Exploring the world through grasping: a developmental approach*

Lorenzo Natale, Francesco Orabona, Giorgio Metta, Giulio Sandini
LIRA-Lab, DIST
University of Genoa
Viale Causa 13, 16145 Genoa, Italy
{nat, bremen, pasa}@liralab.it, sandini@unige.it

***Abstract* – This paper is about the implementation of grasping skills in a humanoid robot. Following a developmental approach the robot is initially equipped with little perceptual and motor competencies whose role is to bootstrap learning through the exploration of the external environment. This crude form of sensorimotor coordination consists of a set of control systems and explorative behaviors as well as simple visual routines. The developmental path leads the robot from the exploration of the physics and geometry of its body to the probing of the external environment. The robot experience builds and modifies continuously its internal representations of the environment, being this its body or the objects it encounters. We discuss the implications of our approach to the study of cognition and our effort to build a cognitive artificial system.**

Index Terms – Humanoid robotics, development, cognitive systems.

I. INTRODUCTION

There is a growing scientific and practical interest to the study of cognitive systems being these biological (to analyze them) or in trying to replicate their capabilities (engineering them). Unfortunately though, there is not even an accepted definition of what a cognitive system is. Lying at the two ends of the spectrum we find cognitivism [1] on one side and enactive systems [2, 3] on the other. According to the first, cognition is a computational process carried out on a symbolic representation of the world. Symbols represent the world and can be shared across different entities: they are a complete characterization of the world in which the entity is located, and as such are independent of the entity itself and its past experience. In enactive approaches instead, cognition is the result of the interaction and co-development of the agent's body and the environment in which the agent lives. Although the definitive answer is still to be found, physical realizations of artificial systems can prove to be a valid test-bed for different theories and models.

In the debate we see our approach more sympathetic toward the enactive system's end. Two aspects seem thus crucial: i) embodiment and ii) development. The two requirements are obviously intertwined, as the interaction between the agent and the environment is possible only by means of a body. As a consequence, representations of the environment depend on the particular embodiment and, more importantly, on the past experience of the system. It is worth noting that representation here is taken in its most neutral and common-sense meaning.

To see how extreme it could be, consider for example, the visual system of primates. This is efficient only as long as the animal moves the eyes to place the high-resolution fovea at interesting places in the environment. Attention regulated by a sophisticate control system allows only specific stimuli to be selected among the wealth of possible choices. Through action, the system's embodiment and the environment codetermine the resulting representations.

Motivated by these considerations we propose a developmental approach to the implementation of cognitive abilities in a humanoid robot. We identified the minimum embodiment as possessing a head, arm, and hand. Although limited, this robot can perform goal-directed actions on objects, like reaching and grasping, it can visually scan the environment, and actively explore it by taking forceful actions.

A grossly simplified though plausible developmental process can be described in three distinct phases – they do not necessarily need to be completely separated. The first stage is devoted to learning the internal models of the robot's body (know also as the *body-map*) which eventually provides basic motor and perceptual skills like gaze control, eye-head coordination and reaching. Based on these abilities the interaction with the external world is investigated in the second phase where the robot discovers properties of objects and ways to handle them (learning to interact). The robot actively explores the environment by taking actions (e.g. reaching) which allows starting the acquisition of information about the physical coherence and the functioning of the environment and, at the same time, discovering new, more efficient, ways of interaction (for example different grasp types). Finally the third stage is about learning to understand and interpret dynamical events possibly including other agents: the robot has associated its actions with the resulting perceptual consequences. Interpretation of other people's actions is achieved by inverting this association.

In the past we have addressed aspects related to the third phase [4, 5]. This paper is focused only on the first two phases.

II. EXPERIMENTAL SETUP

The experiments setup consists of the robotic platform shown on Fig 1. It is an upper torso humanoid robot

* Funding for the research described in this paper has been provided by the EU projects ADAPT (IST-37173) and ROBOTCUB (FP6-004370).

composed of a 5 degree of freedom (DOF) head, 6 DOF arm and a 6 DOF hand. The head has 5 DOF, two of which control the neck pan and tilt, whereas the other three actuate two cameras to pan independently and to tilt on a common axis. The arm is a Unimate PUMA 260, an industrial manipulator with 6 degrees of freedom; it is mounted with the shoulder horizontal (typically vertical) to better mimic the human kinematics. The hand consists of 5 fingers; each finger has three phalanges, the thumb has an additional degree of freedom which allows it to perform a rotation toward the palm. Overall the hand has 16 joints controlled by only six motors. Two motors are connected to the index fingers: they are linked to the first (proximal) and second phalanges. The distal (small) phalange is mechanically coupled to the preceding one so that the two bend together (see Fig 1). Two motors control the motion of medium, ring and little finger. As in the case of the index finger, the proximal phalanges are linked to the first motor, while the second and third phalanges are actuated by the second motor. The mechanical coupling between the joints is realized by means of springs to allow a certain amount of adaptation of the grasp type to the object shape.

From the point of view of the sensors, the head is equipped with two cameras and two microphones for visual and auditory feedback. Tactile feedback is provided by 17 force sensing resistors mounted on the hand, 5 of which are placed on the palm and the remaining 12 evenly distributed on the thumb, index, medium and ring fingers. A JR3 force sensor provides torque and force feedback at the wrist. Further proprioceptive information is provided to the robot by optic and magnetic encoders mounted on all joints of the head, arm and hand.

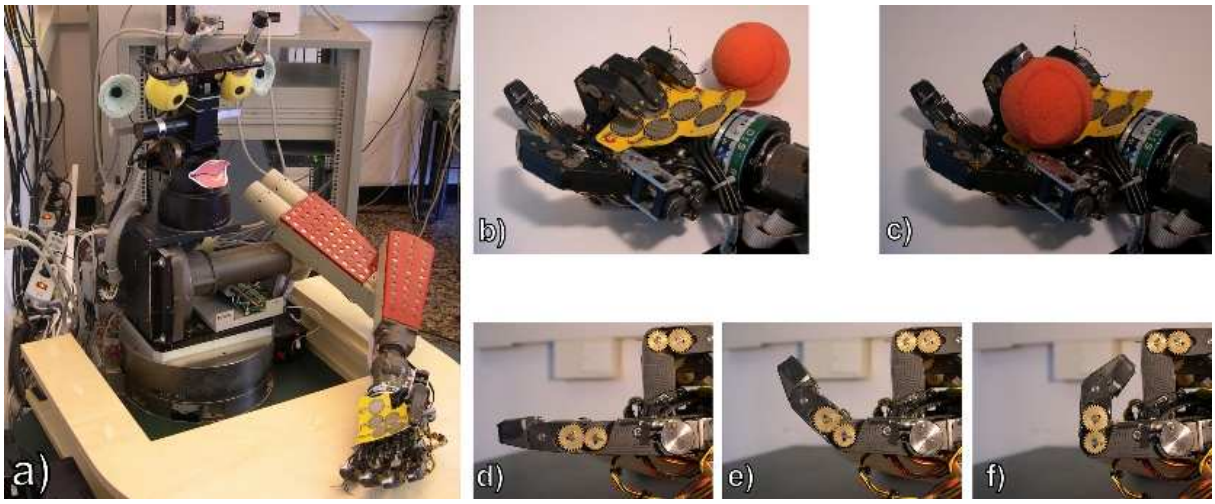

Fig 1: a) the experimental setup, the Babybot. Left: details of the hand. b) and c): elastic compliance. d)-f): mechanical coupling between phalanges.

III. VISION

One of the first steps of any visual system is that of locating suitable interest points in the scene ("salient regions" or events) and eventually direct gaze toward these locations. Human beings and many animals do not have a uniform resolution view of the visual world but rather only a series of snapshots acquired through a small high-resolution sensor (e.g. our fovea). This leads to two questions: i) how to move the eyes efficiently to important locations in the visual scene, and ii) how to decide what is important and, as a consequence, where to look next.

There is accumulating biological evidence that attention is directed to an object or a group of objects, and that the attention system processes properties of objects, rather than regions of space. In fact, it has been shown that selective attention frequently operates on an object-based representational medium in which the boundaries of segmented objects, and not just their spatial position, determine what is selected and how attention is deployed (see [6] for a review). In other words, the visual system seems optimized for segmenting complex three-dimensional scenes into representations of (often partly occluded) objects for recognition and action. Indeed, perceivers must eventually interact with objects in the world and not with disembodied abstract locations.

The literature on visual attention in artificial vision is vast and several models have been proposed [7, 8]; most of them are derived from Treisman's Feature Integration Theory (FIT) [9]. The FIT model employs a set of low-level feature maps which are combined together by a spatial attention window operating in a master saliency map.

The visual attention model we propose starts from approximately the same type of considerations but then uses a concept of salience based on *proto-objects* defined as blobs of uniform color in the image. The robot by acting on objects can then figure out how to combine these proto-objects into coherent wholes, i.e. full-fledged object. Once an object is grasped, in fact, the robot can move and rotate it and build a statistical model of the color blobs and their spatial relationship. This internal representation can subsequently be used to instantiate the top-down component of the attention system.

Throughout the paper we employed log-polar images as defined for example in [10], which mimic the distribution of

cones, i.e. the photoreceptors of the retina involved in diurnal vision. Cones have a higher density in the central region called *fovea*, while they are sparser in the periphery. We never used standard rectangular images.

As a first step the input image is applied a smoothing, by taking the average between the current frame and the output of the color quantization (see later in text) on the previous frame (see Fig 2). Then the red, green, blue channels of each image are separated, and the yellow channel is calculated as the average of the red and green one. These four channels are combined to generate three color opponent channels, similar to those of the retina. Each of these channels, typically indicated as (R+G-, G+R-, B+Y-), has center-surround receptive field (RF) with spectrally opponent color responses. That is, for example, a red input in the center of a particular RF increases the response of the channel R+G- while a green one in the surrounding decreases its response. The spatial response profile of the RF is expressed by a Difference-of-Gaussians (DoG). A RF centered on each pixel of the input image is considered, so the output of the RF filtering is obtained with a convolution with DoG kernel, generating an output image of the same size of the input. The DoG filters are not balanced and similarly to what happens in the human retina the unbalanced ratio implicitly code achromatic information.

Edges are then extracted on the three channels separately by employing a generalization of the Sobel filter due to [11]. The resulting edge maps are combined together to generate a single map. A watershed transform [12] is then applied on the edge map to isolate blobs and to generate *proto-objects*. As a result the image is segmented in blobs with either uniform color or uniform gradient of color.

At this point each blob is tagged with the mean color of the pixels within its internal area (this leads to a color-quantized image). The result is blurred with a Gaussian filter and stored: it will be averaged with the sequent frame to obtain a temporal filtering and reduce the effect of noise.

As discussed above, it is known that a feature or stimulus is salient if it differs from its immediate surrounding area. The bottom-up salience is calculated as the Euclidean distance in the color opponent space between each blob and its surroundings. Moreover, the size of the spot or focus of attention is not fixed but rather linked to the size of the blobs. In the same way the definition of *immediate surrounding area* is relative to the size of the focus of attention. In other words, we compute the salience of each blob in relation to a neighborhood region whose size is proportional to that of the blob itself. In our implementation we use a rectangular region 2.5 times the size of the bounding box of the blob. Blobs that are too small or too big are discarded from the saliency computation and ignored and they will not be considered as possible candidates to be part of objects (proto-objects).

The top-down influence on attention is calculated in relation to the visual search task. When the robot has acquired a model of the object and begins searching for it, it uses the knowledge of the object's appearance to bias the saliency map. In practice, the top-down saliency map is computed as the distance between the average color of each blob and that of the target:

$$S_{top-down} = \sqrt{\left(\langle R^+G^-\rangle_{blob} - \langle R^+G^-\rangle_{object}\right)^2 + \left(\langle G^+R^-\rangle_{blob} - \langle G^+R^-\rangle_{object}\right)^2 + \left(\langle B^+Y^-\rangle_{blob} - \langle B^+Y^-\rangle_{object}\right)^2} \quad (1)$$

where $\langle\ \rangle$ indicates the average of the image values over a certain area (indicated as subscripts). The total salience is simply estimated as the linear combination of the two terms above:

$$S = \alpha S_{top-down} + \beta S_{bottom-up} \quad (2)$$

The total salience map S is eventually normalized in the range 0-255, as a consequence the salience of each blob in the image is relative to the most salient one.

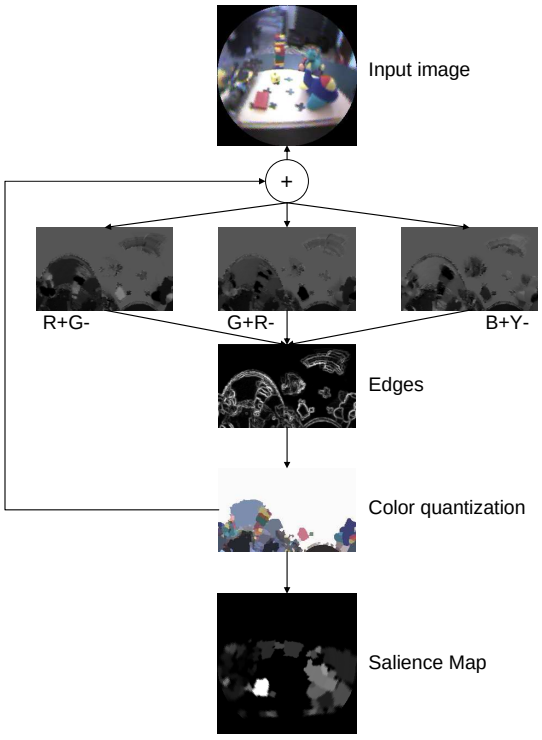

Fig 2: The visual attention system, block schema.

Local inhibition is transiently activated in the salience map. This prevents the focus of attention to be redirected immediately to a location that was previously attended. Such an *inhibition of return* (IOR) has been also demonstrated in human psychophysics.

Our system implements a simple object-based IOR. The robot maintains a list of the last five positions visited, coded in a body centered coordinate system. When the robot redirects gaze it keeps memory of the blobs it has visited. Inhibition occurs only if the blob presents the same color of that stored in the list; in case the object moves or its color changes the location becomes available again for fixation.

IV. LEARNING ABOUT THE SELF

In general the set of the internal models that represent the body is called a *body-schema* or body-map: it involves, for example, the relative position of the limbs, the weight of the body segments and their size. The existence of a body-schema in the brain has gained some support thanks to the work of Graziano [13, 14] and Rizzolatti [15] who found

neurons in the primate motor cortices coding the position of the hand in the visual field irrespective of the relative position of the head and hand.

In biological systems the internal representation of the body is shaped during development and maintained fit to the physical modification occurring in life. In artificial systems (where the body does not normally change with time) adaptation can avoid the tedious operation of manually tuning the system's internal models and their parameters. The latter might be required to compensate changes in the visual appearance of the body or drift in the sensors (e.g. motor encoders).

We propose here an approach similar to Fitzpatrick and Arsenio [16] and Metta and Fitzpatrick [17]. Repeated, self-generated actions are initiated by the robot during the learning phase. In particular we controlled the robot to execute a periodic movement of the wrist. The resulting motion of the hand was detected by computing image difference between the current frame and an adaptive model of the background. The period of motion of each pixel in the resulting motion image was then computed with a zero-crossing algorithm; similar information was extracted on the proprioceptive feedback of each motor encoders. As a result the hand of the robot was segmented by selecting, among the pixels that moved periodically, those whose period matched that of the joints. Conversely non-periodic moving pixels or pixels moving with different periods were identified as being originated by other sources and discarded. Fig 3 shows an example of the detection for two different pixels whose motion was (a) correlated and (b) uncorrelated with that of the robot's hand.

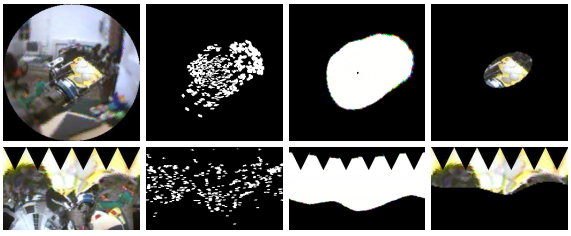

Fig 3: An example of the detection procedure. From left to right: the original image at the beginning of the procedure, the result of the detection, the result of a further low-pass filtering, and the final segmentation.

This information about the hand is employed to learn the position of the hand in the visual field given the current arm and head posture and, simultaneously, the shape and orientation of the hand (in this case represented by an ellipse). Learning is carried out by using a neural network specifically designed for online learning [18]. Eventually these models could predict the position of the robot's hand in the image given a future position of the head and arm.

Although of some utility, the real goal of segmenting the hand is for guiding reaching. The solution we propose is based on the use of a direct mapping between the eye-head motor plant and the arm motor plant [19]. In other words, reaching for an object starts by looking at it. Under this assumption, the fixation point can be seen as the "end-effector" of the eye-head system. The position of the eyes with respect to the head, determines uniquely the position of the fixation point in space relative to the shoulder. The arm motor command can be obtained by a transformation of the eye-head motor/positional variables. We called this approach *motor-motor coordination*, because the coordinated action is obtained by mapping motor variables into motor variables:

$$q_{arm} = f(q_{head}) \quad (3)$$

where q_{head} and q_{arm} are head and arm posture respectively (joint space). What is interesting in this approach is not equation (3) *per se*, which, after all, implements the inverse kinematics of the arm, but the mechanisms we use to learn it. In fact this mapping can be easily learnt if the robot can look at its hand, which is not incidentally the by product of knowing the position of the hand in the image.

The robot explored the workspace by moving the arm randomly, while at the same time, it tracked the hand; whenever the head fixated the hand a new pair arm-head posture was acquired and used as a training sample to the neural network that approximates *f* in equation (3). When a sufficient number of samples were acquired (and the network trained), the robot started using the mapping to reach for visually identified objects.

Once the robot has computed the final arm posture it is still required to plan the actual movement. This was done with a simple linear interpolation between the current and final arm configuration. The trajectory was divided in steps which were effected by the low-level controller; to this purpose we employed a low-stiffness PD controller with gravity compensation. The gravity load term for each joint was also learnt online as described in [20].

V. BUILDING OBJECT MODELS

Either by reaching and randomly grasping or because of a cooperating peer the robot eventually will grasp an object. Both solutions are valid bootstrapping behaviors for the acquisition of an internal model of the object. When the robot holds the object it can explore it by moving and rotating it. In short, the idea is to represent objects as collections of blobs as generated by the visual attention system and their relative position (neighboring relations). The model is created statistically by looking at the same object several times from different points of view. The system estimates the probability for each blob to belong to the object by counting the number of times each blob appears during this exploration phase.

We used the probabilistic framework proposed by Schiele and Crowley [21]. It consists of estimating the probability of the object *O* given a certain local measurement *M*. This probability *P(O|M)* can be calculated using Bayes' formula:

$$P(O \mid M) = \frac{P(M \mid O)P(O)}{P(M)} \quad (4)$$

$$O_{MAP} = \arg\max_{O,\sim O} \{P(O \mid M), P(\sim O \mid M)\}$$

where: *P(O)* is the *a priori* probability of the object *O*, *P(M)* the *a priori* probability of the local measurement *M*, and *P(M|O)* is the probability of the local measurement *M* when the object *O* is being fixated. In the following experiments we carried out only a single detection experiment per object;

there are consequently only two classes, one representing the object and another representing the background. *P(O)* and *P(~O)* are simply set to *0.5* since this choice does not affect the maximization of equation 4.

Since a single blob is not discriminative enough, we considered the probabilities of observing pair of blobs, i.e. the local measure *M* is the event of observing both the central and a surrounding blob:

$$P(M \mid O) = P\left(B_i \mid B_c \text{ and } \left(B_i \text{ adjacent } B_c\right)\right) \quad (5)$$

where *Bi* is the *i-th* blob surrounding the central blob *Bc* which belongs to the object *O*. We exploit the fact the robot is fixating the object and assume *Bc* to be constant across fixations. In practice this corresponds to estimating the probability that all blobs *Bi* adjacent to *Bc* (which we take as a reference) belong to the object. This procedure, although requiring the "active participation" of the robot (through gazing) is less computationally expensive compared to the estimation of the probability for all possible pairs of blobs of the fixated object. Estimation of the full joint probabilities would require a larger training set than the one we used in our experiments. The probabilities *P(M|~O)* are estimated during the exploration phase considering all the blobs not adjacent to the central blob. The local measurements are independent, because they refer to different blobs, so the total probability $P(M_1,\ldots,M_n|O)$ can be factored into the product of the probabilities $P(M_i|O)$. An object is detected if the probability $P(O|M_1,\ldots,M_n)$ is greater than a fixed threshold.

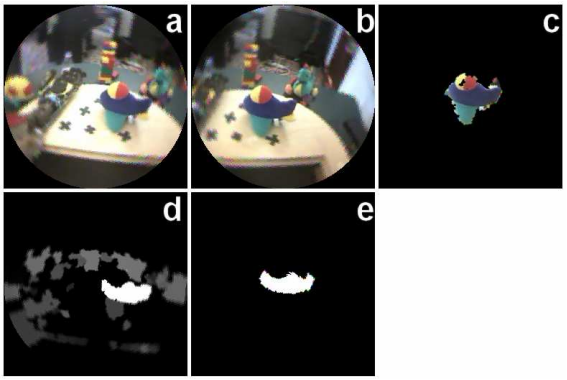

Figure 4: Visual search. The robot has acquired a model of a toy airplane during the exploration phase (not reported); this information primes the attention system which assigns a high saliency to the blue blob at the center of the airplane. A saccade is performed, the object is foveated. (a) and (b) show the visual scene before and after the saccade. (d) and (e) show the result of the attention system at the same instant of time. The result of the segmentation after the saccade is in (c).

When an object is detected after visual search, a possible figure-ground segmentation is attempted, using the information gathered during the exploration phase. Each blob is segmented from the background if it is adjacent to the central blob and if its probability to belong to the object is greater than *0.5*. This probability is approximated using the estimated probability above with the following approximation:

$$P\left(B_i \in O \mid B_c \text{ and } \left(B_i \text{ adiacent } B_c\right)\right) \simeq P\left(B_i \mid B_c \text{ and } \left(B_i \text{ adiacent } B_c\right)\right) \quad (6)$$

Fig 4 shows the result of the segmentation procedure.

VI. DISCUSSION AND CONCLUSION

To recapitulate, we have shown how two phases of autonomous development could be crafted into a humanoid robotic system. It is important to note that the combination of the various components was designed in by the experimenter and not acquired by the robot. Still each of the components showed some component of learning. Fig 5 shows this exemplar behavior through a sequence of pictures taken from the robot's point of view.

The action starts when an object is placed in the robot's hand and the robot detects pressure in the palm (picture 1). This elicits a clutching action of the fingers; the hand follows a preprogrammed trajectory, the fingers bend around the object toward the palm. If the object is of appropriate size, the intrinsic elasticity of the hand facilitates the action and the grasping of the object. The robot moves the arm to bring the object close to the cameras and begin the object exploration. The object is placed in four different positions (as for instance in pictures 2 and 3). During the exploration phase the robot tracks the hand/object; when the object is stationary and fixation is achieved, a few frames are acquired and the model of the object is constructed. At the end of the exploration the object is released (picture 4).

At this point the robot has acquired the visual model of the object and starts searching for it in the visual scene. To do this, it selects the blob whose features better match those of the object's main blob and perform a saccade. After the saccade the model of the object is matched against the blob that is being fixated and its surrounding. If the match is negative the search continues with another blob, otherwise grasping starts (pictures 7-8-9). At the end of the task the robot uses simple haptic information to detect if it is holding the object or rather the action failed. In this process the weight of the object and its consistence is checked (proprioception from the fingers holding the object). If the action is successful the robot waits for a new object to start again, otherwise it performs another reaching-grasping trial.

It is fair to mention that part of the controller for this experiment was preprogrammed. For example, the hand was controlled with stereotyped motor commands. Three primitives were used: one to close the hand after a certain amount of pressure was detected on the palm, and two during grasping to pre-shape the fingers and actually seize the object. The robot relied on the elasticity of the hand to achieve the correct grasping. To facilitate grasping, the trajectory of the arm was also programmed beforehand; waypoints relative to the final position of the arm were included in joint space to approach the object from above.

In spite of this limitations, we have presented results on two important phases of the acquisition of sensorimotor coordination in a relatively sophisticate humanoid robot. We have shown the implementation of a visual attention system employing top-down and bottom-up information. More importantly, we demonstrated how the robot can actively explore the visual appearance of the objects it happens to grasp. This information is also fed to the attention system as a bottom-up primer to control the search of the object. Thus

the robot experience allows building a representation of the objects it interacts with while, at the same time, modulating the attention system. The robot's ability to act is used together with the body internal model to drive the exploration of the environment. This facilitates learning in different ways. At first, it helps the system to focus attention in both space and time. During the acquisition of the object visual model, in fact, the robot can track the object because it knows the position of the hand from its proprioceptive feedback. Proprioception is also useful to detect when the acquisition of the model can be initiated since the object motion is under direct control by the robot. Finally the fact that the object is being held by the hand guarantees the link between different sensory modalities (for example the sight of the object and the kinesthetic information from the hand). The object model makes use of visual information; in [20] we show instead how it is possible to build a model of the object based on haptic information only. In the future we would like to investigate possible ways to integrate the two approaches.

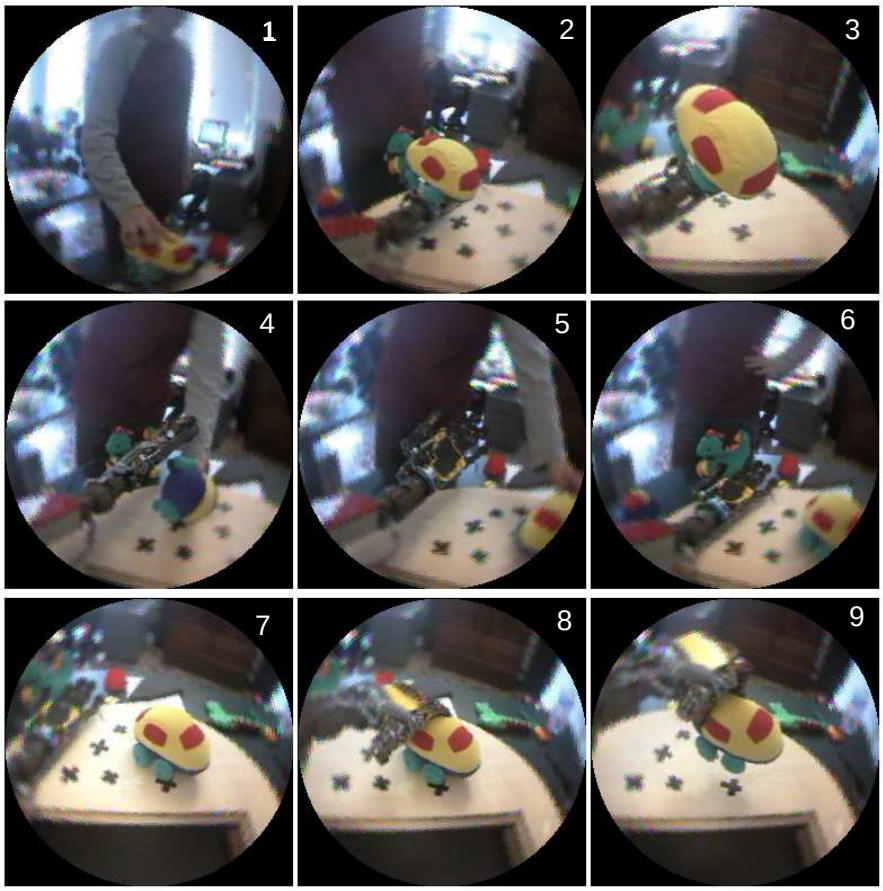

Fig 5: A sequence of the robot grasping an object. The action starts when an object is placed on the palm (1). The robot grasps the object and moves the eyes to fixate the hand (2). The exploration starts in (3) when the robot brings the object close to the camera. The object is moved in four different positions while maintaining fixation; at the same time the object model is trained (3-6). The robot drops the object and starts searching for it (7). The object is identified and a saccade performed (7-9). The robot eventually grasps the toy (10-12).

This work supports the view of *cognition* emerging from the embodied interaction between the system and the environment. Cognition requires a body and the ability to autonomously build the representation of the external world through this interaction. Even a simple set of behaviors has been sufficient to bootstrap the exploration of the environment and the acquisition of a representation of it. We have shown how this initial interaction is sufficient to start linking action with its consequences to form prediction about the behavior of the body and the environment.

Very often prospective control is required to plan a successful action. During grasping, for example, the correct timing of preshaping and closure of the fingers is required; the lags in the sensory processing (visual and tactile) typical of artificial and natural systems make feedback control ineffective. To be able to anticipate the impact of the hand with the object is required to control the timing between preshaping and actual grasping without relaying on visual and tactile feedback. Prospective control, however, is not only important for action. It gives the agent the possibility

to create expectations on which to base the interpretation of the world and the actions performed by others. Through interaction with the world the agent builds a model of how the entities involved behave and what is the resulting sensorial consequence. This link can be used afterward to anticipate the consequence of a similar action and, eventually, compare it with the real feedback.

In the same way new situations can be interpreted by matching them against the agent's past experience. For example the event of a ball that falls on the floor (and the resulting visual and auditory sensations) can be associated to the action of dropping it. Anticipation and prediction enhance the agent's ability to understand and interact with the environment and, for this reason, are important aspects of cognition. The results of this paper are the first necessary steps into the effort of developing cognitive abilities in an artificial system.

REFERENCES

[1] A. Newell and H. A. Simon, "Computer science as empirical inquiry: Symbols and seach," *Communications of the Association for Computing Machinery*, vol. 19, pp. 113-126, 1975.

[2] R. H. Maturana and F. J. Varela, *The tree of knowledge, the biological roots of human understanding*, Revised Edition ed. Boston & London: Shambhala Publications, Inc., 1998.

[3] R. Beer, D., "Dynamical approaches to cognitive science," *Trends in Cognitive Sciences*, vol. 4, pp. 91-99, 2000.

[4] L. Natale, Rao S., and G. Sandini, "Learning to act on objects," presented at Second International Workshop, BMCV 2002, Tubingen, Germany, 2002.

[5] P. Fitzpatrick, G. Metta, L. Natale, S. Rao, and G. Sandini, "Learning About Objects Through Action: Initial Steps Towards Artificial Cognition," presented at IEEE International Conference on Robotics and Automation (ICRA 2003), Taipei, Taiwan, 2003.

[6] B. J. Scholl, "Objects and attention: the state of the art," *Cognition*, vol. 80, pp. 1-46, 2001.

[7] R. Milanese, "Detecting Salient Regions in an Image: From Biological Evidence to Computer Implementation," in *Department of Computer Science*. Geneva: Univerity of Geneva, 1993.

[8] L. Itti, C. Kock, and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254-1259, 1998.

[9] A. M. Treisman and G. Gelade, "A feature integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97-136, 1980.

[10] 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.

[11] X. Li, T. Yuan, N. Yu, and Y. Yuan, "Adaptive color quantization based on perceptive edge protection," *Pattern Recognition Letters*, vol. 24, pp. 3165-3176, 2003.

[12] P. D. Smet and R. Pires, "Implementation and analysis of an optimized rainfalling watershed algorithm," presented at IS&T/SPIE's 12th Annual Symposium Electronic Imaging 2000: Science and Technology, Conference: Image and Video Communications and Processing, San Jose, California, USA, 2000.

[13] M. S. A. Graziano, "Where is my arm? The relative role of vision and proprioception in the neuronal representation of limb position," *Proceedings of the National Academy of Science*, vol. 96, pp. 10418-10421, 1999.

[14] M. S. A. Graziano, D. F. Cooke, and C. S. R. Taylor, "Coding the location of the arm by sight," *Science*, vol. 290, pp. 1782-1786, 2000.

[15] G. Rizzolatti and M. Gentilucci, "Motor and visual-motor functions of the premotor cortex," in *Neurobiology of Neocortex*, P. Rakic and W. Singer, Eds. Chichester: Wyley, 1988, pp. 269-284.

[16] P. Fitzpatrick and A. Arsenio, "Feel the beat: using cross-modal rhythm to integrate perception of objects, others and self," presented at Fourth International Workshop on Epigenetic Robotics, Genoa, 2004.

[17] G. Metta and P. Fitzpatrick, "Early Integration of Vision and Manipulation," *Adaptive Behavior*, vol. 11, pp. 109-128, 2003.

[18] S. Schaal and C. G. Atkeson, "Constructive Incremental Learning from Only Local Information," *Neural Computation*, pp. 2047-2084, 1998.

[19] G. Metta, G. Sandini, and J. Konczak, "A Developmental Approach to Visually-Guided Reaching in Artificial Systems," *Neural Networks*, vol. 12, pp. 1413-1427, 1999.

[20] L. Natale, G. Metta, and G. Sandini, "Learning haptic representation of objects," presented at International Conference on Intelligent Manipulation and Grasping, Genoa, Italy, 2004.

[21] B. Schiele and J. L. Crowley, "Probabilistic object recognition using multidimensional receptive field histograms," presented at 13th International Conference on Pattern Recognition, Vienna, Austria, 1996.