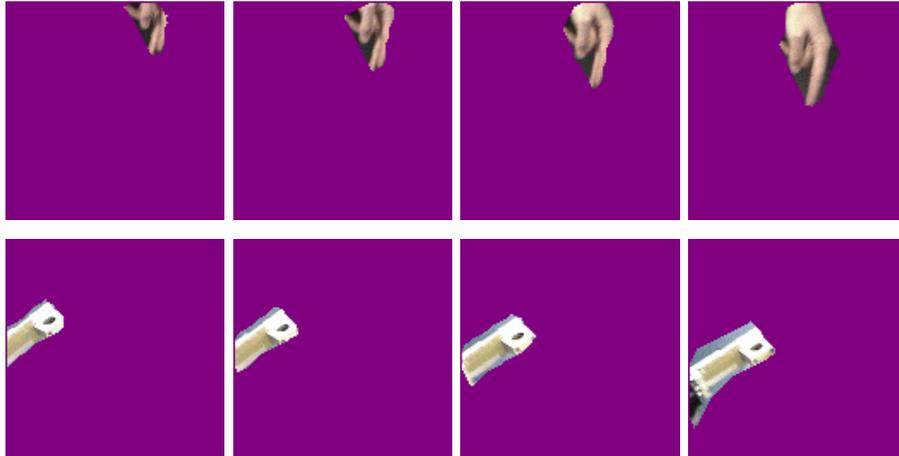

Figure 13. Early experiments on segmenting the robot arm, or a human arm poking an object the robot. The segmentation is performed by working backwards from a collision with a known object.

10. Developing mirror neurons

Poking moves us one step outwards on a causal chain away from the robot and into the world, and gives a simple experimental procedure for segmenting objects.

There are many possible elaborations of this method, all of which lead to a vision system that is tuned to acquiring data about an object by seeing it manipulated by the robot. An interesting question then is whether the system could extract useful information from seeing an object manipulated by someone else. In the case of poking, the robot needs to be able to estimate the moment of contact and to track the arm sufficiently well to distinguish it from the object being poked. We are interested in how the robot might learn to do this. One approach is to chain outwards from an object the robot has poked. If someone else moves the object, we can reverse the logic used in poking – where the motion of the manipulator identified the object – and identify a foreign manipulator through its effect on the object.

We designed two experiments that use poking and the visual segmentation described in the previous sections to probe the structure of objects and control behaviour on the basis of their affordances. Further, although poking gives us a simple procedure for segmenting objects, the procedure would be nevertheless inconvenient in many situations if we had to poke an object every time we needed to grasp it. A better solution is to learn from experience about the behaviour, visual appearance and physical properties of objects.

In the first experiment the robot poked a small set of objects (an orange juice bottle, a toy car, a cube, and a coloured ball) using one of four possible actions (the motor repertoire). Actions are labelled for convenience as pull in, side tap, push away, and back slap (see for example Figure 10). The toy car and the bottle have a definite principal axis that can be easily extracted from the segmented image. They also tend to roll along a definite direction with respect to their principal axis. These visual and physical properties of the objects can be acquired automatically by the robot simply by poking the same object many times (about 100 in our experiment). The results are shown in Figure 14. We plotted there the estimated probability of observing each of the objects rolling along a particular direction with respect to its principal axis. Different trials were clustered using colour information. In fact, in this case, colour is sufficient to distinguish the objects from each other. The next step is to acquire an understanding of poking. This is easily obtained from the same training set. Instead of considering each object separately here, we simply measured the average direction of movement given a certain action. In practise, the robot automatically learns that poking from the left causes the object to slide/roll to the right. A similar consideration applies to the other actions.

After the training stage, if one of the known objects is presented to Cog, the object is recognised, localised and its orientation estimated (principal axis). Recognition and localisation are based on the same colour segmentation algorithm used during learning. Cog then uses its understanding of the affordance of the object (Figure 14) and of the geometry of poking to make the object roll. The whole localisation procedure has an error between $10°$ and $25°$ which proved to be acceptable for our experiment. We performed a simple qualitative test of the overall performance of the robot. Out of 100 trials the robot made 15 mistakes. Twelve of them were due to imprecise control: e.g. the end point touched the object earlier than expected moving the car outside the field of view. The remainders (3) were genuine mistakes due to misinterpretation of the object position/orientation.

This experiment represents an analogue of the response of F5/AIP as explained in Arbib's model (Fagg and Arbib, 1998) in that a specific structure of the robot
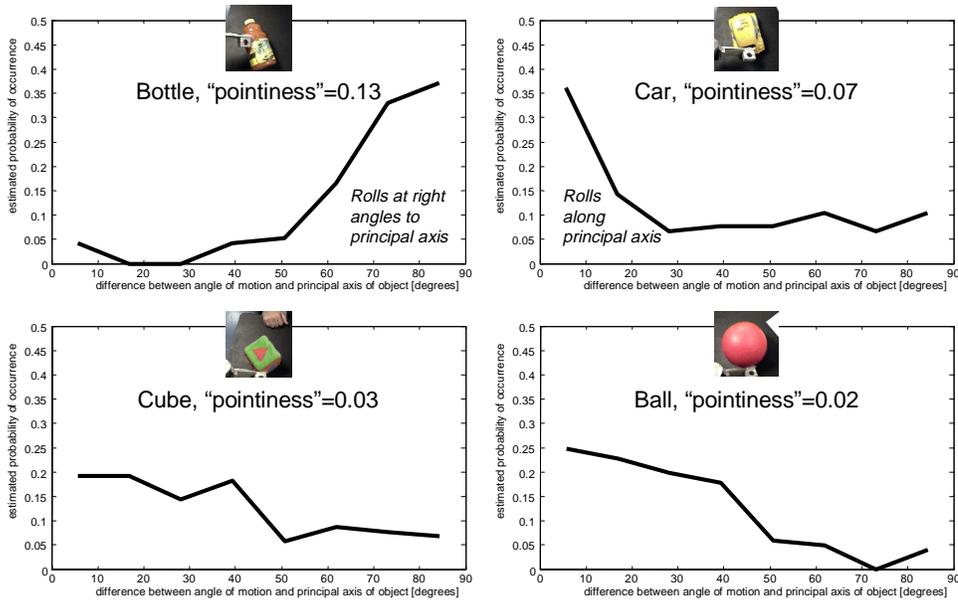
Figure 14. Probability of observing a roll along a particular direction for the set of four objects used in Cog's experiments. Abscissae represent the difference between the principal axis of the object and the observed direction of movement. Ordinates are the estimated probability.

detects the affordance of the object and links it to the generation of behaviour. This is also the first stage of the development of more complex behaviours which rely on the understanding of objects as physical entities with specific properties.

With the knowledge about objects collected in the previous experiment we can then set up a second experiment where the robot observes a human performing an action. In fact, the same visual processing used for analysing a robot-generated action can be used in this situation also, to detect contact and segment the object from the human arm. The first obvious step the robot can take is to identify the action observed with respect to its own motor vocabulary. This is easily done by comparing the displacement of the object with the four possible actions and by choosing the action whose effects are closer to the observed displacement. This procedure is orders of magnitude simpler than trying to completely characterise the action in terms of the observed kinematics of the movement. Here the complexity of the data we need to obtain is somewhat proportional to the complexity of the goal rather than that of the structure/skill of the foreign manipulator.

The robot can also mimic the observed behaviour if it happens to see the same object again. This requires another piece of information. The angle between the affordance of the object (prefered direction of motion) and the observed displacement is measured. During mimicry the object is localised as in the previous experiment and the action which is more likely to produce the same observed angle (relative to the object) is generated. If, for example, the car was poked at right angle with respect to its principal axis Cog would mimic the action by poking the car at right angle, despite the fact that the car's prefered behaviour is to move along its princi-
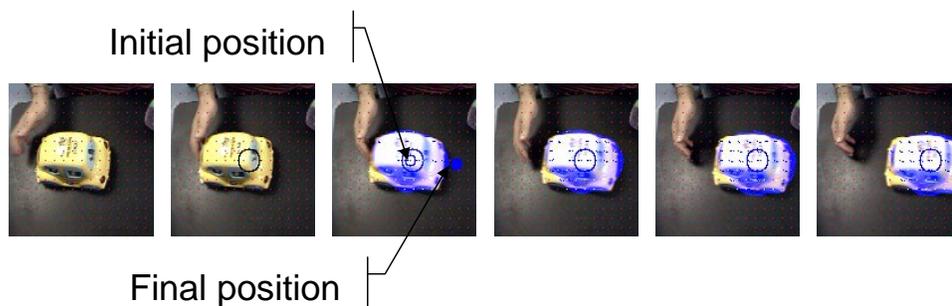
Figure 15. An example of observed sequence. Frames around the instant of impact are shown. Initial and final position after 12 frames is indicated.
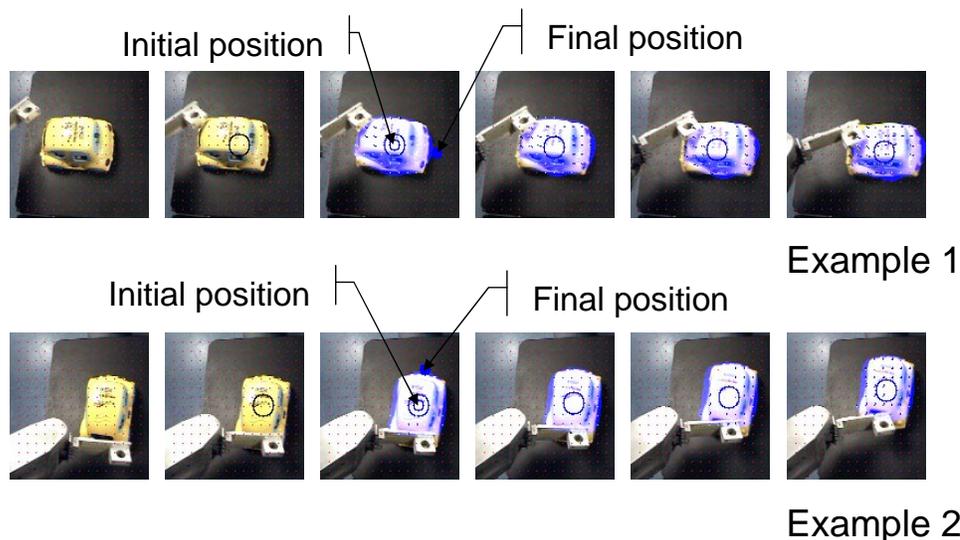

Figure 16. Two examples of mimicry following the observation of Figure 15. Cog mimics the goal of the action (poking along the principal axis) rather than the trajectory followed by the toy car.

pal axis. Examples of observation of poking and generation of mimicry actions are shown in Figures 15 and 16.

As we described before, mirror neurons respond when either watching somebody else performing a manipulative action or when actually manipulating an object. They can be thought of as an association map which links together the observation of a manipulative action performed by somebody else with the neural representation of one's own action. The question of whether a mirror-like representation can be autonomously developed by the robot (or a human for that matter) can then be answered. The association map can be constructed by identifying when the goal and the object are the same irrespective of who is the actor. Actions that lead to the same consequences are thus part of the same equivalence class. This is exactly what mirror neurons represent.

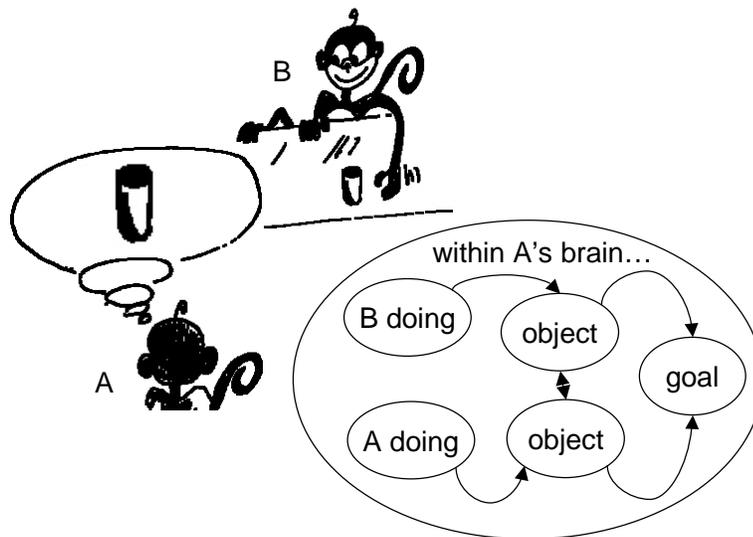Figure 17 shows this causal chain in action. There are a series of interesting

Figure 17. Mirror neurons and causality: from the observer's point of view (A), understanding B's action means mapping it onto the observer's own motor repertoire. If the causal chain leading to the goal is already in place (lower branch of the graph) then the acquisition of a mirror neuron for this particular action/object is a matter of building and linking the upper part of the chain to the lower one. There are various opportunities to reinforce this link either at the object level, at the goal level or both.

behaviours that can be realised based on mirror neurons. Mimicry is an obvious application, since it requires just this type of mapping between other and self in terms of motor actions. Another important application is the prediction of future behaviour from current actions, or even inverting the causal relation to find the action that most likely will get to the desired consequence.

11. Discussion and Conclusions

In this paper, we showed how causality can be probed at different levels by the robot. Initially the environment was the body of the robot itself, then later a carefully circumscribed interaction with the outside world. This is reminiscent of Piaget's distinction between primary and secondary circular reactions (Ginsburg and Opper, 1978). Objects are central to interacting with the outside world. We raised the issue of how an agent can autonomously acquire a working definition of objects.

The number of papers written on techniques for visual segmentation is vast. Methods for characterising the shape of an object through tactile information are also being developed, such as shape from probing (Paulos, 1999) or pushing (Moll and Erdmann, 2001). But while it has long been known that motor strategies can aid vision (Ballard, 1991), work on active vision has focused almost exclusively on moving cameras. There is much to be gained by bringing a manipulator into the equation (Tsikos and Bajcsy, 1991). The implications may be far reaching. For example, we have shown that, without any prior knowledge of the human form, the robot can identify episodes when a human is manipulating objects that are familiar

to the robot purely by the operational similarity of the human arm and its own manipulator in this situation.

We have related our work to some very interesting results from neurobiology that have implications for sensorimotor integration, such as the discovery of mirror neurons. Our view is that while biologists are doing a good job of elucidating *what* mirror neurons are and how they operate, work like ours can more readily clarify *why* they are useful in practise. We believe the answer lies in the developmental process. A creature created fully formed could perhaps operate just fine without mirror neurons, but to reach adult competence from a more primitive stage requires continuously interleaving perception with experimental action – a situation that seems to call for mirror neurons and similar machinery. Our goal has been to build a robot capable of such experimentation, and to identify specific functional advantages of mirror-like representation in the development of its visual competence. Knowledge of functional advantages could suggest new and interesting relationships for biologists to look for that they may not have thought of (since they have never tried to actually build a vision system themselves).

Acknowledgements

References

Aloimonos, J., Weiss, I., and Bandopadhay, A. (1987). Active vision. *International Journal on Computer Vision*, 2:333–356.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, 76(8):996–1005.

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

Berkeley, G. (1972). *A new theory of vision and other writings.* Dent, London. First published in 1734.

Birchfield, S. (1999). *Depth and Motion Discontinuities.* PhD thesis, Dept. of Electrical Engineering, Stanford University.

Boykov, Y. and Kolmogorov, V. (2001). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 359–374.

Brooks, R. A., Breazeal, C., Irie, R., Kemp, C. C., Marjanović, M., Scassellati, B., and Williamson, M. M. (1998). Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence (AAAI-98)*.

Brooks, R. A., Breazeal, C., Marjanovic, M., and Scassellati, B. (1999). The Cog project: Building a humanoid robot. *Lecture Notes in Computer Science*, 1562:52–87.

Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (2000). Visuomotor neurons: ambiguity of the discharge of 'motor' perception? *International Journal of Psychophysiology*, 35:165–177.

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

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

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

Gibson, J. J. (1977). The theory of affordances. In Shaw, R. and Bransford, J., editors, *Perceiving, acting and knowing: toward an ecological psychology*, pages 67–82. Hillsdale NJ: Lawrence Erlbaum Associates Publishers.

Ginsburg, H. and Opper, S. (1978). *Piaget's theory of intellectual development*. Prentice-Hall, Englewood Cliffs, NJ. 2nd edition.

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

Jeannerod, M. (1997). *The Cognitive Neuroscience of Action*. Blackwell Publishers Inc., Cambridge Massachusetts and Oxford UK.

Kovacs, I. (2000). Human development of perceptual organization. *Vision Research*, 40(10-12):1301–1310.

Manzotti, R. and Tagliasco, V. (2001). *Coscienza e realtà: una teoria della coscienza per costruttori di menti e cervelli*. il Mulino.

Marjanović, M. J., Scassellati, B., and Williamson, M. M. (1996). Self-taught visually-guided pointing for a humanoid robot. In *From Animals to Animats: Proceedings of 1996 Society of Adaptive Behavior*, pages 35–44, Cape Cod, Massachusetts.

Maturana, R. and Varela, F. (1998). *The tree of knowledge, the biological roots of human understanding*. Boston & London. Shambhala Publications, revised edition.

Metta, G., Sandini, G., and Konczak, J. (1999). A developmental approach to visually-guided reaching in artificial systems. *Neural Networks*, 12:1413–1427.

Milner, A. D. and Goodale, M. A. (1995). *The visual brain in action*, volume 27. Oxford University Press.

Moll, M. and Erdmann, M. A. (2001). Reconstructing shape from motion using tactile sensors. In *Proc. 2001 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, Maui, HI.

Mussa-Ivaldi, F. A. and Giszter, S. F. (1992). Vector field approximation: a computational paradigm for motor control and learning. *Biological Cybernetics*, 67:491–500.

Paulos, E. (1999). Fast construction of near optimal probing strategies. Master's thesis, University of California, Berkeley.

Rizzolatti, G. and Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, 21:188–194.

Sandini, G. and Tagliasco, V. (1980). An anthropomorfic retina-like for scene analysis. *Computer Vision, Graphics and Image Processing*, 14(3):365–372.

Tarr, M. and Black, M. (1994). A computational and evolutionary perspective on the role of representation in vision. *CVGIP: Image Understanding*, 60(1):65–73.

Tsikos, C. and Bajcsy, R. (1991). Segmentation via manipulation. *IEEE Transactions on Robotics and Automation*.

Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. In *Analysis of visual behavior*, pages 549–586. MIT Press, Cambridge, Massachusetts.

Walter, W. G. (1950). An imitation of life. *Scientific American*, 182(5):42–45.

Whaite, P. and Ferrie, F. P. (1997). Autonomous exploration: Driven by uncertainty. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(3):193–205.

Williamson, M. (1999). *Robot Arm Control Exploiting Natural Dynamics*. PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

Wohlscläger, A. and Bekkering, H. (2002). Is human imitation based on a mirror-neurone system? Some behavioural evidence. *Experimental Brain Research*, 143:335–341.

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