Learning from shared activity

Paul Fitzpatrick*
*MIT AI Lab
Cambridge, Massachusetts, USA
paulfitz@ai.mit.edu

Giorgio Metta*†
†Lira Lab, DIST, Uiversity of Genova
Genova, Italy
pasa@dist.unige.it

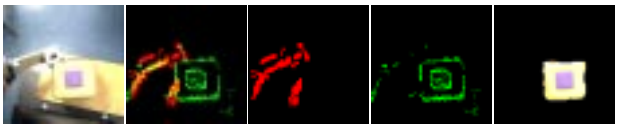
Abstract

Imitative behavior requires a mapping between the action of another and one's own action. This is a challenging perceptual problem. We show how a robot can expand its perceptual abilities far beyond an initial set of primitives by acquiring high-quality visual experience within the context of a simple object manipulation activity. The robot pokes objects and watches a human companion poke the same objects, and uses motion cues to segment the object, its own arm, and the human's hand. From the data collected, the robot learns about object motion and recognition, about the appearance of the human hand, and can train up a low-level orientation filter not present in its primitive set of filters. The representation of objects and motion is analogous to canonical and mirror neurons, and so is fundamentally well suited to imitation.

1 Introduction

To imitate an action, it must first be perceived correctly. The actor must be located and identified, and then tracked throughout the action. The same needs to be done for any objects involved. Perception is an interpretive process that endeavors to capture the essentials of an agent's context and discard incidental, irrelevant details. This differentiation between essential and incidental is crucial to imitation, but simply represents an extreme of the kind of decision the perceptual system faces at every instant. Hence it seems wise to address imitation within a wider context of a fully integrated perceptual system.

We show how a robot can acquire the appropriate percepts for imitation through a shared activity, one that can be done either by the robot or a human. This activity is simply striking objects and watching how they move.



Moment of impact | *Motion in frame immediately after impact* | *Aligned motion from before impact* | *Masking out prior motion* | *Final segmentation*

Figure 1: Active segmentation. The robot arm is deliberately driven to collide with an object. The apparent motion after contact, when masked by the motion before contact, identifies a seed foreground (object) region. Such motion will generally contain fragments of the arm and environmental motion that escaped masking. Motion present before contact is used to identify background (non-object) regions. An optimal object region is computed from the foreground and background information using graph cuts (Fitzpatrick, 2003).

2 A shared activity: poking

A robot equipped with an arm and an active vision head was given a simple "poking" behavior, whereby it selected objects in its environment and struck them (Fitpatrick and Metta, 2002). Since the robot had a limited reach, this activity required the cooperation of a human companion to bring the robot interesting objects to poke. The behavior could also be preempted by the companion. When the robot fixated an object and was about to reach for it, the companion could choose to poke the object instead, in which case the robot would refrain from acting.

This choice of activity has many benefits. *(i)* The motion signature generated by the impact of the arm with a rigid object greatly simplifies segmenting that object from its background, and obtaining a reasonable estimate of its boundary (see Figure 1). This "active segmentation" procedure is key to automatically acquiring training data of sufficient quality to support the many forms of learning described in the remainder of this paper. *(ii)* The poking activity also leads to object-specific consequences, since different objects respond to poking in different ways. For example, a toy car will tend to roll forward, while a bottle will roll along its side. *(iii)* The basic operation involved, striking objects, can be performed by either the robot or its human companion, creating a controlled point of comparison between robot and human action.

3 Learning about motion

When the robot pokes an object and acquires its segmentation, it is straightforward to track how the object moves after impact. Pooling this data over all objects reveals how the movement of the robot's arm – which can ap-

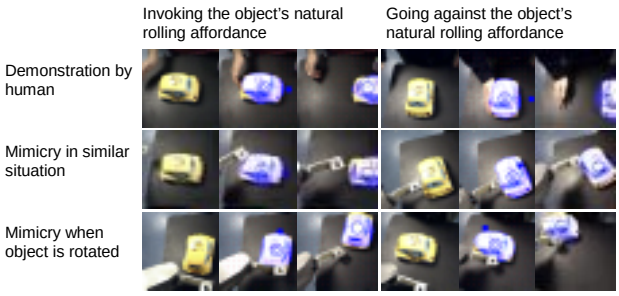


Figure 2: The sequences on the left show the robot mimicking a human exploiting a car's rolling "affordance". The sequences on the right show what happens when the human hits the car in a contrary fashion, going against its preferred direction of motion. The robot mimics this "unnatural" action, suppressing its usual behavior of trying to evoke rolling.

proach the object from several directions – correlates with the final movement of the object.

But not all objects behave the same way when struck. Some objects have a preferred direction of motion – for example, a toy car tends to roll forward along its principal axis, or a bottle might roll along its side. The robot was made sensitive to differences in object motion conditioned on object identity (color histogram) and principal axis. The robot exhibited what it learned in two ways. It would attempt to strike objects so that they would roll. It also engaged in simple form of mimicry (see Figure 2). If the human companion moved an object in a way that exploited an "affordance" such as rolling, the robot would attempt to do the same; if the human moved the object in an "unnatural" way, the robot would attempt to copy this too.

4 Learning object appearance

The affordance learning in the last section depended on a mechanism for differentiating objects based on their color histogram. The data collected over hundreds of pokes contained a significant...

By clustering segmented views of objects based on color histograms, the robot collected about 100 views of each object. If the quality of the clusters generated is sufficiently good, it should be possible to extract reliable consensus prototypes for each object. This is in fact the case, as Figure 3 shows. Using the most naive alignment procedure and averaging process possible, a blurry "mean" view of the objects can quickly be derived. This could be sharpened by better alignment procedures, or just used to pick out the best single match to the mean view for each object. Of course, this paper is not proposing that simple color histograms are how recognition should be done – there are better ways (for example, see Schiele and Crowley (2000)), rather it is giving evi-

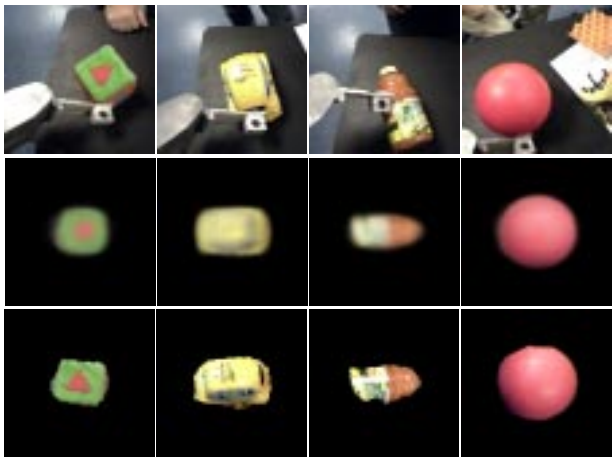


Figure 3: The top row shows the four objects used in this experiment, seen from the robot's perspective. The middle row shows prototypes derived for those objects using a naïve alignment procedure. None of the prototypes contain any part of the robot's manipulator, or the environment. These prototypes are used to find the best available segmentations of the objects (bottom row).

dence that active segmentation can generate data of sufficient quality to train up a recognizer.

Figure 3.

5 Learning actor appearance

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. We can create an effective procedure for learning about manipulators by simply giving the robot a predisposition to fixate familiar objects. This enables 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 of its own arm from that of others simply by checking whether it was commanding its arm to move towards the target at the time. 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 Section **??**, where the foreground is now taken to be regions moving at the desired velocity, and the background is everything else. Figure 4 shows preliminary results for this proce-

Figure 4: 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).

dure. The results are based on relatively little data, yet are already sufficient to pick out good prototype views for the robot and human manipulator. A procedure like this could be used to autonomously train a recognizer for the human hand, which could then be included in further operational definitions, expanding the robot's domain of grounded knowledge ever outwards – but this is very much future work.

Figure 4.

6 Learning low-level vision

Orientation is an important visual cue for many purposes, such as object segmentation, recognition, and tracking. It is associated with neighborhoods rather than individual points in an image, and so is inherently scale dependent. At very fine scales, relatively few pixels are available from which to judge orientation. Lines and edges at such scales are extremely pixelated and rough. Orientation filters derived from analytic considerations, with parameters chosen assuming smooth, ideal straight lines or edges (for example, Chen et al. (2000)) are more suited to larger neighborhoods with more redundant information. For fine scales, an empirical approach seems more promising, particularly given that when the number of pixels involved is low, it is practical to sample the space of all possible appearances of these pixels quite densely.

This paper is an exploration of how edges in "natural" images appear when viewed through an extremely small window (4 by 4 pixels). This window size is chosen to be large enough to be interesting, but small enough for the complete range of possible appearances to be easily visualized. Even at this scale, manual data collection and labelling would be extremely tedious, so a robotic system Brooks et al. (1999) was employed to automatically compile a database of the appearance of oriented features (Section **??**). These features were extracted by sampling image patches along object boundaries, which were in turn determined using active segmentation Fitzpatrick (2003). The resulting "catalog" of edge appearances proved remarkably diverse, although the most frequent appearances were indeed the "ideal" straight, noise-free edge (Section **??**). Finally, it is a simple matter to take this catalog of appearances and use it as a memory-based image processing filter (Section **??**).

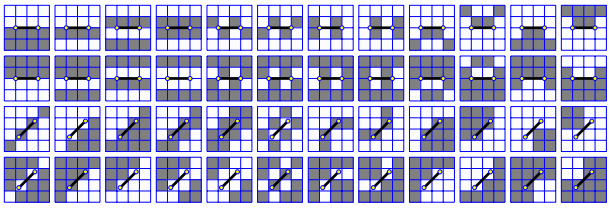


Figure 5: Edges have diverse appearances. This figure shows the orientations assigned to a test suite prepared by hand. Each 4×4 grid is a single test edge patch, and the dark line centered in the grid is the orientation that patch was observed to have in the training data. The oriented features represented include edges, thin lines, thick lines, zig-zags, corners etc. It is difficult to imagine a set of conventional filters that could respond correctly to the full range of features seen here – all of which appeared multiple times in object boundaries in real images.

Figure 5.

7 Discussion and conclusions

Robots and animals are actors in their environment, not simply passive observers. They have the opportunity to examine the world using causality, by performing probing actions and learning from the response. 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.

Acknowledgements

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

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.

J. Chen, Y. Sato, and S. Tamura. Orientation space filtering for multiple orientation line segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

P. Fitpatrick and G. Metta. Towards manipulation-driven vision. In *IEEE/RSJ Conference on Intelligent Robots and Systems*, 2002.

P. Fitzpatrick. First contact: Segmenting unfamiliar objects by poking them. 2003. Submitted to IEEE Conference on Computer Vision and Pattern Recognition.

B. Schiele and J. L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36 (1):31–50, January 2000.