1 Closing the loop

Poking is sufficient to explore some very simple affordances of objects, such as rolling, toppling, breaking etc. We explore the rolling affordance with the four objects shown in Figure 1. Each of the objects rolls in a different way, which the robot can learn about and exploit.

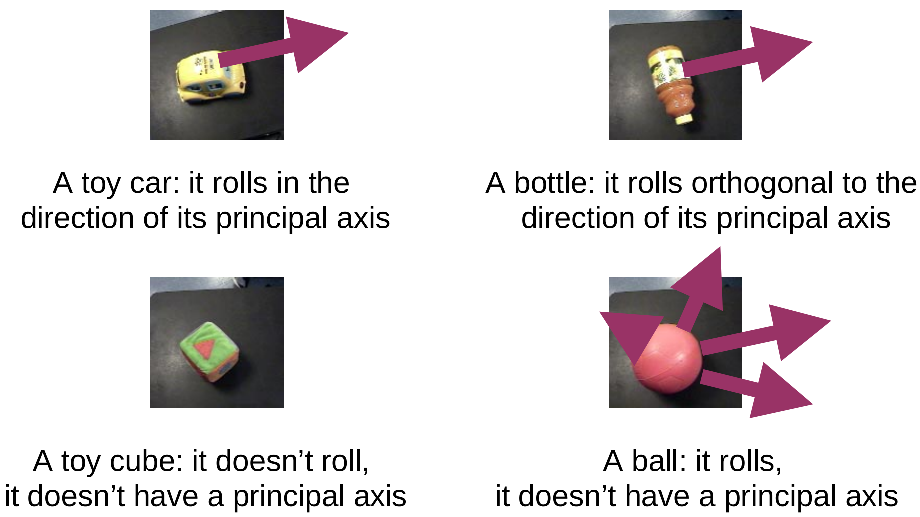


Figure 1: Different objects roll in different ways. A toy car rolls forward, a bottle rolls on its side, a sphere rolls in any direction, and a cube doesn't really roll at all.

The final behavior of the robot after training is: human presents an object, makes it roll. Presents the same object, perhaps in another orientation, and the robot pushes it in the right direction to make it roll. If the human hits the object in a non-canonical direction, so will the robot. This serves to demonstrate the full loop of perception and action.

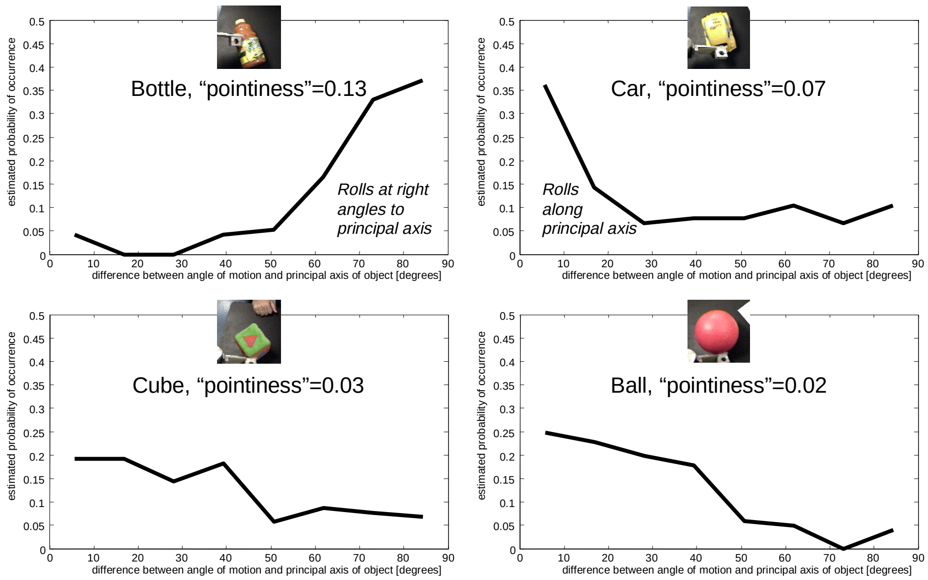


Figure 2: The robot relates the direction an object moves in with the direction of its principal axis. For a car, there is a clear tendency to roll in the direction of the principal axis. A bottle has a clear tendency to roll towards its side. Cars and spheres have no clear principal axis so, whether they roll or not, there is nothing to say.

2 Active Segmentation

Gives a sketch of the min-cut algorithm and how it is applied to this domain.

Use 16-connectivity (8 neighbors, plus cells connected by a “knight” move).

3 Detecting the point of collision

Two components – detecting the moment of impact, and extracting as much data as possible from the frames around it.

There are particular periods when the robot is attentive and fixating. When this is so, it can detect visually when a collision occurs between a moving object and a previously stationary object in view. The principal task, then, is keep the motion of the moving object distinct from that of the impacted object.

Brief introduction to image differencing and background subtraction. Stationary camera assumption to facilitate pixel modeling. Don't want to keep head stationary, but can fixate for significant periods.

Image differencing is a very simple technique for detecting motion by simply subtracting successive frames 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, computer monitors, movement of the camera itself, etc. A related technique called background modeling 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. While these techniques are not ideally suited to a moving platform like our robot, they are short periods during which they can be useful. In particular, when the robot is fixating a target, we can do this.

Within the context of the robot fixating a target, we try to detect the moment of impact precisely, so we can apply the (relatively) slow segmentation optimization to a narrow interval of the video input and maintain close to real-time performance. A moving manipulator colliding with an object will accelerate it, if it is not too massive. If the object is rigid, the motion of the manipulator will be transmitted through it. This transmission can be detected as a spreading motion that is not plausibly generated by the manipulator itself.

Some assumptions that may fail: object not too heavy; object at least semi-rigid; manipulator not moving above a certain speed; manipulator not casting shadows on the object itself. When the robot is poking the object itself, it can control some of this. If the object iself is troublesome, then we potentially diagnose this, or just ignore it.

4 Learning about manipulators

How could a robot find human arms and hands in the environment 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 [cite]. Another approach is possible in our situation. If the robot can detect when an impact event occurs, it can collect segmentations of the object that caused the impact. The set of objects that habitually trigger the motion of other objects is not a bad operational definition of a manipulator, and should include the human hand/arm, and the robot's own arm.

Unconstrained motion in a scene is difficult to parse. But since the robot has become familiar with a set of objects through poking, it can constrain the scenarios in which it may identify the manipulator. In particular, fixating a familiar object is a necessary condition for reliably detecting collision. Fixation is improved by object recognition and localization, trained my the previous poking episodes.

May fail for more complex events – e.g. other motion, shadows cast on object etc.

When the robot fixates an object that it can reach, it will try to poke it. This is inhibited if it sees motion around the object, giving a human the opportunity to poke it instead.

The point of contact detection will still work in this case, since it doesn't rely on the manipulator being the robot's own arm.

The frames leading up to the impact are great for detecting the appearance of the manipulator itself. And give a strong cue that the moving object is in fact a manipulator (something used to impact objects). See Figure 7.

Can't necessarily remove shadow.

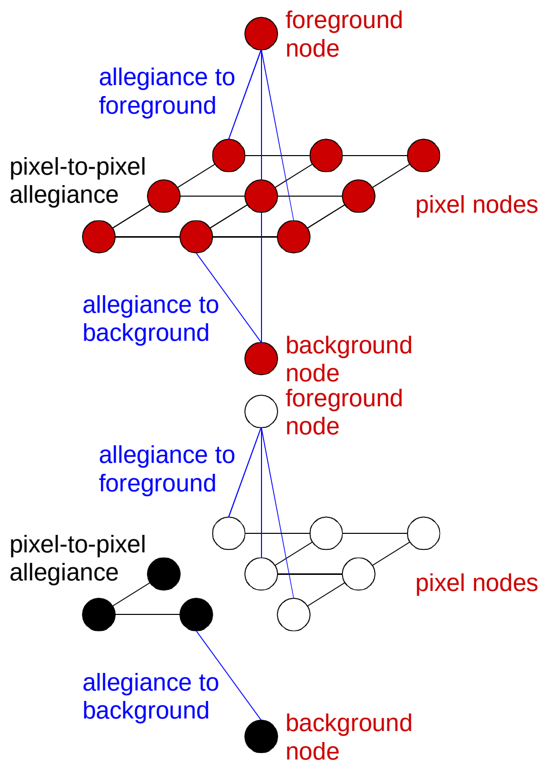


Figure 3: A simple example of the graph cut algorithm in operation. The left graph represents the output of the point-of-contact processing. Edges in the graphs are weighted by how much it will cost to split connected nodes. The bulk of the nodes are in one-to-one correspondence with pixels in the image. There are two extra nodes corresponding to the foreground and background. The goal is to split the graph into two by removing edges. The cost of the split is the sum of the weights on the edges removed. There are good approximate algorithms for finding a minimum cost solution [ref].

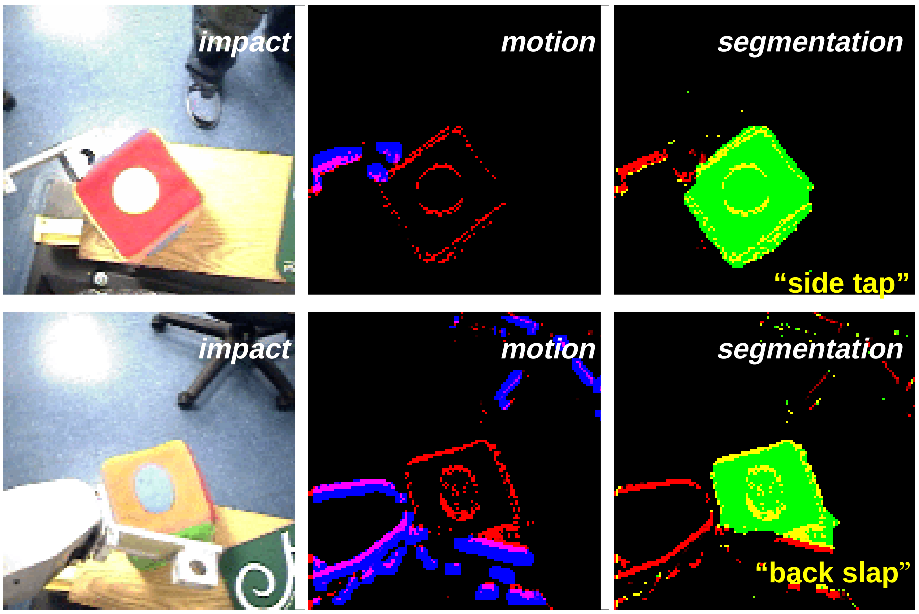


Figure 4: Cog batting a cube around. The top two rows show the flipper poking the object repeatedly from the side, turning it slightly. The third row shows Cog batting an object away. The images in the first column are frames prior to a collision. The second column shows the actual impact. The third 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.

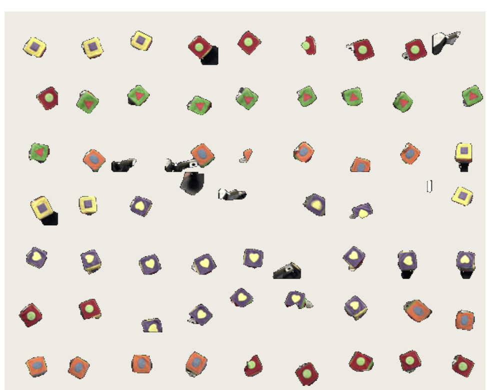


Figure 5: Sample results

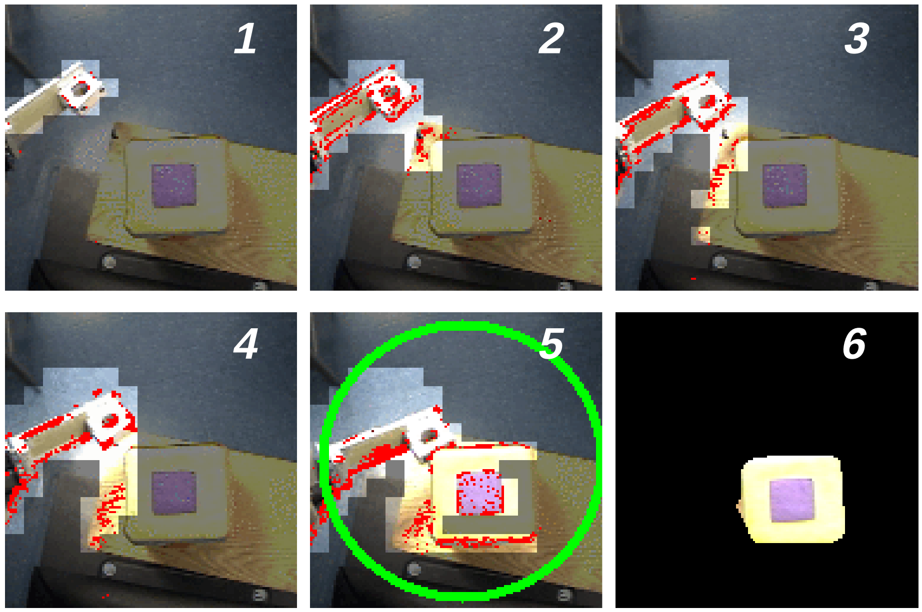


Figure 6: The moment of impact is detected visually by the sudden expansion of motion away from the arm. Motion before and after contact is compared to gather information for segmentation.

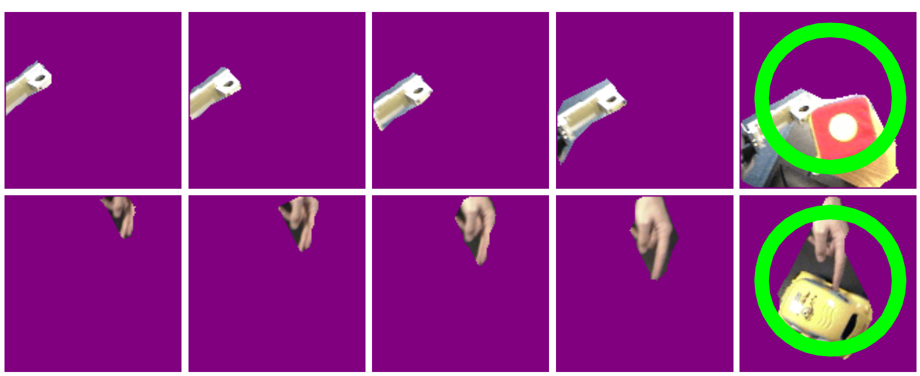


Figure 7: Early experiments on segmenting the robot arm, or a human hand poking an object the robot is familiar with, by working backwards from a collision event.

Abstract

For the purposes of manipulation, we would like to know what parts of the environment are physically coherent ensembles – that is, which parts will move together, and which are more or less independent. It takes a great deal of experience before this judgement can be made from purely visual information. This paper develops active strategies for acquiring that experience through experimental manipulation, using tight correlations between arm motion and optic flow to detect both the arm itself and the boundaries of objects with which it comes into contact. 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.

5 Introduction

Robots and animals are actors in their environment and not simply passive observers. This gives them the potential to examine the 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 system, such as a humanoid robot.

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 many important steps along the way, and describe how they fit in a larger scale implementation. Also, we discuss in what sense our artificial implementation is substantially in agreement with neuroscience.

Table 1 shows three levels of causal complexity. The simplest causal chain that the actor experiences 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 active 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.

Finally we argue that extending this causal chain further will allow the actor to make a connection between her own actions and the actions of another. This is reminiscent of what has been observed in the responses of the monkey's premotor cortex.

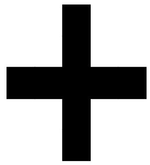
type	*nature of causation*	*time profile*
sensorimotor coordination	direct causal chain	strict synchrony
object probing	one level of indirection	fast onset upon contact, potential for delayed effects
mirror representation	complex causation involving multiple causal chains	arbitrarily delayed onset and 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.

6 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 phenomenon as figure-ground segmentation? Consider the example in Figure 8. It is immediately clear that the drawing on the left is a cross, perhaps because we already have a criterion, which allows segmenting on the basis of the intensity difference. 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, and we do not have the criterion to perform the figure-ground segmentation, we might think this is just 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.

While we have to be inventive to come up with a segmentation problem that tests a human, we don't have to go far at all to find something that baffles our robots. Figure 9 shows a robot's-eye view of a cube sitting on a table. Simple enough, but many rules of thumb used in segmentation fail in this particular case. And even an experienced human observer, diagnosing the cube as a separate object based on its shadow and subtle differences in the surface texture of the cube and table, could in fact be mistaken – perhaps some malicious researcher is up to mischief. The only way to find out for sure is to take action, and start poking and prodding. As early as 1734, Berkeley observed that:



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

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

a cross *a binary cross* *?*

Figure 8: Three examples of crosses, following [Manzotti and Tagliasco, 2001]. The human ability to segment objects is not general-purpose, and improves with experience.

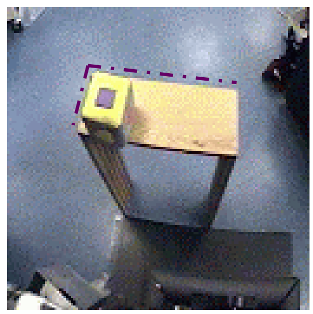


Figure 9: A cube on a table. The edges of the table and cube happen to be aligned (dashed line), the colors of the cube and table are not well separated, and the cube has a potentially confusing surface pattern.

> ...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 in the presence of [touch] physical contact, vision becomes more powerful, and many of its illusions fade away.

Objects and actions

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 results in neural science. 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.

Key to understanding how such capabilities could develop is the well-known result by Ungerleider and Mishkin [Ungerleider and 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 and 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 is emphasized by many commentators, it is significant that there is 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.

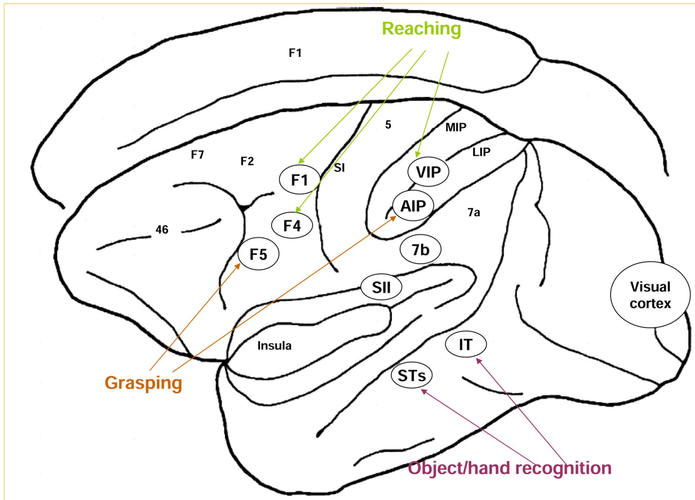


Figure 10: Monkey brain with indication of the main areas participating to object oriented actions (adapted from [Fagg and Arbib, 1998]). As described in the text, three main functions can be indentified: object recognition, reaching, and grasping. These form three parallel yet connected streams of processing. The circuit connecting the visual cortex to the inferior parietal lobule (VIP), F4 and F1 subserves reaching. AIP and F5 are responsible for grasping. Temporal areas (infero temporal, IT) and STs are correlated to the semantic of object recognition. [It's likely we need to ask permission to use this figure].

Grossly simplifying (see also figure 10), the brain circuitry responsible for object oriented actions is thought to consist of at least four interacting regions: the primary motor cortex (F1), the premotor cortex (F4, F5), the inferior parietal lobule (AIP, LIP), and the temporal cortex (TE, TEO) ([Rizzolatti et al., 1997, Fadiga et al., 2000, Jeannerod, 1997] for a review). While this is a useful subdivision, it is worth to bear in mind that the connectivity of the brain is much more complex, that bidirectional connection are present, and that the 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.

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

Reaching responsive neurons are present in the inferior parietal lobule. For example, Jeannerod et al. [Jeannerod et al., 1995] reported that the temporary deactivation of the caudal part (VIP) - by injecting a GABA agonist - of the intraparietal sulcus 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 RF often overlap (e.g. a central visual RF corresponds to a tactile RF in the nose or mouth). The parietal cortex contains also cells related to eye position/movements that are likely to be involved into 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 [Graziano et al., 1997, Fogassi et al., 1996]. 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, for example, the forearm. The somatosensory RF is usually in register with the visual one. 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 et al., 2000] described neurons that keep memory of the position of objects for the purpose of reaching. They found neurons that change their firing rate after a brief illumination of an object in the reaching space. The neurons returned to their baseline firing rate only after showing the monkey that the object had been actually taken away or moved to a different position.

Sakata and coworkers [Sakata et al., 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 the object alone 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, yet others responded indifferently to the presentation of a broad class of objects. Area AIP is interesting because contains both motor and visual 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 just another big motor area. Recent studies (see [Jeannerod, 1997, Fadiga et al., 2000]) 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 target of the projection from AIP (to which it sends back recurrent connections), was thouroghly investigated by Rizzolatti and colleagues [Gallese et al., 1996].

F5 neurons can be classified in at least two different categories: canonical and mirror. Canonical and mirror neurons are indistinguishable one from another 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 [Fadiga et al., 2000], becomes active under 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 human, area F5 is thought to correspond to Broca's area: there is an intriguing link between gesture understanding, language, imitation, and mirror neurons [Rizzolatti and Arbib, 1998].

The superior temporal sulcus region (STs) and part of TE contain neurons that are similar in response to mirror neurons [Perrett et al., 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 to the processing of the visual information and then commnunicate to F5 [Gallese et al., 1996].

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. reaching, grasping, recognition); ii) the ability to probe longer chains triggers the emergence of a new functionality and/or a new set of behaviors. The next section delves a bit more into the description of the ontogenesis of object oriented action and provides a hypothesis amenable to implementation.

A working hypothesis

Taken together these results from neuroscience suggest a very basic role for motor action. 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 to object recognition in cases where an agent has to develop categorization autonomously. Of course in standard supervised learning action is not required since the trainer does 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.

If our hypothesis is correct then the first developmental step has to be that of transporting the hand close to the object. This function is accomplished mostly by the circuit VIP-F4-F1. Reaching requires at least the detection of the object and the transformation of its position 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 et al., 2002] and to [Flanders et al., 1999] the gaze direction (the eye motor plant) seems to be the priviledged reference system used to code reaching. The signals required for calibrating/learning reaching can be perhaps acquired completely by a random exploration of the workspace [Metta et al., 1999, Marjanović et al., 1996].

Once reaching is realiable enough, the attention can be concentrated 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 abilities (maybe poking and prodding initially). Arbib and coworkers [Fagg and Arbib, 1998] proposed the FARS model as a possible description of the computation in AIP/F5. They did not consider however how affordances can be actually learned during the interaction with the environment.

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 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 conditions why this is feasible are a consequence of active manipulation. During a manipulative act there are a number of additional costraints that can be factored in to simplify perception/computation. For example, detection of useful events is simpler at least because of touch, because it is known when the reaching started, and because the location of the object is known in the first place.

The last subsystem to develop is object recognition. Object recognition can build on manipulation in finding the boundaries of objects and segment 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. This functions are carried out by the infero-temporal cortex (IT). The same considerations apply to the recognition of the manipulator (either self or foreign). In fact, STs specialized to this task. Information about object identity is also sent to the parietal cortex and contributes to the formation of the affordances.

nature of causation	*main path*	*function and/or behavior*
direct causal chain	VC-VIP-F4-F1	reaching
one level of indirection	VC-AIP-F5-F1	poking, prodding, 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	object recognition

Table 2: Degrees of causal indirection vs. function in the brain.

7 The experimental platform

This work is implemented on the robot Cog, an upper torso humanoid [Brooks et al., 1999]. The robot has previously been applied to tasks such as visually-guided pointing [Marjanović et al., 1996], and rhythmic operations such as turning a crank or driving a slinky [Williamson, 1998]. Cog has two arms, each of which has six degrees of freedom – two per shoulder, elbow, and wrist. The joints are driven by series elastic actuators [Williamson, 1995] – essentially a motor connected to its load via a spring (think strong and torsional rather than loosely coiled). 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 characterized environment, where collisions are frequent and informative events.

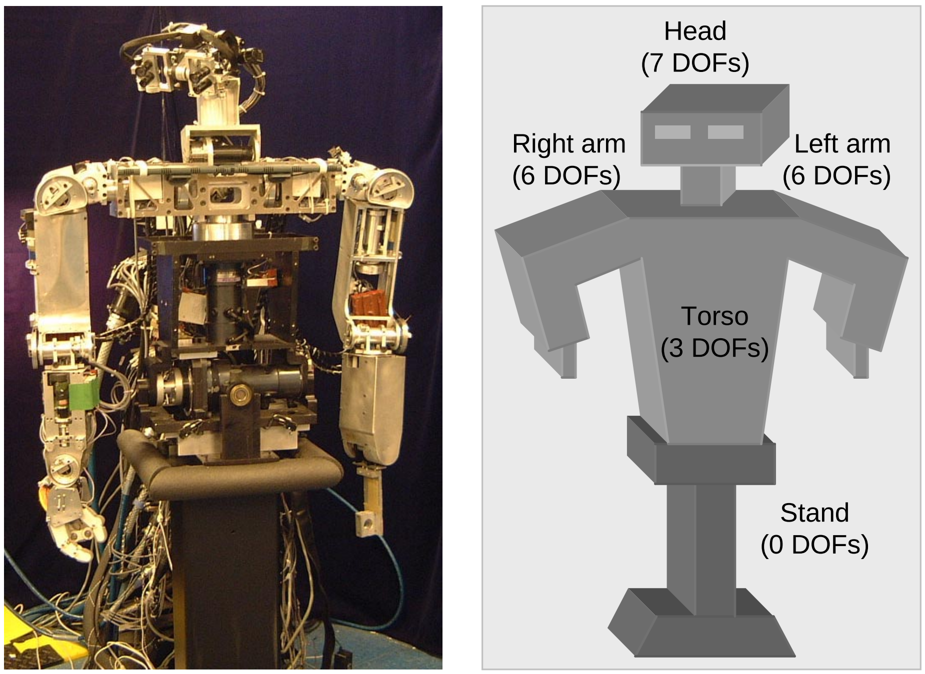


Figure 11: Degrees of freedom (DOFs) of the robot Cog. The arms terminate either in a primitive "flipper" or a four-fingered hand. The head, torso, and arms together contain 22 degrees of freedom.

8 Perceiving direct effects of 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 extremely complex, the correlation in time of the two events will generally 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 localize the arm in the image without any prior information about visual appearance. In the next section we will show that once the arm has been localized we can go further, and identify the boundaries of objects with which the arm comes into contact.

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 the 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 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-

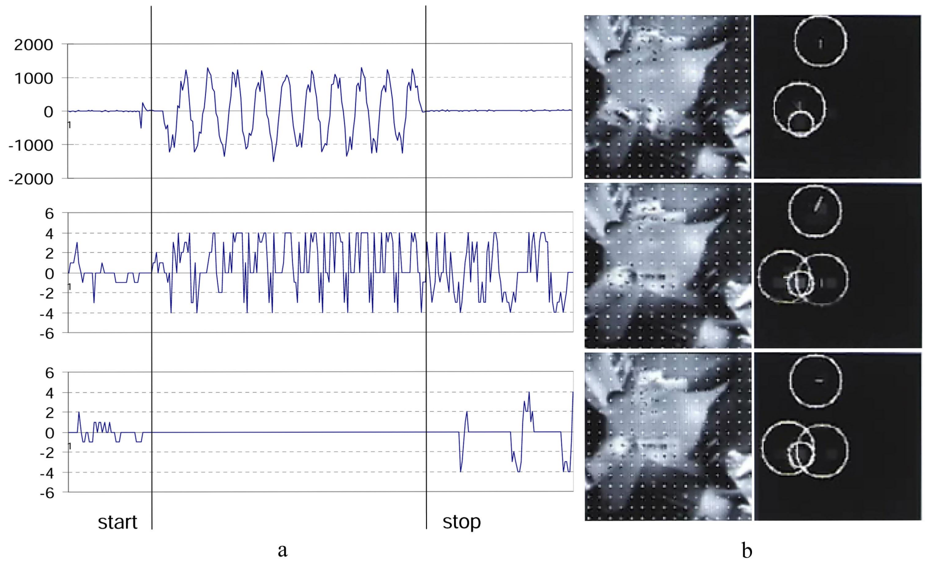


Figure 12: (a) An example of the correlation between optic flow and arm movement. The traces show the movement of the wrist joint (upper plot) and optic flow sampled on the arm (middle plot) and away from it (lower plot). (b) The robot's point of view and the optic flow generated are shown on the left. On the right are the results of correlation. Large circles represent the results of applying a region growing procedure to the optic flow. The small circle marks the point of maximum correlation, identifying the regions that correspond to the robot's own arm.

loop control, and to training up a feed-forward path. We shall demonstrate that this localization can be performed without knowledge of the arm's appearance, and without assuming that the arm is the only moving object in the scene.

Localizing 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 where 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. This probability could also be reduced by higher resolution (particularly in time) visual processing.

Figure 12 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 12 shows examples of detection of the arm and rejection of a distractor.

Localizing the arm using proprioception

The localization 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 localization during actual operation.

As a function approximator we simply fill a look-up table, implemented as a list of nodes allocated dynamically. This implementation was chosen to reduce memory consumption; the input space is six dimensional and even a coarse discretization of this space would require memory in the order of several Mbytes. Rather than using all the joint angles the current direction of gaze is first coded in terms of only two angles representing the global pan (θ) and tilt (ϕ) of one of the cameras. This is easily computed from the kinematics of the head and the joint angles. The end-point position is coded considering only the first four joints ($q_1 \dots q_4$). The position of joint q_5 and q_6 is not employed because the wrist does not significantly contribute to the end-point position. The output of the approximator is the position of the end-point (the forearm) on the image plane. Figure 13 shows the resulting behavior after about twenty minutes of real-time learning.

Reaching for the object

Reaching is implemented as a direct mapping between the direction of gaze (θ, ϕ) and the command required to reach the fixation point. This procedure is consistent because we are interested in reaching a point on a plane in front of the robot (a table). The resulting map is thus 2D $\rightarrow$ 2D. The same argument could be extended to the 3D case by augmenting the encoding of gaze with, for example, the vergence angle. The arm motor commands are represented in terms of joint positions, and the mapping is linear:

$$\begin{pmatrix} \hat{q}_1 \\ \vdots \\ \hat{q}_6 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ \vdots & \vdots \\ a_{61} & a_{62} \end{pmatrix} \cdot \begin{pmatrix} \theta \\ \phi \end{pmatrix} \tag{1}$$

where $\hat{q}$ are the desired joint positions. The coefficient a_{nm} are estimated following a brief calibration procedure from a small number of training pairs of the form $(\hat{\mathbf{q}}, (\theta, \phi))$. The linear approximation is justified in our case because of the relatively small region of the workspace where the reaching is expected to operate. The complete robot workspace is much bigger because the torso can also move to keep the operational point of the linear approximation within reasonable limits.

At a lower level a low-stiffness position control and a simple trajectory generator interpolate the motion of the arm from the current position to the commanded one. Gravity compensation for the shoulder joint has been implemented to further improve accuracy.

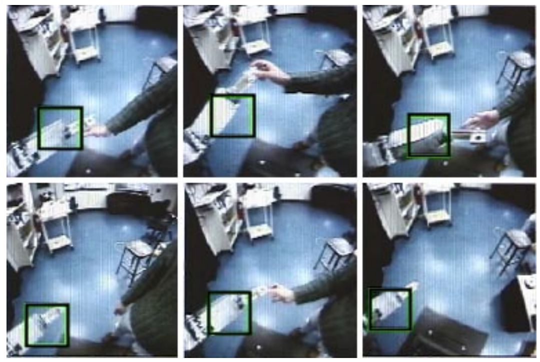


Figure 13: Predicting the location of the arm in the image as the head and arm change position. The rectangle represents the predicted position of the arm using the map learned during a twenty-minute training run. The predicted position just needs to be sufficiently accurate to initialize a visual search for the exact position of the end-effector.

9 Perceiving indirect effects of action

We have assumed that the target of a reaching operation is chosen visually. As discussed in the introduction, visual segmentation is not easy, so we should not expect a target selected in this way to be a correctly segmented. For the example scene in Figure 9 (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 develop a procedure for refining the segmentation using the same idea of correlated motion used earlier to detect the arm.

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 can

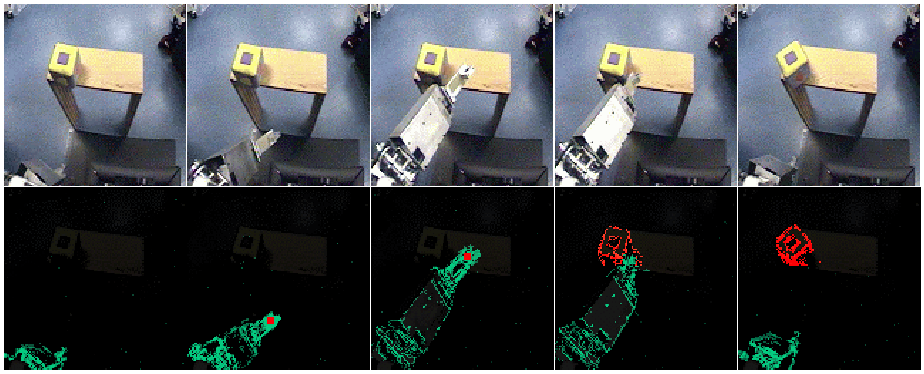


Figure 14: The upper sequence shows an arm extending into a workspace, tapping an object, and retracting. This is an exploratory mechanism for finding the boundaries of objects, and essentially requires the arm to collide with objects under normal operation, rather than as an occasional accident. The lower sequence shows the shape identified from the tap using simple image differencing and flipper tracking.

simply be ignored, since the robot will not be able to manipulate them. But if the object is smaller, it is likely to move a little 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 coincides with the physical boundaries of the object.

Figure 14 shows how a "poking" movement can be used to refine a target. During a poke operation, the arm begins by extending outwards from the resting position. The end-effector (or "flipper") is localized 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 neighborhood 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 minimizes 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.

For simplicity, the head is kept steady throughout the poking operation, so that simple image differencing can be used to detect motion at a higher resolution than optic flow. Because a poking operation currently always starts from the same location, the arm is localized using a simple heuristic rather than the procedure described in the previous section – the first region of optic flow appearing in the lower part of the robot's view when the reach begins is assumed to be the arm.

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

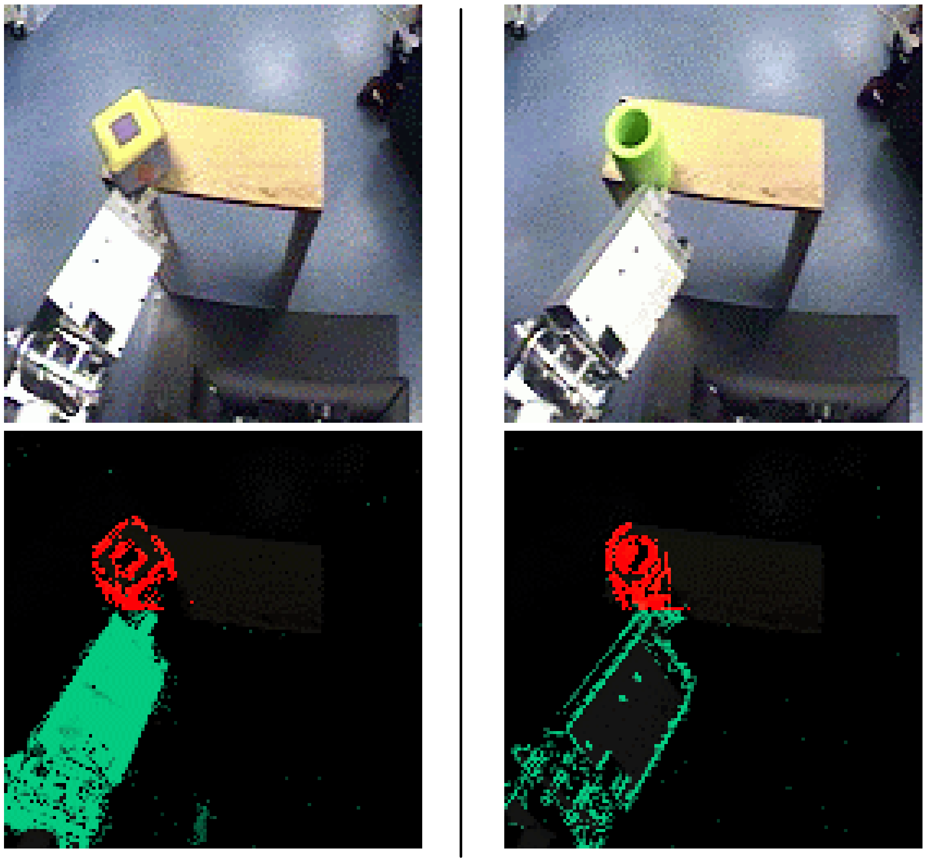


Figure 15: Poking can reveal a diffence in the shape of two objects without any prior knowledge of their appearance.

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.

10 Experimenting with object affordances

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.

Segmentation alone is still unconvenient in many situations if not coupled with a mechanism to learn from experience. For example, it would be terribly unefficient to poke the object first and then try to grasp it. It would be much better if the robot could learn about objects and [if it] had a way to identify a previously encountered object. A further difficulty, at least for a robot with a simple manipulator (e.g. as COG's flipper), is that "affordances" are scarce: most of the time the object simply move from one position to another if we are willing to discount when it falls from the table.

However, for objects that roll there is a cue the robot can exploit to understand their behavior. An object that rolls tends to do so even if it is not poked precisely. We selected a small set of objects to experiment with: a cube, a toy car, an orange juice bottle, and a ball. Affordances are not only a property of the mechanics of the object, but rather a combination of visual appearance, of the object's physical constituent, and of the ability of the actor. We selected a measure of the principal axis of the object (easily obtained from the segmentation) as visual component of the affordance. Table 3 shows the expected behavior:

A further elaboration is required to group the data belonging to the same

object	*angle between principal axis and preferred direction of rolling*	*behavior*
cube	n.a.	no principal axis, does not roll
car	$0°$	rolls along the principal axis
bottle	$90°$	rolls at right angle
ball	n.a.	no principal axis, does roll

Table 3: Behavior of a small set of objects when poked at random by the robot manipulator.

object as obtained from many poking acts into coherent clusters. As clustering mechanism we employed color histograms. After each poking action, a color histogram of the pixels of the segmented region is built and used as criterion to judge whether the object belongs to an existing group (e.g. if it is mostly yellow, it is likely to be the toy car). This works well for a small set of objects but sophisticated methods are required for a more general case with a large set of objects. The data structure that simulates the AIP-F5 affordance computation maintains all the instances of poking grouped by object, all the prototypes of the segmented object, the direction of movement, and the action applied by the robot in each trial.

An alternative to the vision-based clustering procedure is to try to come to grips with the behavior of the objects after a single encounter, and use the behavior itself as clustering criterion. This is more difficult because of noise: e.g. there is still a non-zero probability that the object would not roll at all.

Figure 16 shows the results of the segmentation, clustering and estimation of the affordance of the same set of four objects. The training set consists of about 100 actions per object. The motor vocabulary of the robot consists of four possible directions of poking. We labeled them for convenience as: pull in, side tap, push away, and back slap, depending on the effect they have on the object from the point of view of the robot. Actions were generated at random during this training stage. During a poking action, the object is tracked for 12 frames after the time of contact and the overall displacement is computed.

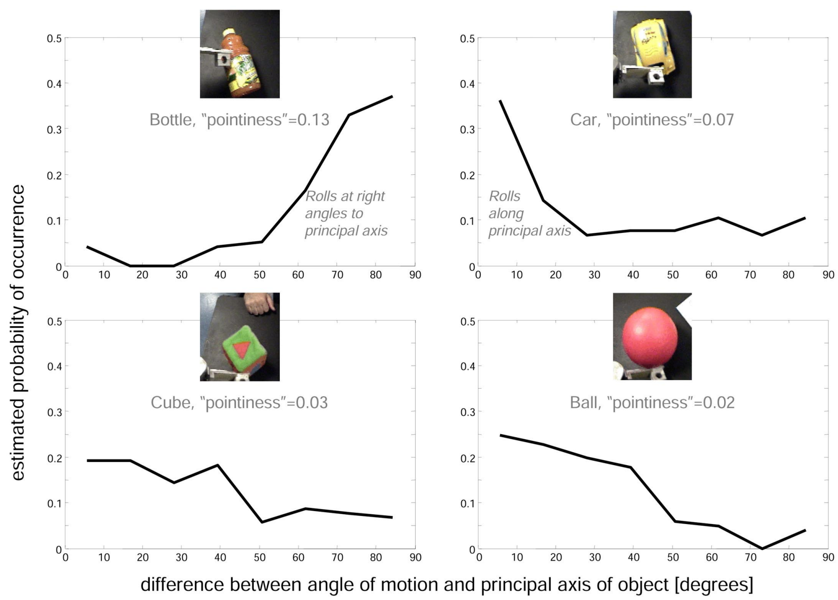


Figure 16: Probability of observing a roll along a particular direction for the set of four objects used in our experiments. Abscissae represent the difference between the principal axis of the object and the observed direction of movement. Ordinates the estimated probability.

Yet this description of the affordances does not have any usable quantity to take action once an object is observed. For this purpose a description of the geometry of poking is required: i.e. the description of the properties of objects (figure 16) has to be connected to a description of the behavior of the object. This information can be derived from the same training set we collected for learning about rolling. Figure 17 shows the histograms of the direction of movement of the object for each possible action. For example, the back slap moves the object mostly upward (about $-100°$ on average) and away from the robot. A similar consideration applies to the other poking gestures. Figure 17 was obtained from the data of about 500 poking events.

The last step is to connect all these elements together. If a known object is presented to COG, the object is recognized, localized, and its orientation estimated (principal axis). Recognition is based on the color histograms. The same procedure used to form the clusters is employed here. Localization is simply implemented by histogram backprojection and a search across the image. The current orientation of the object is then estimated by comparing the current image with all the prototypes contained in the cluster. The whole procedure has an error on the estimation of the principal axis in the range $10°$ to $25°$ depending on the object.

To actually exploit the understanding of the affordance we need to connect vision to behavior. The robot looks for the preferred rolling direction of the object (see figure 16) and adds it to its current orientation. The action whose effects are closer (on average) to the combination of the orientation and affordance is selected.

We performed a simple qualitative test of the robot's behavior presenting randomly two of the objects (the toy car and the bottle) - note that the ball and

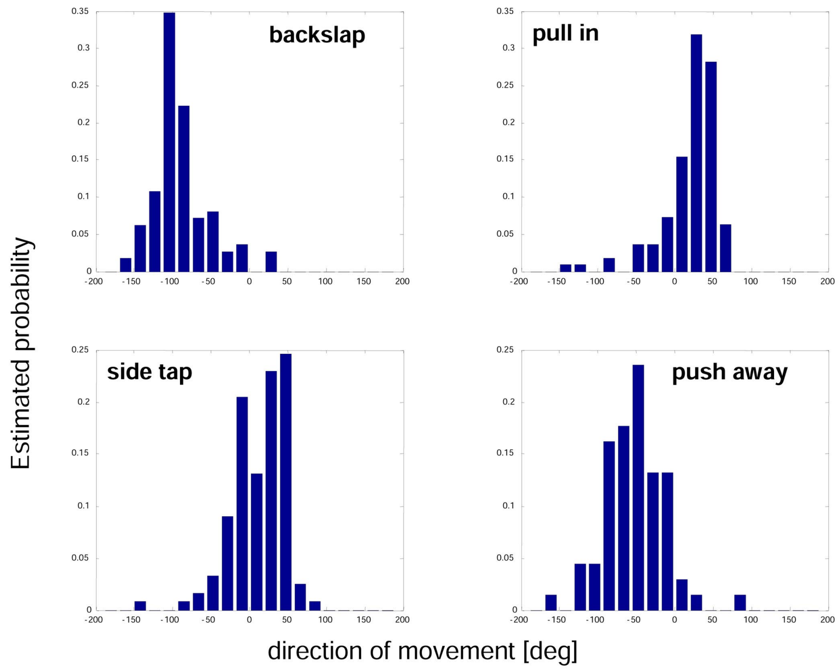


Figure 17: Histogram of the direction of movement of object for each possible poking action.

the cube do not have a well defined principal axis so there is no point in running the experiment. Out of 100 trials the robot made XY mistakes. Analysis of the errors reveals that they are mainly due to misinterpretation of the orientation of the object or to unprecise control.

11 Developing mirror neurons

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. The next experiment was designed to explore this aspect.

In fact, the same processing used for analyzing an active poking can be used to detect a contact and segment the object from the manipulator. This is not different from what used for learning. While one might argue then that learning can be carried out just by mere observation, it is worth noting that: i) this situation is not as well defined [circumscribed] as the active one, and ii) there is no connection to the motor aspects of the action and consequently it is difficult to link the observation to the behavior. There is no physical contact, thus there is plenty of room for getting confused by false positives. The temporal aspect, so well constrained during active manipulation, is more vague here - the robot, for example, does not know when the foreign manipulator starts or stops the action. If missing a contact event or getting a false or mistaken segmentation is not much of a problem in "observation mode", it is much more troublesome is we corrupt the training data with unreliable/noisy observations. Further, we should not assume the human "teacher" is truly collaborative. There is no guarantee that actions suited to the robot perceptual system and/or goal are performed at all. More seriously, the link to behavior is completely missing. Even if visual information about objects can be collected as before, tracing back which action causes a particular consequence cannot be autonomously learnt by the robot. Conversely, in the case the robot has already learned about objects, as e.g. we have shown in the previous section, this information can be factored in to help the observation of somebody else's action. Touch (not in COG) and physical contact are additional bits of information about the ongoing activity.

In our case, if any activity is detected close to the object - measured by the amount of motion in a neighborhood of the fixation point corresponding to the robot's foveal camera - reaching is inhibited and the whole action observed (assuming there is one at all). An example of human poking is shown in figure 18.

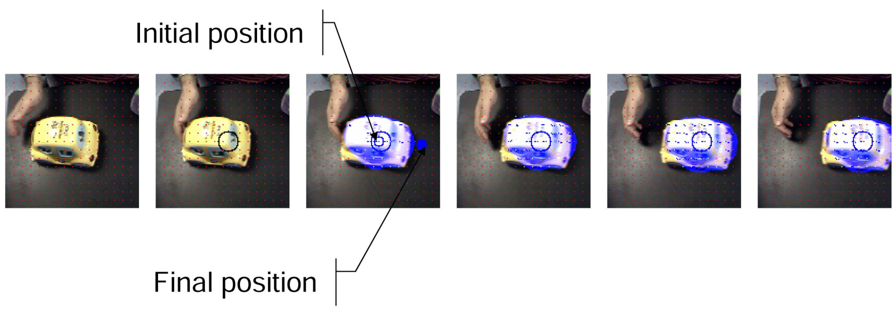


Figure 18: An example of observed sequence with tracking superimposed. Frames around the moment of contact are shown. The object, after segmentation, is tracked for 12 frames using a combination of template matching and optic flow. The big circles represent the position of the toy car in successive frames. The two small circles (outline and solid) displayed on the frame of contact (3^{rd} from the left) are the position at the time of contact and at the 12^{th} respectively.

The first obvious thing the robot can do is to identify the action just observed with respect to its motor vocabulary. It is easily done, in this case, 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. Indeed it works well and it allows - even if in this limited setting - recognizing a complex action by interpreting its consequences on the environment. This is orders of magnitude simpler than trying to completely characterize the action in terms of the observed kinematics of the movement. Here, the complexity of the data we need to obtain from the observations is somehow proportional to the complexity of the goal rather than that of the structure/skills of the foreign manipulator. In our case, because the action, the goal, and the object are relatively simple, the only information required is about the displacement of the object.

Therefore, the next question is whether we can use this "understanding" of observed actions to implement mimicry behavior. It would be easy now to try to replicate the action just observed if the same object were presented again. However, there is still a bit of ambiguity in that we can choose to mimick either the observed displacement of the object or the way the object was poked with respect to its rolling affordance.

We chose to implement the latter. It is clear that poking along a particular observed direction requires trivial modifications. In practice, after an action is observed the angle between the affordance (see table 3) and the actual displacement is measured and stored. If it happens to see the same object again, the robot chooses the action that has the greatest probability of poking the object along the previously stored angle. Figure 19 shows to examples of mimicry as a consequence of the action shown in figure 18.

This response is exactly what we would expect from a "mirror-type" rep-

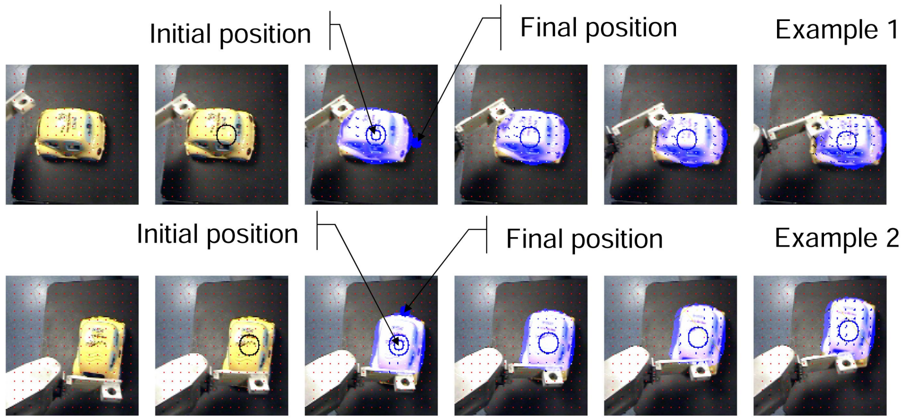


Figure 19: Two examples of mimicry following the observation in figure 18 where a human manipulator pokes the toy car exploiting the affordance (the car rolls). In example 1 (top row), the toy car has the same orientation it had in the demonstrated action and the robot repeats the observed action. In example 2 (bottom), the car is 90° with respect to example 1. The appropriate action to exploit the affordance and make the car roll is thus a back slap.)

resentation. The observed action is interpreted on the basis of the robot own motor code. The same data structure is also used/activated when performing an action in response to the sight of a known object. The causal link between the two events that could be separated by several seconds is the object, the goal, and the object's affordances. There is considerable precedent in the literature for a strong connection between viewing object manipulation performed by either oneself or another [Wohlscläger and Bekkering, 2002]. There is also a growing evidence that imitation is goal-directed [Bekkering and Wohlschlager, 2000] and that the object of the action is explicitly coded (e.g. during reaching) [Woodward, 1998].

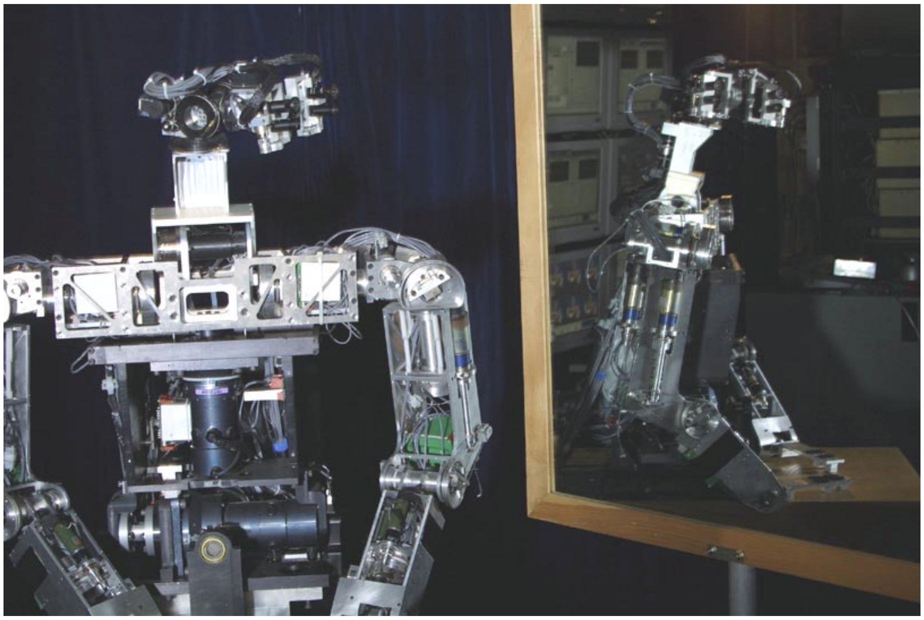


Figure 20: The ultimate goal of this work is for our robot to follow chains of causation outwards from its own simple body into the complex world.

12 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.

In computer vision there is much to be gained by bringing a manipulator into the equation. Many variants and extensions to the experimental "poking" strategy explored here are possible. For example, a robot might try to move an arm around *behind* the object. As the arm moves behind the object, it reveals its occluding boundary. This is a precursor to visually extracting shape information while actually manipulating an object, which is more complex since the object is also being moved and partially occluded by the manipulator. Another possible strategy that could be adopted as a last resort for a confusing object might be to simply hit it firmly, in the hopes of moving it some distance and potentially overcoming local, accidental visual ambiguity. Obviously this strategy cannot always be used! But there is plenty of room to be creative here.

Acknowledgements

This work benefited from discussions with Charles Kemp and Giulio Sandini. Many people have contributed to developing the Cog platform [Brooks et al., 1999]. Funds for this project were provided by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" project under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement.

References

[Bekkering and Wohlschlager, 2000] Bekkering, H. and Wohlschlager, A. (2000). Imitation in children is goal-directed. *The quarterly journal of experimental psychology*, 53A(1):153–164.

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

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

[Brooks et al., 1999] 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 et al., 2000] 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 and Arbib, 1998] Fagg, A. H. and Arbib, M. A. (1998). Modeling parietal-premotor interaction in primate control of grasping. *Neural Networks*, 11(7–8):1277–1303.

[Flanders et al., 1999] Flanders, M., Daghestani, L., and Berthoz, A. (1999). Reaching beyond reach. *Experimental Brain Research*, 126(1):19–30.

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

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

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

[Graziano et al., 2000] Graziano, M. S. A., Cooke, D. F., and Taylor, C. S. R. (2000). Coding the location of the arm by sight. *Science*, 290(December):1782–1786.

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

[Jeannerod et al., 1995] Jeannerod, M., Arbib, M. A., Rizzolatti, G., and Sakata, H. (1995). Grasping objects: the cortical mechanisms of visuomotor tranformation. *Trends in Neurosciences*, 18(7):314–320.

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

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

[Marjanović et al., 1996] 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.

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

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

[Mussa-Ivaldi and Giszter, 1992] 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.

[Perrett et al., 1990] Perrett, D. I., Mistlin, A. J., Harries, M. H., and Chitty, A. J. (1990). Understanding the visual appearance and consequence of hand action. In *Vision and action: the control of grasping*, pages 163–180. Ablex, Norwood, NJ.

[Pouget et al., 2002] Pouget, A., Ducom, J.-C., Torri, J., and Bavelier, D. (2002). Multisensory spatial representation in eye-centered coordinates for reaching. *Cognition*, 83:B1–B11.

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

[Rizzolatti et al., 1997] Rizzolatti, G., Fogassi, L., and Gallese, V. (1997). Parietal cortex: from sight to action. *Current Opinion Neurobiology*, 7(4):562–567.

[Sakata et al., 1997] Sakata, H., Kusunoki, M., Taira, M., Murata, M., and 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 and Mishkin, 1982] Ungerleider, L. G. and Mishkin, M. (1982). Two cortical visual systems. In *Analysis of visual behavior*, pages 549–586. MIT Press, Cambridge, Massachusetts.

[Williamson, 1995] Williamson, M. (1995). Series elastic actuators. Master's thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.

[Williamson, 1998] Williamson, M. (1998). Neural control of rhythmic arm movements. *Neural Networks*, 11(7-8):1379–1394.

[Wohlscläger and Bekkering, 2002] 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, 1998] Woodward, A. L. (1998). Infant selectively encode the goal object of an actor's reach. *Cognition*, 69:1–34.

13 Extras

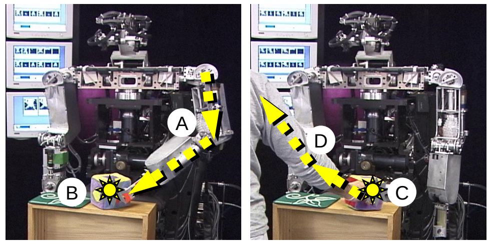


Figure 21: foo