Learning haptic representation of objects

Lorenzo Natale, Giorgio Metta and Giulio Sandini
LIRA-Lab, DIST – University of Genoa
viale Causa 13, 16145 Genova, Italy
Email: nat, pasa, sandini @dist.unige.it

Abstract

Results from neuroscience suggest that the brain has a unified representation of objects involving visual, haptic, and motor information. This representation is the result of a long process of exploration of the environment linking the sensorial appearance of objects with the actions that they afford. Somewhat inspired by these results, in this paper, we provide support to the view that such representation is required for certain skills to emerge in an artificial system and present the first experiments along the route.

1 Introduction

All biological systems share the capability of actively interacting with the environment; however, among all species only primates have the ability to actually manipulate objects and to elevate some of them to the status of tools. This includes the ability to handle small as well as relatively bigger objects, to grasp them in many diverse ways, and to select the most appropriate depending on the task to be fulfilled. Grasping allows primates to gather information about objects that otherwise would not be available (e.g. physical properties like softness, roughness, or weight) and, in addition, to relate this information to cues coming from other sensory modalities (such as vision). This not only because tactile and proprioceptive information is available trough direct contact but, more interestingly, because of the causal link between one's own actions and the entities acted upon. That is, acting produces consequences that can be sensed and properly associated to objects' properties. Recent neurophysiological findings started to probe how deep and intricate it is the link between action, the interaction of the physical body with the environment, and the emergence of cognition in humans [1, 2]. According to these results the representation of objects, of our object directed actions, and of our body's skills and shape are deeply intertwined [3, 4]. While this is true in general, it is even truer when manipulation is considered.

In robotics, dexterous manipulation has been studied extensively and there have been many attempts to build and control articulated hands [5]. Although exceptionally important this effort may still be of limited scope if our true aim is rather that of implementing cognitive abilities in an artificial system.

In previous experiments we showed how a robot could exploit self-generated actions to explore object properties [6-8]. However in those cases, the robot did not have a dexterous hand and very simple actions were used instead (such as poking and prodding). In the same spirit but with a more sophisticate hardware we present here a preliminary experiment with an upper torso humanoid robot equipped with a binocular head, an arm and a five-finger hand. The goal is to explore the possibility of gathering physical properties of objects from very little prior knowledge and to understand what kind of parameters can be extracted from proprioceptive/tactile feedback. We show that given an extremely simple explorative strategy the robot is able to build a representation of objects that happen to touch its hand. The motor action is defined in advance and elicited by tactile stimulation. The explorative strategy and the hand's passive compliance suffice in starting to acquire structured information about the physical properties of objects drawn from a small set. In particular, we will show that the system categorizes objects by exploiting differences on their shape and weight.

The paper is organized as follows. In the next section we present our motivations for pursuing this particular approach. The robotic setup and the experiments are described in section 3 and 4 respectively. We conclude in section 4 by discussing the results and drawing the conclusions.

2 A unified representation of objects

The reconstruction of a visual scene based on visual information alone is an ill-posed problem [9]. This notwithstanding it seems that the brain is able somehow to dispel all possible illusions and provide us with a consistent 3D picture of the outer world. The overall process that makes this possible is far from being understood although it has been widely investigated by neuroscientists, physiologists, roboticists, and by computer scientists. Many agree on the fact that the brain takes advantage not only of visual cues, but also of the wealth of multimodal information from other senses and from the

kinaesthetic experience derived from the interaction of the body with the environment. The representation of the world in adults is the result of a long active process of collecting information which starts in infancy and continues all along our life. We use the word *active* to stress the fact that we are not passive observers in the world. If on the one hand it is only by acting that we can access objects' properties that otherwise would not be available (like weight, roughness or softness), on the other actions allow us to learn the consequences of the interaction between the body morphology and the object. According to Jannerod [10] the brain has a pragmatic representation of the attributes relevant for action. This is somehow different from the semantic representation grouping together all information necessary for object recognition and categorization. The former includes parameters relevant for shaping the hand according to the size, weight and orientation of the object we are going to grasp. The latter has the function of forming a *perceptual image* of the object in order to identify it. In dealing with an object the brain has to solve the following questions: *what* the object is, *where it is* and *how* to handle it. The representation of *where* and *how* constitutes the pragmatic representation which is directly related to action. The representation of *what* is related to the conscious perception of the object and corresponds to its semantic representation.

The *where* representation is completely different from the others and does not directly involves knowledge of objects. The representation of *what* the object is and *how* it can be manipulated are normally integrated but under certain conditions can be dissociated. There seems to be two independent circuits in the brain dealing with the two types of cues. This is suggested by behavioral studies about reactions time in humans, by anatomical studies performed in monkeys, and from the observation of patients with lesions in the posterior parietal cortex (for a review see [10]).

Although separated both representations are based on knowledge that is acquired (learned) by interacting with objects. Even when answering the *what* question, information about shape, size and weight might prove helpful to bias the recognition in cases when only ambiguous cues are available. Similarly, the same cues are used during grasp to anticipate the shape of the hand thus to achieve a stable grasp. Visual information in this case activates the brain circuitry responsible for the pragmatic representation of the object to be grasped which controls the orientation of the hand, its maximum aperture and the opposition space.

Recent studies on the monkey motor cortex have revealed the existence of neurons which code a similar pragmatic representation of objects [2]. A group of neurons located in the monkey premotor cortex (area F5) are activated both when producing a motor response to drive an object-directed grasping action and when only fixating a graspable object. This population of neurons seems to constitute a vocabulary of motor actions that could be applied to a particular object. This response is somewhat reminiscent of Gibsonian affordances because it represents the ensemble of grasping actions that an object affords [3]. The link between action and perception is important because it may be involved in the process of understanding the actions performed by others. This is supported by the discovery of another class of neurons (mirror neurons [11]) which not only fire when the monkey performs an action directed to an object, but also when the monkey sees another conspecific (or the experimenter in this case) performing the same action on the same object. Clearly knowing in advance the range of affordances given the object facilitates the interpretation of the observed gesture by constraining the space of possibilities to those suited for the context.

In the following sections we describe experiments showing the acquisition of some of the building blocks of this neural representation in a biomorphic artificial system. In the discussion we will finally review the connection between the experimental results and the present section.

3 The robotic setup

The work presented here was implemented on the Babybot, a humanoid torso with a 5 degree of freedom (dof) head, a 6 dof arm and a 5 finger hand. The robot has two cameras which can independently pan and tilt around a common axis. The head has two further dof providing additional pan and tilt movements to the neck. The arm is an industrial manipulator mounted horizontally as illustrated in Figure 1. Previous works on Babybot have addressed the problem of orienting the head toward visual as well as auditory targets [12, 13], the development of reaching behavior [14] and the use of visual and vestibular information for visual stabilization [15]. Attached to the arm end point is a 5 finger robotic hand. Each finger has 3 phalanges; the thumb can also rotate toward the palm. Overall the number of degrees of freedom is hence 16. Since for reasons of size and space it is practically impossible to actuate the 16 joints independently, only six motors were mounted on the palm. Two motors control the rotation and the flexion of the thumb. The first and the second phalanx of the index finger can be controlled independently. Medium, ring and little finger are linked mechanically thus to form a single virtual finger controlled by the two remaining motors. No motor is connected to the fingertips; they are mechanically coupled to the preceding phalanges in order to bend naturally as shown in Figure 3. The mechanical coupling between gears is realized by

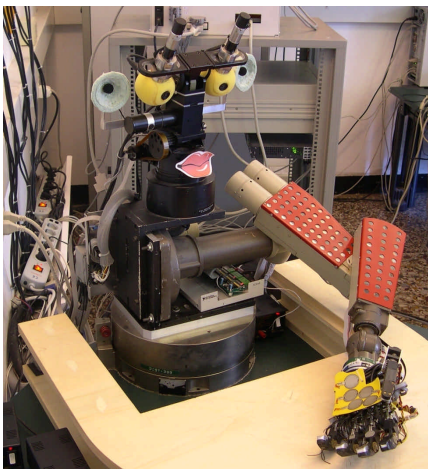


Figure 1 The robotic setup the Babybot.

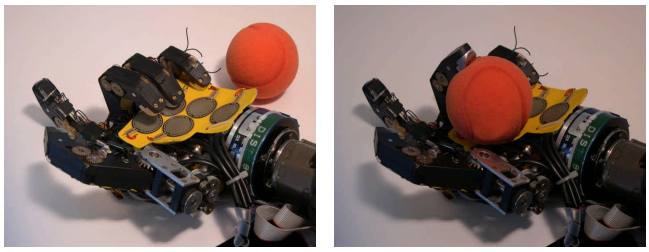


Figure 2 Elastic shape adaptation.

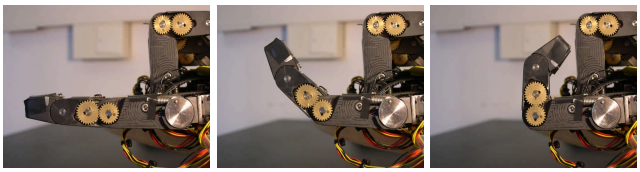


Figure 3 Mechanical coupling of the fingertips.

springs. This has the following advantages:

- The action of the external environment (the object the hand is grasping) can result in different hand postures (see Figure 2).
- Low impedance, intrinsic elasticity. Same motor position results in different hand postures depending on the object being grasped.
- Force control: by measuring the spring displacement it is possible to gauge the force exerted by each joint.

Hall-effect encoders at each joint measure the strain of the hand's joint coupling spring. This information jointly with that provided by the motor optical encoders allows estimating the posture of the hand and the tension at each joint. In addition, force sensing resistor (FSRs) sensors are mounted on the hand to give the robot tactile feedback. These commercially available sensors exhibit a change in conductance in response to a change in pressure. Although not suitable for precise measurements, their response can be used to detect contact and measure to some extent the force exerted to the object surface. Five sensors have been placed in the palm and three in each finger (apart from the little finger, see Figure 2).

4 The experiment

In this case the robot does not yet explore the world by actively reaching for objects but grasps toys that either are placed in the palm or touch the fingers. Since the robot has no knowledge about the object to be grasped tactile sensors are used to elicit a clutching action every time the hand is touched. Whenever pressure is applied to the fingers the hand closes by using a predefined motor command (synergy). The fingers stop when the maximum torque value – e.g. the motor error in the controller – exceeds a certain threshold for a certain amount of time (Figure 4).

Objects in a set are randomly chosen and given to the robot; the robot closes the hand and after a certain amount of time the grasp is released. The motor action does not change from trial to trial; owing to the intrinsic elasticity of the joints the action of the object on the fingers is exploited to adapt the hand to the target of the grasp. For each grasp the posture of the hand reflects the physical size of the object; the corresponding joint angles are then fed to a self-organizing map (SOM).

Initially we employed a set of 6 objects with different shapes (see Figure 5 left). The condition where no object is actually placed in the hand was included in the experiment. For each object about 30 grasp actions were performed; the result of the clustering is reported in Figure 5 (right). The network in this case had two layers with 15 units each (total of 225 neurons).

For each input pattern we report the unit which was activated the most on the 15x15 grid; different markers are used for different objects.

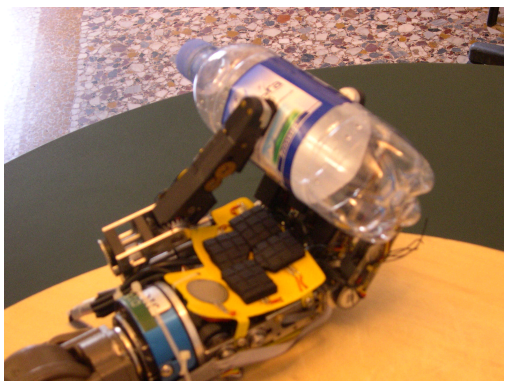


Figure 4 A picture of the hand grasping an object**.**

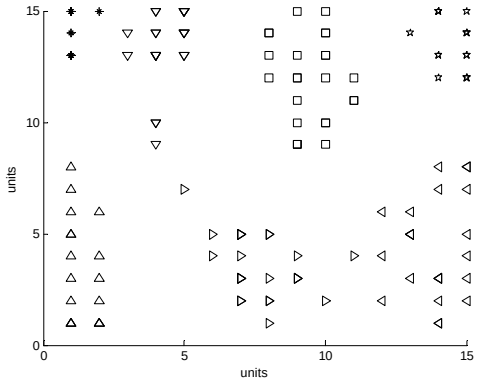


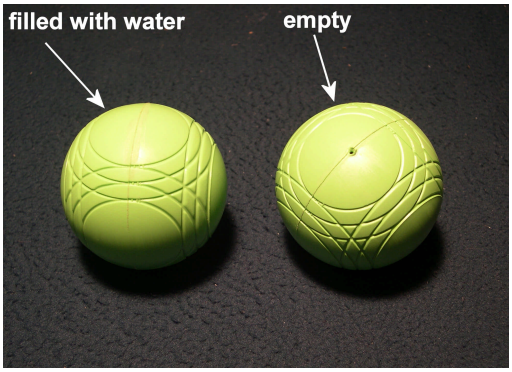
Figure 5 Experiment 1. Left: 6 objects were used, a bottle, a brick, a rod, a wooden ball, a small tennis ball made of foam rubber and a small plastic bowl. Right: result of the clustering. 6 classes are formed, one for each object plus one for the no-object condition. The map shows the grid of units (15x15), markers correspond to the neuron which resulted activated the most when a particular input pattern was applied; different markers correspond to different objects.

The SOM forms 7 clusters, each for a different object plus the no-object condition. Although some objects were quite different in terms of shape, the two small spheres – the plastic bowl and the tennis ball – had almost the same size. These two objects were nonetheless correctly separated by the SOM; this is due to the fact that the tennis ball is softer than the rigid plastic covering of the bowl. As the fingers bend around the soft object they slightly squeeze it thus creating a different category.

A second experiment was carried out with two object having identical shape and size, but different weight. At this purpose we used two plastic bowls, one of which filled with water to increase its weight (Figure 6, left). The hand is oriented upwards, the palm facing the ceiling, so gravity affects the force exerted by the fingers during grasp. The robot grasped each object about 60 times and the collected information was used by the SOM. In this case, since two objects were used, the network consisted only in 25 units (two layers of 5 neurons each). The result of the clustering reported in Figure 6 (right) shows that the network is able to separate the two set as being originated from different objects. As the two spheres have exactly the same size, the capacity of the network to categorize the input patterns is due to the fact that the fingers apply different forces; the hand posture thus implicitly code objects' weight.

5 Conclusions

We described two experiments where the robot uses its hand to explore physical properties of objects drawn from a set. Objects are placed in the palm or between the opposing fingers; the grasping action is elicited by pressure either on the palm or on the fingers. We showed that given the specific design of the hand, and very little prior knowledge, the robot is able to collect some physical features of the objects it receives. A self organizing map was employed to



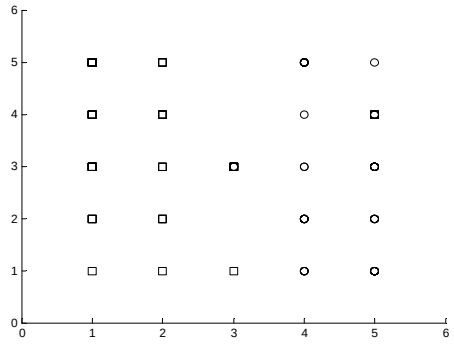


Figure 6 Experiment 2. Left: two identical sphere of different weight were used. Right: result of clustering. Markers represent the unit which was activated the most for each input pattern. Different markers correspond to different objects. In this case touch sensors were not used.

categorize the postural information obtained from the grasping. The clustering is not surprising in itself, being just a natural result of the mechanical design of the hand (the elasticity components connecting the joints) and the motor synergy exploited by the robot. Nevertheless the network implicitly codes not only physical features like shape (that in principle could be visually extracted) but also intrinsic properties like weight. Other physical features, like the object's compliance, might facilitate recognition. However we believe that the results are important; they prove that an active, embodied system can easily solve problems that otherwise would be hard (in the case of the balls of similar size), or even impossible (like in the case of the two identical small bowls having different weight).

The experiment as it is does not employ visual information yet, but it is not hard to conceive possible ways to include it. Visual parameters like color and shape (central moments) could be extracted from the objects and included in the network input vector.
The resulting representation would then link together the appearance of the object with the haptic information acquired during previous grasps.

The implications of this unified visuo-haptic representation may be twofold: improve recognition of objects and control of preshaping before actual grasping. In the first case although object recognition is based on visual cues only, haptic information can help to disambiguate in cases where vision is illusive (e.g. the distance-size ambiguity). In the second case motor information could be used to improve grasp stability by anticipating the posture of the hand during reaching according to the size and weight of the object to be grasped (preshaping).

Finally, physical properties like softness, weight and texture extend the internal representation of objects and allow generalizing their use based on their affordances. In fact by learning the effect of repetitive actions on different objects it is possible to identify important regularities between their physical properties and the way they behave when acted upon This ability to group different objects according to their possible use is a necessary step toward a truly cognitive system [8, 13].

Acknowledgments

The work described in this paper has been supported by the EU Projects ADAPT (IST 2001-37173), COGVIS (IST-2000-29375) and MIRROR (IST-2000-28159).

References

1. Rizzolatti, G. and M.A. Arbib, *Language within our grasp.* Trends in Neurosciences, 1998. **21**(5): p. 188-194.
2. Gallese, V., et al., *Action recognition in the premotor cortex.* Brain, 1996. **119**: p. 593-609.
3. Gibson, J.J., *The theory of affordances*, in *Perceiving, acting and knowing: toward an ecological psychology*, R. Shaw and J. Bransford, Editors. 1977, Lawrence Erlbaum: Hillsdale. p. 67-82.
4. Jeannerod, M., *The Cognitive Neuroscience of Action*. Fundamentals of Cognitive Neuroscience, ed. M.J. Farah and M.H. Johnson. 1997, Cambrige, MA and Oxford UK: Blackwell Publishers Inc. 236.
5. Coehlo, J., J. Piater, and R. Grupen, *Developing haptic and visual perceptual categories for reaching and grasping with a humanoid robot.* Robotics and Autonomous Systems, 2001. **37**: p. 195-218.
6. Natale, L., Rao Sajit, and G. Sandini. *Learning to act on objects*. in *Second International Workshop, BMCV 2002*. 2002. Tubingen, Germany: Springer.
7. Fitzpatrick, P., et al. *Learning About Objects Through Action: Initial Steps Towards Artificial Cognition*. in *IEEE International Conference on Robotics and Automation (ICRA 2003)*. 2003. Taipei, Taiwan.
8. Metta, G. and P. Fitzpatrick, *Early Integration of Vision and Manipulation.* Adaptive Behavior, 2003. **11**(2): p. 109-128.
9. Ballard, D.H. and C.M. Brown, *Principles of Animate Vision.* Computer Vision Graphics and Image Processing, 1992. **56**(1): p. 3-21.
10. Jeannerod, M., *Object Oriented Action*, in *Insights into the Reach to Grasp Movement*, K.M.B. Bennet and C. U., Editors. 1994, Elsevier Science. p. 3-15.
11. Fadiga, L., et al., *Visuomotor neurons: ambiguity of the discharge or 'motor' perception?* Internation Journal of Psychophysiology, 2000. **35**(2-3): p. 165-177.
12. Metta, G., *Babybot: a Study on Sensori-motor Development*, in *DIST*. 2000, University of Genova: Genova. p. 176.
13. Natale, L., G. Metta, and G. Sandini, *Development of Auditory-evoked Reflexes: Visuo-acoustic Cues Integration in a Binocular Head.* Robotics and Autonomous Systems, 2002. **39**(2): p. 87-106.
14. Metta, G., G. Sandini, and J. Konczak, *A Developmental Approach to Visually-Guided Reaching in Artificial Systems.* Neural Networks, 1999. **12**(10): p. 1413-1427.
15. Panerai, F., G. Metta, and G. Sandini, *Learning Stabilization Reflexes in Robots with Moving Eyes.* Neurocomputing, 2002. **48**(1-4): p. 323-337.