CHAPTER 6

Reaching out: discovering one's own (and other) manipulators

> *He'd never bitten a hand that fed him. After all, this made it so much harder for the hand to feed you tomorrow.* (Pratchett, 1999)

In a sense, poking provides the robot with an operational definition of what objects are by giving it an effective procedure for learning about them. It is not perfect – for example, the robot is effectively blind to objects that are too small or too large – but for objects at an appropriate scale for manipulation, it works well. Once the robot is familiar with a set of such objects, we can go further and provide an operational definition of a *manipulator* as something that acts upon these objects. This chapter develops and effective procedure for grounding this definition.

To get training images of the manipulator, we need to find an opportunity when we can both segment it from the background and be sure that the segmented region is in fact the manipulator. Without constraining the environment, or having a prior training period in which a hand-eye mapping was trained, this is quite hard to do. However, there is one ideal opportunity – the moments before a poking event. This is a fairly narrow time window, when the robot is fixating and actively trying to move the arm into view. There is also an independent measure of whether the arm is in fact in view and whether everything is proceeding smoothly, which is the contact detection algorithm. Together, these can identify a short period of time in which the manipulator is very likely to be visible and moving across the field of view.

6.1 Hand to eye coordination

A robot has privileged knowledge of the state of its arms. Hence it can in principle predict from proprioceptive feedback where the arms will appear in the images from its cameras. Learning the mapping from joint angles to retinotopic coordinates is a favorite task in robotics (Fitzpatrick and Metta, 2002; Marjanović et al., 1996; Metta et al., 1999). Detection of the endpoint of the manipulator during training is made trivial by either giving it a special color, or by shaking it repeatedly. Once the mapping is learned, this simplification is no longer necessary, since the mapping does not depend on visual appearance. But what if we did want to learn a mapping that depends on appearance? For example, it would be useful if the robot could independently estimate the location of the arm from visual evidence rather than motor feedback, so that it could do precise closed-loop visually-guided

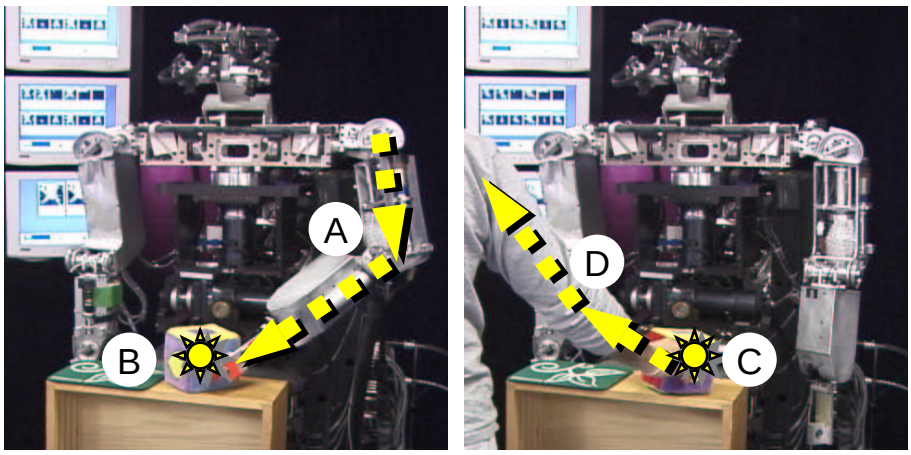

Figure 6-1: On the left, the robot establishes a causal connection between commanded motion and its own manipulator (A), and then probes its manipulator's effect on an object (B). The object then serves as a literal "point of contact" (C) to link robot manipulation with human manipulation (on the right, D).

control, rather than just 'blind' open-loop reaching. Now if we make the endpoint obvious using color or repeated motion in order to detect the manipulator, we must be careful that we can actually move away from that constraint after training.

The solution adopted is to identify a special situation in which the robot's arm can be identified in the visual field under its normal behavior. Consider the basic poking behavior introduced in Chapter 3. The visual collision detection mechanism developed in that chapter operates without any appearance model of the arm. When it does detect a collision near the point of fixation towards which the arm is being driven, that collision is very likely to be between the arm and an object. In Chapter 3 the motion caused by this collision was used to segment the object, with any motion regions that appeared to originate before the collision being discarded as being due to the arm, its shadow, or background motion. We can turn this reasoning around and try to segment the object that was moving before the collision. These segmentations are likely to contain the arm. As was shown in Chapter 5, if sufficiently good segmentations can be collected, then they can be refined – they don't need to be perfect.

If behaviors like poking are possible using open-loop reaching, is there any real reason to develop visually-guided reaching? As will be seen Chapter 7, although open-loop poking is fine for segmentation purposes, it leaves a lot to be desired when actually trying to move an object in a controlled way.

6.2 Objects as intermediaries

There is another motivation behind the choice of the poking behavior as the vehicle for detecting the manipulator. Clearly contact with objects is not the only possible way to locate the robot's arm, if we are willing to construct some special training behavior, such as a simple analogue of a human infant's hand-regard behavior. But it does have the advantage that objects can also serve as a point of contact between robot and human (Kozima et al., 2002). This thesis has shown that a robot can use its manipulator to familiarize itself with objects by poking them; if the robot then observes those known objects behaving as if they are being poked – but without it actually poking them itself – then

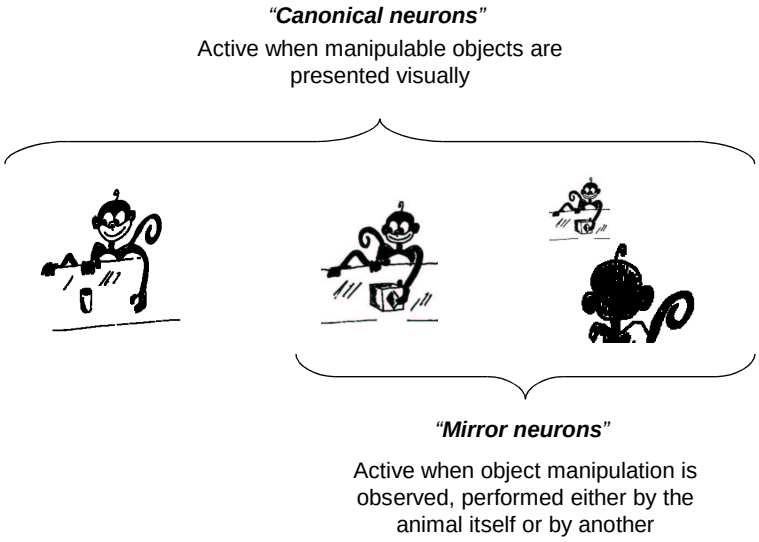

Figure 6-2: Canonical neurons are active when a manipulable object is observed, but this is not sufficient to activate mirror neurons. Mirror neurons are active only when goal-directed manipulation of an object is observed. This manipulation may be performed either by the animal itself, or by another (Gallese et al., 1996). These classes of neuron have been observed in area F5 of both monkeys and humans. (Monkey drawings by Giorgio Metta).

it can reasonably deduce that they are being acted on by some entity like itself (see Figure 6-1). For this to work, the robot must be able to perceive poking carried out by another. This is not easy to do in general, since the robot's gaze is unlikely to be directed at the appropriate location. But once it is familiar with an object, it can maintain fixation on it for a long period of time. Or if it sees a reachable object, familiar or unfamiliar, and begins to poke it, it will also fixate. It is simple to add a behavior that suppresses poking if unexpected motion is detected around the object (not due to the robot's own movement). This means that a human, noticing that the robot is preparing to poke an object, can take over and poke it themselves. Then the same machinery developed for active segmentation to operate when a foreign manipulator (such as the human hand) pokes the fixated object. Of course the robot can easily distinguish segmentations produced using its own arm from that of others simply by checking whether it was commanding its arm to move towards the target at the time. In this way it can also build up a model of foreign manipulators belong to others in its environment.

6.3 Canonical and mirror neurons

Is there any strong reason to identify the representation of the robot's own arm with the arms of others from the very beginning? Much of vision research is motivated by hints from biology. For primates, many neurons are concerned with both vision and motor control. For example, Fogassi et al. (1996) and Graziano et al. (1997) observed neurons that have receptive fields in somatosensory, visual, and motor domains in area F4. Motor information appears to be used to keep the somatosensory and visual receptive fields anchored to a particular part of the body, such as the forearm, as the body moves. So-called 'canonical' neurons have been located in area F5 which respond when the host acts upon an object in a particular way (such as grasping it with a precision grip), or when it

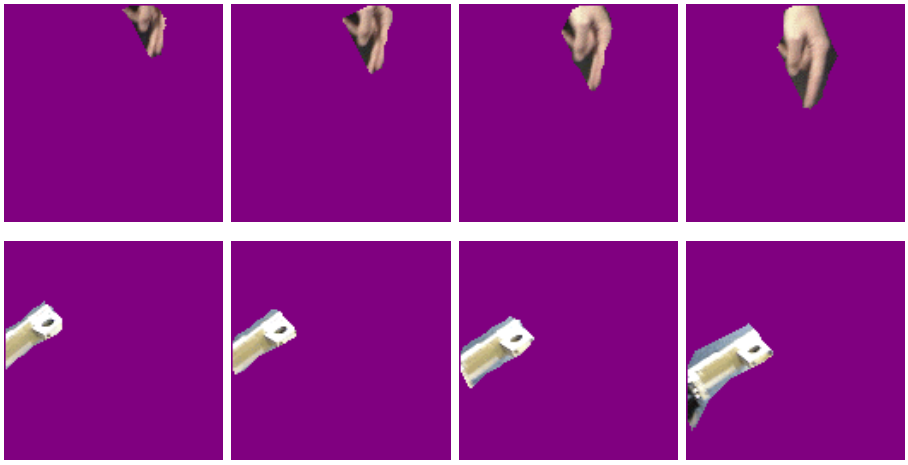

Figure 6-3: Experiments on segmenting the robot arm, or a human arm poking an object. The segmentation is performed by working backwards from the point of collision the object, which occurs in the frame immediately following the ones shown.

simply fixates the object (Jeannerod, 1997). These responses are very specific to the type of object and type of action, and so have been interpreted as being the neural analogue of the affordances of Gibson (Gibson, 1977). Affordances are discussed in detail in Chapter 7; briefly, they are simply possible actions that a particular actor can apply to a particular object. A related class of neurons in F5 called 'canonical' neurons have also been identified in primates (including humans) which respond when the host is performing a particular action (such as grasping) or when the host observes someone else performing that same action (Gallese et al., 1996). Again this response is very specific and hard to fool – if a neuron is selective for a particular type of grasp, it does not respond if a tool such as a pliers in used instead. These neurons have been interpreted as a possible basis for imitative behavior, and perhaps even language (Rizzolatti and Arbib, 1998). The existence of these neurons are a motivation for this work.

6.4 Implementation details

The manipulator can be segmented by hypothesizing that it moves towards the object at a constant velocity in the period immediately preceding the moment of contact. Estimating the velocity from the gross apparent motion allows the segmentation problem to be expressed in the form introduced in Chapter 3, where the foreground is now taken to be regions moving at the desired velocity, and the background is everything else. The apparent motion is computed by finding at each frame the translation that best aligns the differences between successive frames (rotation is ignored on this short timescale). Each pixel is assigned either to the stationery background or to a layer moving at the estimated rate. Typical results of this segmentation procedure are shown in Figure 6-3 for both the robot's arm and a human hand. Although it isn't relevant to the current discussion, segmentation of the object happens in the same way after human action as it does after robot action (Figure 6-4 shows an example).

Figure 6-4: The segmentation algorithm will work for human poking operations, if the robot is fixating the object the human pokes. The robot can be made to fixate objects by bringing its attention to it by any of the means discussed in Chapter 8.

6.5 Modeling the manipulator

As a first test to make sure the segmentation data was of usable quality, it was passed through the same alignment and averaging procedure described for objects in Chapter 5. The only modification made was that segmentation masks were now right-aligned rather than center-aligned, since differing lengths of the manipulator can be in view. This builds in the assumption that the manipulator is long and thin. Figure 6-5 shows the results of alignment. The segmentation that best matches the averaged prototype is a good segmentation. So the basic goal of the procedure, to acquire good images of the manipulator, seems to be reasonably met.

The segmentations can be passed directly to the object localization system developed in Chapter 5, with one modification. Again, the 'center' of the manipulator is not well-defined, we are better off working with its endpoint. This could be detected from where the manipulator strikes the object during the segmentation procedure, but as a shortcut the endpoint was simply defined as the rightmost point in the object (this works fine since the robot only pokes with its left arm). Typical localization results are shown in Figure 6-6.

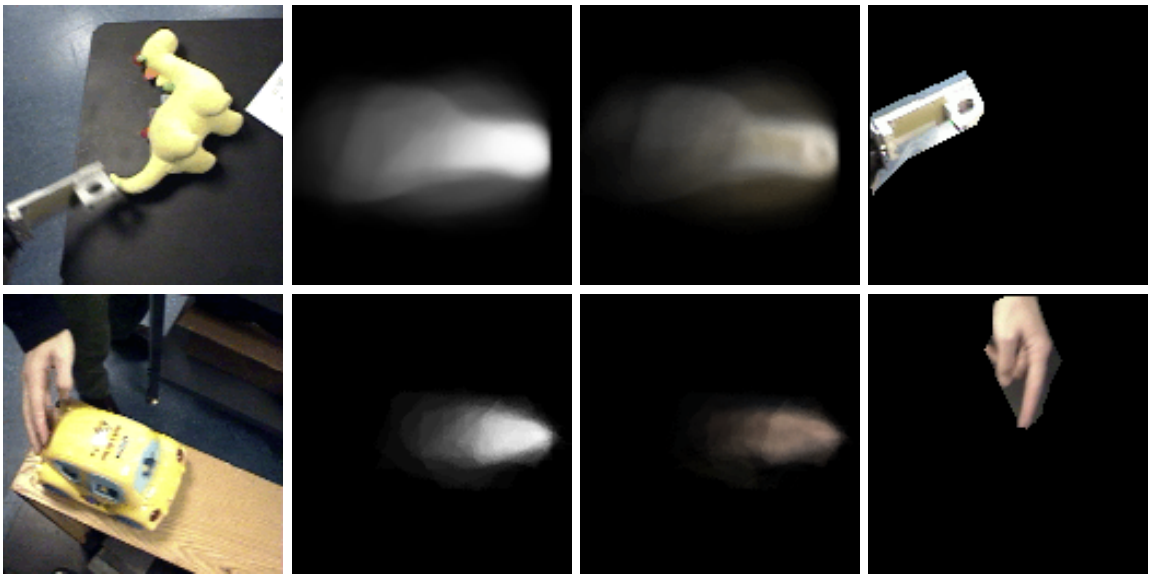

Figure 6-5: The robot manipulator (top left) was automatically segmented during 20 poking sequences. The segmentations were aligned and averaged, giving the mask and appearance shown in the adjacent images. The best matching view is shown on the top right. A similar result for the human hand is shown on the bottom, based on much less data (5 poking sequences, hands of two individuals).

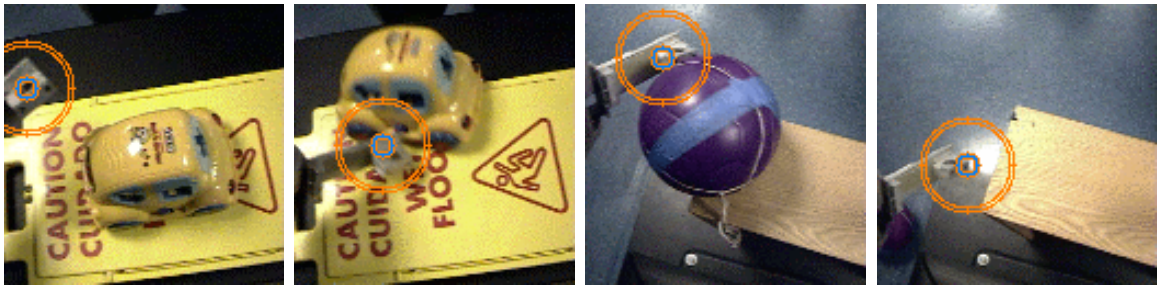

Figure 6-6: After training, the endpoint of the robot's arm can be reliably detected when it is in view (as indicated by the orange circle), despite variations in distance to the arm and how much of the arm is in view.