Towards Manipulation-Driven Vision

Paul M. Fitzpatrick* and Giorgio Metta*,†

*MIT AI Lab – Massachussetts Institute of Technology – USA
†Lira Lab, DIST – University of Genova – Italy

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.

1 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 1. 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 2 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 –

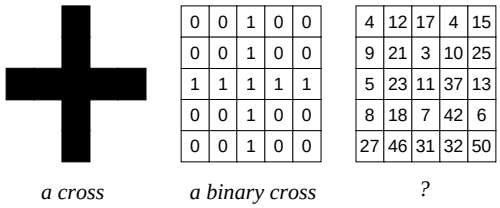

Figure 1: *Three examples of crosses, following [12]. The human ability to segment objects is not general-purpose, and improves with experience.*

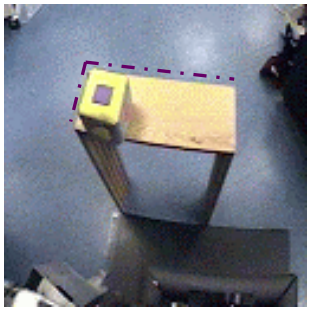

Figure 2: *A cube on a table. The edges of the table and cube happen to be aligned (dashed line), the colors of the cube and table are not well separated, and the cube has a potentially confusing surface pattern.*

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:

> ...objects can only be known by touch. Vision is subject to illusions, which arise from the distance-size problem... [2]

In this paper, we provide support for a more nuanced proposition: that in the presence of touch, 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 un-

likely) 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 [11] 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 [20] 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. 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[15].

The dorsal stream goes through the parietal lobe and premotor cortex, which heavily project into the primary motor cortex to eventually control movements. For many years the premotor cortex was considered just another big motor area. Recent studies [9] have demonstrated that this is not the case. Visual responsive neurons have been found: some purely visual but many with interesting visuo-motor characteristics. In area F5 in the monkey neurons responding to object manipulation gestures are found.

Area F5 contains a neural analogue to the affordances of Gibson [8]: neurons that activate both when grasping an object and when fixating the same object. Other neurons in the same area activate when grasping an object or when watching someone else manipulating that object. This "mirror" representation may be important for mimicry behaviors [6] and perhaps even language [19]. Another important class of neurons in premotor cortex is found in area F4 [7]. While F5 is more concerned with the distal muscles (i.e. the hand), F4 controls more proximal muscles (i.e. reaching). A subset of neurons in F4 exhibit a joint somatosensory, visual, and motor receptive field. The visual receptive field (RF) extends in 3D from a given body part – for example, the forearm. The somatosensory RF is usually in register with the visual one. Finally, 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.

A working hypothesis. Taken together, these results from neuroscience suggest that relating motor action to visual perception is fruitful. Certainly they are intertwined at a very basic level in human vision. 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 presegmenting 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.

Neurons in area F4 are thought to provide a body map useful for generating arm, head, and trunk movements. Our robot learns autonomously a crude version of this body map by fusing vision and proprioception. As a step towards establishing the kind of visuomotor representations seen in F5, we then develop a mechanism for using reaching actions to visually probe the connectivity and physical extent of objects without any prior knowledge of the appearance of the objects (or indeed of the arm itself).

2 The experimental platform

This work is implemented on the robot Cog, an upper torso humanoid [4]. The robot has previously been applied to tasks such as visually-guided pointing [13], and rhythmic operations such as turning a crank or driving a slinky [21]. 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 [22] – 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.

3 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

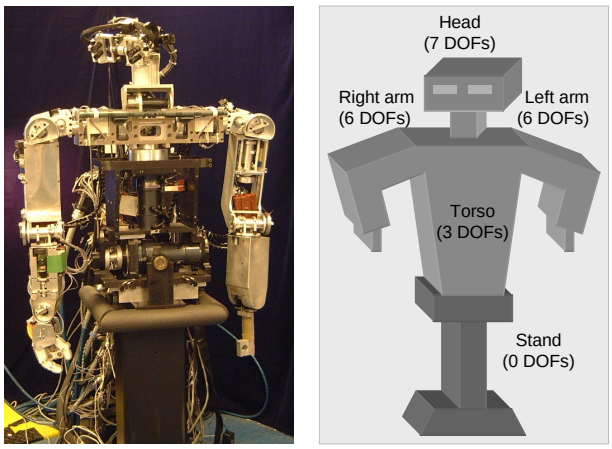

Figure 3: *Degrees of freedom (DOFs) of the robot Cog. The arms terminate either in a primitive "flipper" or a four-fingered hand. The head, torso, and arms together contain 22 degrees of freedom.*

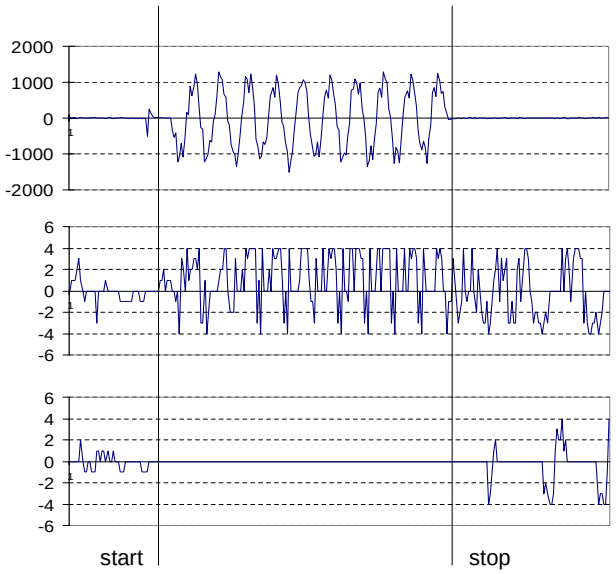

Figure 4: *An example of the correlation between optic flow and arm movement. The traces show the movement of the wrist joint (upper plot) and optic flow sampled on the arm (middle plot) and away from it (lower plot). As the arm generates a repetitive movement, the oscillation is clearly visible in the 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.*

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 [14] using a stereotyped trajectory, perhaps using postural primitives [17]. 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 [13, 14]. 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 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 [13] 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

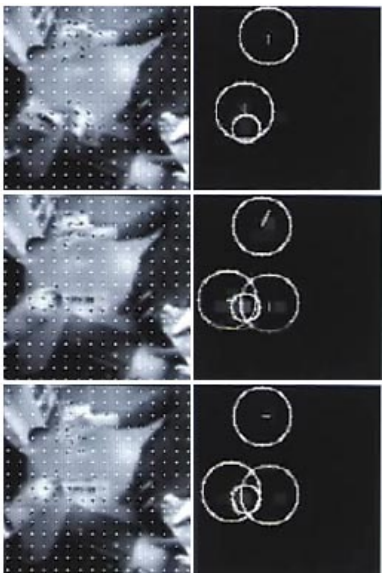

Figure 5: *Detecting the arm/gripper through motion correlation. The robot's point of view and the optic flow generated are shown on the left. On the right are the results of correlation. Large circles represent the results of applying a region growing procedure to the optic flow. Here the flow corresponds 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.*

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 4 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 5 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, reducing the 11-dimensional input space of joint angles based on the much lower number of degrees of freedom used in controlling them (see Figure 6). Figure 7 shows the resulting behavior after about twenty minutes of real-time learning.

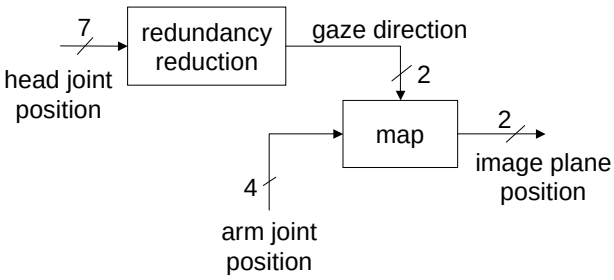

Figure 6: *Mapping from proprioceptive input to a visual prediction. Head and arm joint positions are used to estimate the position of the projection of the hand in the image plane. Redundant configurations of the (7 DOF) head are mapped to a simpler (2D) representation, and the wrist-related DOFs of the arm are ignored.*

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 initialize a visual search for the exact position of the end-effector.*

4 Perceiving indirect effects of action

We have assumed that the target of a reaching operation is chosen visually. As discussed in Section 1, 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 2 (a

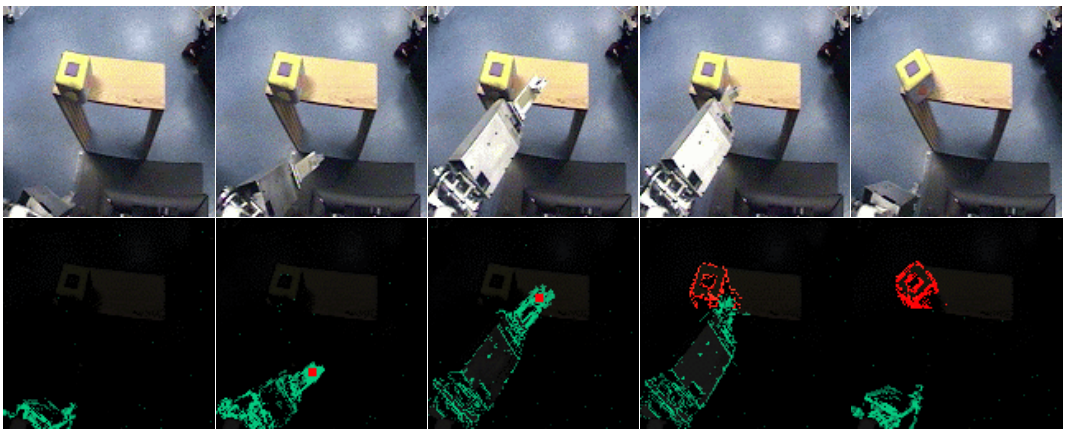

Figure 8: *The upper sequence shows an arm extending into a workspace, tapping an object, and retracting. This is an exploratory mechanism for finding the boundaries of objects, and essentially requires the arm to collide with objects under normal operation, rather than as an occasional accident. The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

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 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 [3]. 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 8 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 of physical assemblages.

5 Discussion and Conclusions

The number of papers written on techniques for visual segmentation is vast. Methods for characterizing the shape of an object through tactile information are also being developed, such as shape from probing [5, 18] or pushing [16, 10]. But while it has long been known that motor strategies can aid vision [1], work on active vision has focused almost exclusively on moving cameras. There is much to be gained by bringing a manipulator into the equation, as we have shown in this paper. 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.

Acknowledgments

This work benefited from discussions with Charles Kemp and Giulio Sandini. Many people have contributed to developing the Cog platform [4]. 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

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

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

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

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

[5] R. Cole and C. Yap. Shape from probing. *Journal of Algorithms*, 8(1):19–38, 1987.

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

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

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

[9] M. Jeannerod. *The Cognitive Neuroscience of Action*. Blackwell Publishers Inc., Cambridge Massachusetts and Oxford UK, 1997.

[10] Yan-Bin Jia and Michael Erdmann. Observing pose and motion through contact. In *Proceedings of the IEEE International Conference on Robotics and Automation*, May 1998.

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

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

[13] Matthew J. Marjanović, Brian Scassellati, and Matthew M. Williamson. 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, 1996.

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

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

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

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

[18] E. Paulos. Fast construction of near optimal probing strategies. Master's thesis, University of California, Berkeley, 1999.

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

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

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

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