Learning About Objects Through Action - Initial Steps Towards Artificial Cognition

Paul Fitzpatrick*; Giorgio Metta*†; Lorenzo Natale†; Sajit Rao†; Giulio Sandini†
* AI-Lab, MIT, Cambridge, MA, U.S.A.
†LIRA-Lab, DIST, Univ of Genova, Viale Francesco Causa 13, Genova 16145, Italy
paulfitz@ai.mit.edu; {pasa, nat, sajit, sandini}@dist.unige.it

***Abstract*— Within the field of Neuro Robotics we are driven primarily by the desire to understand how humans and animals live and grow and solve every day's problems. To this aim we adopted a "learn by doing" approach by building artificial systems, e.g. robots that not only look like human beings but also represent a model of some brain process. They should, ideally, behave and interact like human beings (being situated). The main emphasis in robotics has been on systems that act as a reaction to an external stimulus (e.g. tracking, reaching), rather than as a result of an internal drive to explore or "understand" the environment. We think it is now appropriate to try to move from acting, in the sense explained above, to "understanding". As a starting point we addressed the problem of learning about the effects and consequences of self-generated actions. How does the robot learn how to pull an object toward itself or to push it away? How does the robot learn that spherical objects roll while a cube only slides if pushed? Interacting with objects is important because it implicitly explores object representation, event understanding, and can provide definition of objecthood that could not be grasped with a mere passive observation of the world. Further, learning to understand what one's own body can do is an essential step toward learning by imitation. In this view two actions are similar not only if their kinematics and dynamics are similar but rather if the effects on the external world are the same. Along this line of research we discuss some recent experiments performed at the AI-Lab at MIT and at the LIRA-Lab at the University of Genova on COG and Babybot respectively. We show how the humanoid robots can learn how to poke and prod objects to obtain a consistently repeatable effect (e.g. sliding in a given direction), to help visual segmentation, and to interpret a poking action performed by a human manipulator.**

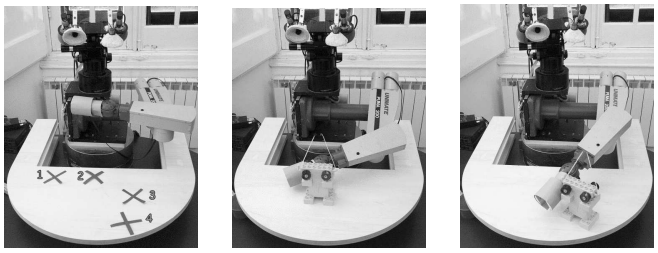
I. Introduction: Learning to Act on Objects

In order to explore the role of sensory information and motor skills in cognition, we are pursuing a developmental approach. The basic idea is that assembling something as complex as a cognitive system from scratch as a collection of modules is virtually impossible. Rather, as in humans, cognitive abilities develop over time layering over previous stages of development. Development may therefore not be just an artifact of biological systems, but a necessary way to manage complexity [1]. To this end we see three broad stages in the development of our humanoid robot platform: The first stage involves learning a *body-image* or body-schema [2] [3]. Having learned to distinguish its body from the rest of the world the robot can move on to the second stage, which is interaction with external objects. The third and final stage involves interpreting object-object interactions. An essential feature of this developmental program is that each stage strictly *requires* and *layers upon* the previous stages. In this paper we present results for the second and third stages in this developmental schema.

Results from neuroscience suggest that action and manipulation are fundamental for acquiring knowledge about objects [4] [5] [6]. These results point out that action and perception are possibly much more intertwined than what was once believed. It would be difficult to draw a separation for where perception ends and action starts. Even the distinction between motor and sensory areas tends to be very blurred.

Drawing more from the neural science literature, we now know that areas active when reaching and/or grasping present a mixed structure containing action and sensory related neurons. Arbib and colleagues [6] interpreted these responses as the neural analogue of the affordances of Gibson [7]. In Gibson's theory an affordance is a visual characteristic of an object which can elicit an action without necessarily involving an object recognition stage. It seems that areas AIP (parietal) and F5 (premotor/frontal) are active in such a way to provide the individual with a mechanism to detect affordances. F5 projects to F1 (primary motor cortex) and can therefore control behavior. In particular area AIP contains neurons that respond both when generating a grasping movement and when observing the object being grasped [5] [8]. Responses are congruent: e.g. a precision grip responsive neuron would also fire during the observation of a small object (for which the precision grip is a likely action). Similar neurons with slightly different temporal characteristics have been observed in F5.

Rizzolatti and coworkers [5] extensively probed area F5. They found another class of grasping neurons that also responded during observation of somebody else's



(a) initial arm-positions for target approach

(b) At the beginning of a pushing movement

(c) At the end of the movement

Fig. 1. The experimental setup

action. They called this newly discovered cells mirror neurons. This activation of F5 is coherent with the idea that the brain internally reproduces/simulates the observed actions. This can possibly form the basis for recognition of complicated biological motion and for mimicry of observed behaviors.

Taken together, these results suggest that the ability to visually interpret the motor-goal or behavioral/purpose of the action may be helped by the monkey's ability to perform that action (and achieve a similar motor goal) itself. Therefore, active exploration of the environment may be critical not only to subsequent performance of an action, but also for the interpretation of the actions of others. We argued at the beginning of this section that this "probing/exploring activity" is the second stage of a developmental sequence. It is at this stage when the ability to interact with object is formed, and includes the detection of affordances and the manipulation of objects. In the following sections we present an implementation of this stage. We will show how a humanoid robot

- Learns the effect of pushing/pulling actions on objects and uses this to drive goal-directed behavior.
- Acquires a particular affordance and behaviorally demonstrates this form of "understanding" about objects.
- Uses the knowledge of affordances, gained through exploration, to interpret human action and mimic the last observed action.

II. Learning to Push/Pull/Poke objects

The goal of this experiment is to learn the effect of a set of simple pushing/pulling actions from different directions on a toy object, and then use the learned knowledge to move a new object in a desired goal-direction.

Figure 1 A Shows the experimental platform "Babybot",an upper torso humanoid robot at the LIRA Lab, Univ of Genova. Babybot has a 5 DOF head, and a 6 DOF arm, and 2 cameras whose Cartesian images are mapped to a log-polar format [9]. The robot also has a force sensitive wrist, and a metal stub for a hand. The target is placed directly in front of the robot on the play-table. Babybot starts from any of four different initial positions (shown in the figure) at the beginning of a trial run.

In a typical trial run the robot continuously tracks the target while reaching for it. The target (even if it is moving) is thus ideally always centered on the fovea, while the moving hand is tracked in peripheral vision. Figure 1 (B) shows the arm at one of its initial positions and (C) shows the end of the trial with the target having been pushed to one side.

During each such trial run, the time evolution of two event variables are continuously monitored: the initial proprioceptive hand-position, and then at the moment of contact (when the hand first touches the object) the direction of the *retinal* displacement of the target.

A. The Target Representation and Results of Learning

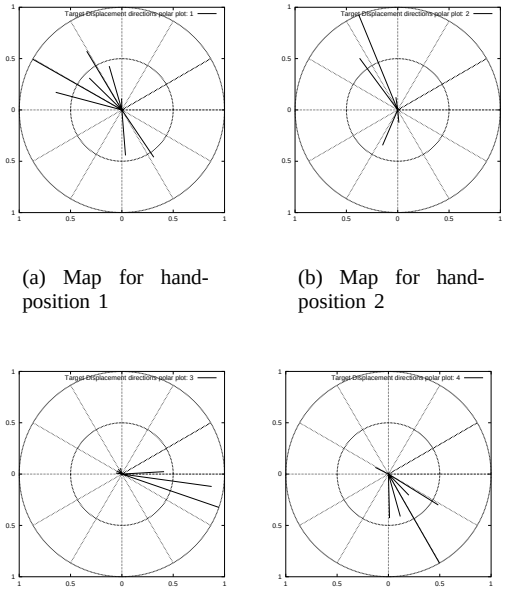
The purpose of the training phase is to learn a mapping from initial hand position to direction of target movement. Therefore, associated with each initial hand position is a direction map (a circular histogram) that summarizes the directions that the target moved in when approached from that position. After each trial the appropriate direction map is updated with the target motion for that particular trial.

Approximately 70 trials, distributed evenly across the four initial starting positions, were conducted. Figure 2 shows the four direction maps learned, one for each initial arm position considered. The maps plot the frequency with which the target moved in a particular direction at the moment of impact. Therefore longer radial lines in the plot point towards the most common direction of movement. As we can see, the maps are sharply tuned towards a dominant direction.

B. Testing the Learned Maps

The learned maps are used to drive motor planning in a straightforward manner as shown in Figure 3. The robot is presented with the usual target as before, but this time also with another toy nearby (Figure 3 a). The goal is to push the target towards the new toy. The robot first extracts the desired displacement vector from the scene, finds the direction map which is most tuned in that direction. The initial-hand position corresponding to that map, is used and the dynamics takes care of the rest, resulting in the motion of the target towards the desired direction, (Figure 3b,c).

A quantitative measure of the performance before and after learning is to look at the error angle between the



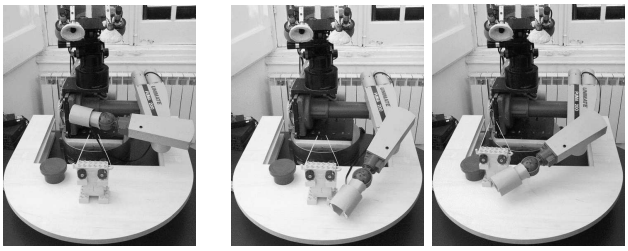
(a) Map for hand-position 1

(b) Map for hand-position 2

(c) Map for hand-position 3

(d) Map for hand-position 4

Fig. 2. The learned target-motion direction maps, one for each initial hand-position

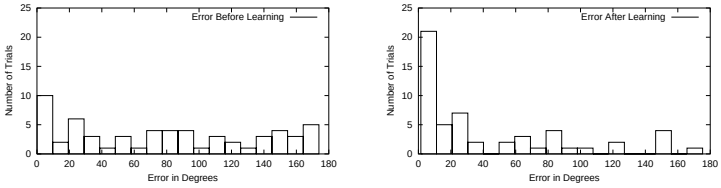


(a) The round toy is the new desired target position

(b) The learned maps are used to re-position the arm in preparation for pushing

(c) At the end of the movement

Fig. 3. The learned direction maps are used to drive goal-directed action



(a) Error distribution before learning

(b) Error distribution after learning

Fig. 4. Improvement in performance: Plots of the distribution of the angle between the desired direction and actual direction, before and after learning. Zero degrees indicates no error, while 180 degrees indicates maximum error.

desired direction of motion (of the target towards the goal) and the actual direction that the target moved in when pushed. First, as a baseline control case, 54 trials were run with the goal position (round toy) being varied randomly and the initial hand-position being chosen randomly among the four positions (i.e learned maps are not used to pick the appropriate hand position). Figure 4(a) shows the error plot. The distribution of errors is not completely flat as one would expect because the initial hand-positions are not uniformly distributed around the circle. Nevertheless, the histogram is not far from uniform. Doing the same experiment, but using the learned map to position the hand, yields the error plot shown in 4(b). As we can see the histogram is significantly skewed towards an error of 0, as a result of picking the correct initial-hand position from the learned map.

III. Learning Object Affordances

In the previous section, we ignored both the identity of the object, and the gross shape (elongation) of the object, and learned possibly the simplest property of the behavior of the object; namely the instantaneous direction of motion as a result of a push/pull action.

In this section we take both identity, and some shape properties of the object into account and thereby have a more detailed characterization of the behavior. The uniqueness of the motion-signature of the object can used to identity the object itself, and can be associated with the visual appearance of the object. We describe an experiment where a robot acts on a small set of objects using a small motor repertoire consisting of four actions indicated for convenience as pull in, side tap, push away, and back slap.

During a training/exploration stage the robot performs several trials for each ⟨object, action⟩ pair and learns the motion signature of the behavior of the object for that action. We show how, in a more goal-directed mode, the robot uses it's knowledge of the object affordances to choose the appropriate action to make a given object roll. Finally we show how the robot is able to interpret the effect of a human action on an object, and respond with it's own action that produces a similar effect, i.e. mimicry.

The robot used for this experiment was "Cog", an upper torso humanoid robot [10] at the MIT A.I. Lab. Figure 5 shows the robot, Cog has two arms, each of which has six degrees of freedom. The joints are driven by series elastic

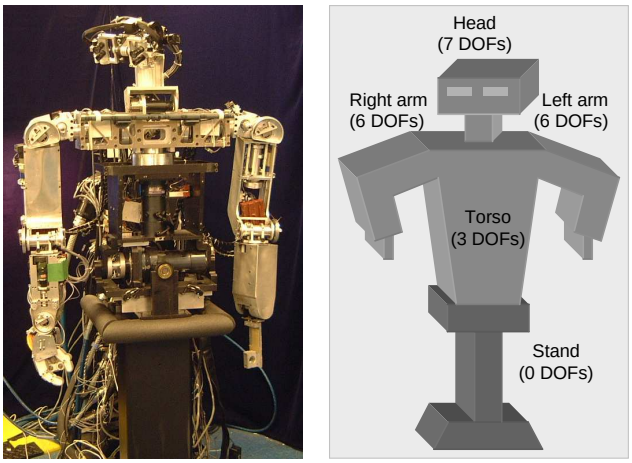


Fig. 5. Degrees of freedom of the robot Cog. The arms terminate in a primitive flipper. The head, torso, and arms together contain 22 degrees of freedom.

actuators - essentially a motor connected to its load via a spring. Cog's head has seven degrees of freedom and mounts four cameras and a gyroscope simulating part of the human vestibular system.

A. Characterizing Object Behavior

During the training phase, each of the four objects in this experiment (an orange juice bottle, a toy car, a cube, and a colored ball) is "poked" about a 100 times, i.e. roughly 25 repetitions of each action for each object. The visual feature of the resulting object behavior that is extracted is the instantaneous direction of motion of the object (as in the previous section) but this time *the angle is relative to the principle axis of inertia of the object*. The direction of motion of the object and principle axis of inertia are extracted as follows.

During a single poking operation, the arm is identified as the first object to move in the scene. Once the arm is identified, a sudden spread of motion in the image (spatially correlated with the end-effector) identifies a contact. Whatever was moving before this instant is considered background (the arm, other disturbances). Thereafter, the newly moving "blob" can be identified as the object. A further refinement is needed to fill in the gaps since motion detection is in general sparse and depends on the visual appearance/texture of the object. Two examples of poking and segmentations are shown in Figure 6.

Having segmented the object the principle axis of inertia is easily extracted, and its instantaneous direction of motion is gathered from the optical flow information of the pixels belonging to the object. Then the difference between these two angles is used to update a "motion-signature probability map" of the kind shown in Figure

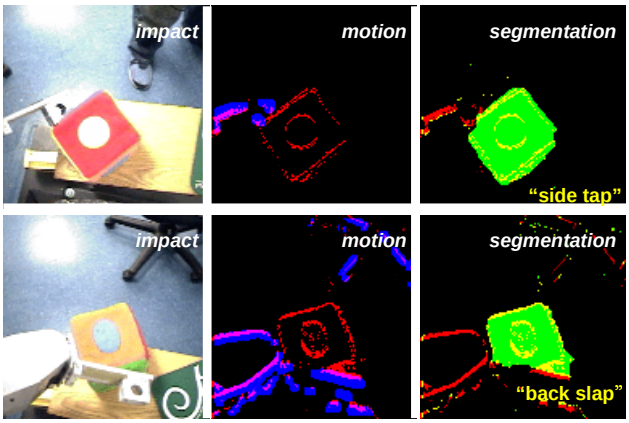


Fig. 6. Cog poking a cube. The top row shows the flipper poking the object from the side, turning it slightly. The second row shows Cog batting an object away. The images in the first column are frames at the moment of impact. The second column shows the the motion signal at the point of contact. The bright regions in the images in the final column show the segmentation produced for the object.

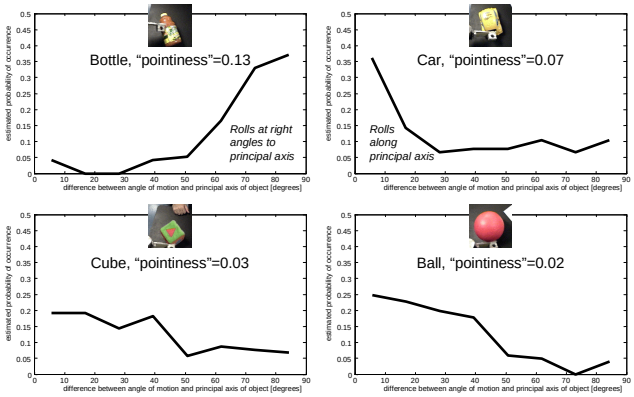


Fig. 7. Probability of observing a roll along a particular direction for the set of four objects used in Cog's experiments. Abscissas represent the difference between the principal axis of the object and the observed direction of movement. Ordinates are the estimated probability.

7.

B. Results and Demonstration of Learning

For each ⟨object, action⟩ pair the representation of the object affordances or movement signature is in terms of a histogram of probabilities for each relative angle of motion. In other words the probabilities represent the likelihood of observing each of the objects rolling along a particular direction with respect to their principal axis. Figure 7 shows one set of such estimated probability maps learned from the exploration/training stage.

To provide a behavioral demonstration of the learned affordances, one of the known objects was placed in front of Cog, with Cog's task being to choose the action that would most likely make that particular object roll.

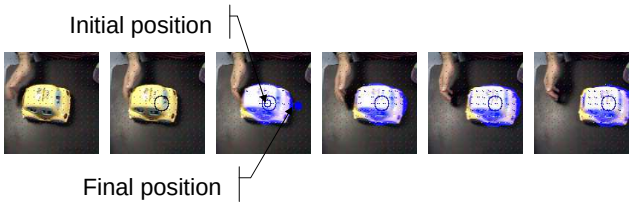


Fig. 8. An example of observed sequence. Frames around the instant of impact are shown. Initial and final position after 12 frames is indicated.

The object is recognized, localized and its orientation estimated (principal axis). Recognition and localization are based on the same color information collected during learning. Cog then uses its understanding of the affordance of the object (Figure 7) and of the geometry of poking to choose the action that is most likely to make the object roll. The object localization procedure has an error between $10°$ and $25°$ which proved to be tolerable for our experiment. We performed a simple qualitative test of the overall performance of the robot. Out of 100 trials the robot made 15 mistakes. Twelve of them were due to imprecise control: e.g. the end point touched the object earlier than expected moving the car outside the field of view. The remainders (3) were genuine mistakes due to misinterpretation of the object position/orientation.

This experiment represents an analogue to the response of F5/AIP as explained in Arbib's model [6] in that a specific structure of the robot detects the affordance of the object and links it to the generation of behavior. This is also the first stage of the development of more complex behaviors which relies on the understanding of objects as physical entities with specific properties.

C. Understanding the actions of others

An interesting question then is whether the system could extract useful information from seeing an object manipulated by someone else.

In fact, the same visual processing used for analyzing an active poking has been used to detect a contact and segment the object from the manipulator. The first obvious thing the robot can do is identify the action just observed with respect to its motor vocabulary. It is easily done by comparing the displacement of the object with the four possible actions and by choosing the action whose effects are closer to the observed displacement. This procedure is orders of magnitude simpler than trying to completely characterize the action in terms of the observed kinematics of the movement. Here the complexity of the data we need to obtain is somewhat proportional to the complexity of the goal rather than that of the structure/skill of the foreign manipulator.

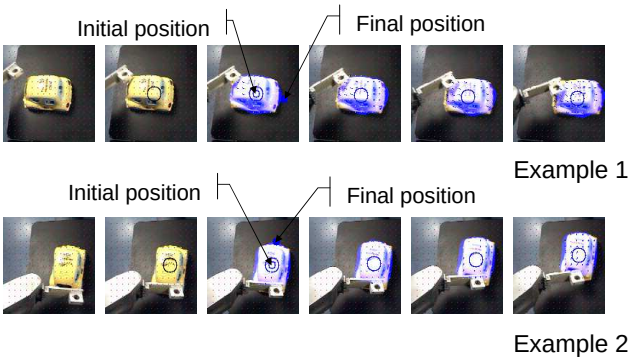


Fig. 9. Two examples of mimicry following the observation of Figure 8.Cog mimics the goal of the action (poking along the principal axis) rather than the trajectory followed by the toy car.

The robot can also mimic the observed behavior if it happens to see the same object again. This requires another bit of information. The angle between the affordance of the object (preferred direction of motion) and the observed displacement is measured. During mimicry the object is localized as in section III-A and the action which is more likely to produce the same observed angle (relative to the object) is generated. If, for example, the car was poked at right angle with respect to its principal axis Cog would mimic the action by poking the car at right angle. In spite of the fact that the car preferred behavior is to move along its principal axis. Examples of observation of poking and generation of mimicry actions are shown in Figure 8 and 9.

IV. Conclusion

All biological systems are embodied systems, and an important way they have for recognizing and differentiating between objects in the environment is by simply acting on them. Only repeated interactions (play!) with objects can reveal how they behave when acted upon (e.g. sliding vs rolling when poked). We have shown two experiments where, in a discovery mode, the visual system learns about the consequences of motor acts in terms of such features, and in a goal-directed mode the mapping may be inverted to select the motor act that causes a particular visual change. These two modes of learning; the consequences of a motor act, and selecting a motor act to achieve a certain result, are obviously intertwined, and together are what we mean by "learning to act". Furthermore, the same information can also be used to interpret the effect of a human-action on an object (as seen in mirror neurons), and thereafter select an appropriate action to mimic the effect on the object. The experiments together underline the central theme that learning to act on objects is very important, not only to get better at interacting with future objects, but also to interpret the actions of others.

V. Acknowledgments

Work on Cog was funded 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.

Work on BabyBot was funded by European Commission Information Society Technologies branch, as part of the "Cognitive Vision Systems" project under contract number IST-2000-29375, and also as part of the MIRROR project under contract number IST-2000-28159.

VI. References

[1] G. Metta, G. Sandini, L. Natale, R. Manzotti, and F. Panerai, "Development in artificial systems," in *Proc. EDEC Symposium at the International Conference on Cognitive Science*, (Beijing, China), Aug. 2001.

[2] J. Lackner and P. DiZio, "Aspects of body self-calibration," *Trends in Cognitive Sciences*, vol. 4, July 2000.

[3] A. Sirigu, J. Grafman, K. Bressler, and T. Sunderland, "Multiple representations contribute to body knowledge processing. evidence from a case of autopagnosia," *Brain*, vol. 114, pp. 629–642, Feb 1991.

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

[5] V. Gallese, L. Fadiga, L. Fogassi, and G. Rizzolatti, "Action recognition in the premotor cortex," *Brain*, vol. 119, pp. 593–609, 1996.

[6] A. Fagg and M. Arbib, "Modeling parietal-premotor interaction in primate control of grasping," *Neural Networks*, vol. 11, no. 7–8, pp. 1277–1303, 1998.

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

[8] M. Jeannerod, M. Arbib, G. Rizzolatti, and H. Sakata, "Grasping objects: the cortical mechanisms of visuomotor tranformation," *Trends in Neurosciences*, vol. 18, no. 7, pp. 314–320, 1995.

[9] G. Sandini and V. Tagliasco, "An anthropomorphic retina-like structure for scene analysis," *Computer Vision, Graphics and Image Processing*, vol. 14, pp. 365–372, 1980.

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