Proceedings of the 2003 IEEE International Workshop on
Robot and Human Interactive Communication
Millbrae, California, USA, Oct. 31 - Nov. 2, 2003

Embodied Vision - Perceiving Objects from Actions

Artur M. Arsenio
MIT AILab
200 Technology Square, Room NE43-936
Cambridge, MA02139, USA
arsenio@ai.mit.edu

Abstract

For the purposes of visual manipulation of objects by a robot, the latter has to learn about object properties, as well as actions that the robot may apply on it. This paper presents strategies to acquire such competencies based on human-robot interactions. Perception is driven by manipulation from an actor, either human or robotic. Interaction with human teachers facilitates robot learning of new objects and their functionality, or the acquisition of new competencies as an actor. Self-exploration of the world extends the robot's knowledge concerning object properties, and consolidates the execution of learned tasks.

1 Introduction

Embodied vision [2] consists of boosting the vision capabilities of an artificial creature by fully exploiting the concepts of an embodied agent situated in the world. Although biological organisms are particularly good at solving visual tasks, this is still a challenge in computer vision. In active vision [4] the percepts can be acquired in a purposive way by the active control of a camera. This approach has been successfully applied to problems such as stereo vision, by dynamically changing the baseline distance between the cameras, or by active focus selection.

We argue for solving a visual problem by not only actively controlling the perceptual mechanism, but also and foremost actively changing the environment through experimental manipulation. Scene features that can be modified by a robotic agent (using for example a robotic arm) or a human teaching a robot, are the image spatial content, the motion (the optical flow computed from the image), the frequency content, the color content, or even the structure of a scene in an image. In previous work [1], we explored interactions with an agent to create information concerning the grouping of color regions that form an object. This strategy was valid even for objects drawn on books or heavy, fixed objects like a sofa. This paper exploits instead changing motion content at several frequency/spatial scales, as described in Section 2.

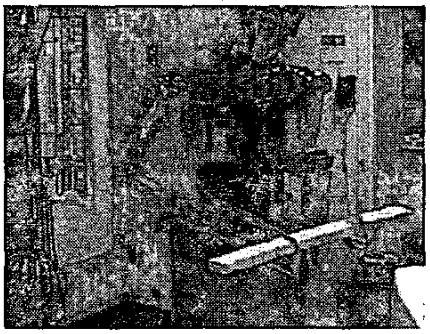

Figure 1: *The humanoid robot Cog has two, six-degree-of-freedom arms. Each joint is driven by a series elastic actuator [10]. This compliant arm is designed for human/robot interaction, but they are not reliable for the execution of precise motions.*

Visual Tasks The work presented in this paper was developed on the humanoid robot Cog, shown in Figure 1 sawing a piece of wood using neural oscillators to control the rhythmic movements, [10]. However, according to [10], the robot did not know how to grab the saw or the underlying task sequence. The neural oscillator parameters need to be inserted off-line, using a trial-and-error approach. An automatic approach for selecting the parameters was proposed in [3]. But it remains necessary to recognize the object - a saw, identify it with the corresponding action, and learn the sequence of events and objects that characterize the task of sawing (described in Section 3). Furthermore, a general strategy should apply to any object the robot interacts with, and for any task executed, such as hammering or painting. Other research work [8] described robotic tasks such as a robot juggling or playing with a devil stick (hard tasks even for humans). However, the authors assumed off-line specialized knowledge of

0-7803-8136-X/03/$17.00©2003 IEEE

the task, and simplified the perception problems by engineering the experiments.

This paper describes an embodied approach for learning task models while simultaneously extracting information about objects. Through social interactions of a robot with an instructor (see Figure 2), the latter facilitates robot's perception and learning, in the same way as human teachers facilitate children perception and learning during child development phases. The robot will then be able to further develop its competencies (as introduced in Section 4), and to learn more about objects (Section 5) by itself, using simple actions such as poking.

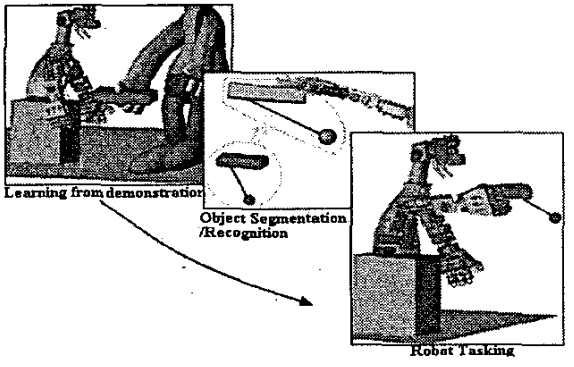

Figure 2: *At initial stages, object properties (e.g. appearance, size or shape) are unknown to the robot. As a human teacher manipulates objects, the robot build object models and learn the teacher actions. Robot learning may then continue autonomously.*

2 Perception Driven by Action

The world surrounding us is full of information. A critical capability to an intelligent system is filtering the information that is relevant to a task. This capability is developmentally acquired by humans. A fundamental question is then *How to Select the Relevant Information?*

We argue for both humans and the robot to take an active role on this selection. A human teacher is able to describe objects to the robot using his arm, hand or finger as an actuator. Indeed, primates have specific brain areas to process the hand visual appearance [7].

Infants are genetically predisposed to select skin-tone stimulus as soon as their first minutes of life, and they start early in life drawing attention to objects under periodic motion. It has also been shown that infants prefer bright-colored objects [6]. Working by these principles, the humanoid robot Cog was equipped with an attentional system that actively selects attentional targets, by building an attentional spatial map. This map receives data from basic feature maps: skin-tone maps, color saliency maps and motion maps. This may be viewed as a pre-filter stage.

Whenever attention is draw to an object, the robot then selects relevant temporal/spatial information. For simplicity, the head is kept steady throughout operation, so that image differences can be used for motion detection at higher frame rates than optical flow. Periodic motions with a high frequency content over a narrow portion of the frequency spectrum are extracted, since repetitive action patterns are indicators of good sources of information.

Vision at Multiple Spacial/Temporal Scales. Time information is lost by transforming a signal to the frequency domain. Indeed, it is not possible to detect time events on the Fourier transform of a signal.

However, signals generated by most interesting moving objects and actors contain numerous transitory characteristics, such as the occurrence of events: abrupt changes of the object velocity; appearance of a new object on a scene (or disappearance of an existing one); positional contact or separation between objects; and correlations among the position and velocity signals of multiple objects. These characteristics are often the most important part of the signal, and Fourier analysis is not suited to detect them.

The Short-Time Fourier Transform (STFT) maps a signal into a two-dimensional function of time and frequency. It represents a sort of compromise between the spatial and frequency based views of a signal, with the drawback of a fixed window size, which remains the same for all frequencies. Gabor filters and wavelets overcome this limitation by varying the window size at multiple scales. The approach is this paper applies STFTs with window sizes varying in a logpolar fashion. Long spatial intervals (and hence small window sizes) provide more precise spatial, local information, while larger windows increase frequency resolution.

Newton's Third Law: Action-Reaction. As just described, spatial information is lost when converting a signal into its frequency content. Therefore, it remains imperative to detect events localized over the temporal sequence. In the field of video compressing, standard systems do not store all information in a sequence of images. Instead, only changes with a significant content over a stationary scene are stored at each frame. From this perspective, good spatial event candidates are moving regions of the image that change velocity abruptly under contact, as illustrated by Figure 3. For instance, if a robot pokes an object,

the reaction to the action exerted by the robot causes the object to move.

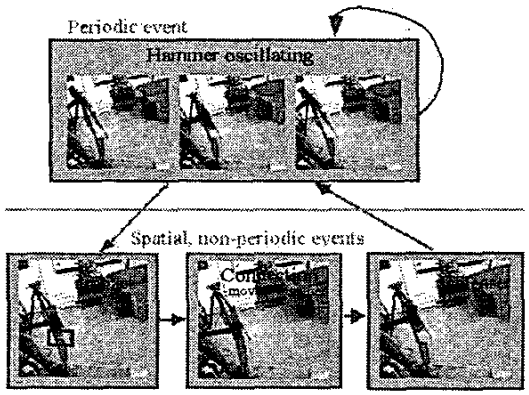

Figure 3: *Hammering task. This task is both characterized by periodic motions, and also by spatial (time) events. Through observation of the task being executed, the robot should be able learn the sequence of events that compose a task, as well as the objects involved.*

Developmental Learning. Although a human can interpret visual scenes perfectly well without acting on them, such competency is acquired developmentally by linking action and perception. Actions are not necessary for standard supervised learning, since off-line data is segmented manually. But whenever an actor has to autonomously acquire object categories using its own body to generate informative percepts, actions become indeed very useful.

3 Perceiving Actions on Objects

This Section presents the algorithms for identifying events at multiple spatial/frequency resolutions.

3.1 Spatial Domain Events

The algorithm to identify and track multiple objects in the image is described herein. A motion mask is first derived by subtracting gaussian filtered versions of successive images and placing non-convex polygons around any motion found. A region filling algorithm is applied to separate the mask into regions of disjoint polygons (using a 8-connectivity criterion). Each of these regions is used to mask a contour image computed by a Canny edge detector. The contour points are then tracked using the Lucas-Kanade pyramidal algorithm. An affine model is built for each moving region from the position and velocity of the tracked points. Outliers are removed using the covariance estimate for such model.

Four categories of spatial events will be hereafter defined. Whenever an event occurs, the objects involved in such event are inserted into a short term memory, with a span of two seconds after the last instant the object moved. Multiple image regions are tracked. A region is tracked if it is moving or it maps into an object in memory.

Implicit Contact. This class of events is originated by such actions as grabbing a stationary object, or assembling it to a moving object. It corresponds also to the initial event of poking/slapping a stationary object. The velocity of the stationary object rises abruptly from zero under contact with the actor actuator (human arm/hand or robotic arm/grip) or moving object. This creates a sudden expansion of the optical flow (because both the actor and the object move instantaneously together). Therefore, this event is triggered by:

- ▷ velocity of static object rises abruptly (flow history kept for 0.5 seconds)
- ▷ the are of the actuator's motion flow grows abruptly
- ▷ actuator velocity is large
- ▷ connection maintained over two consecutive frames

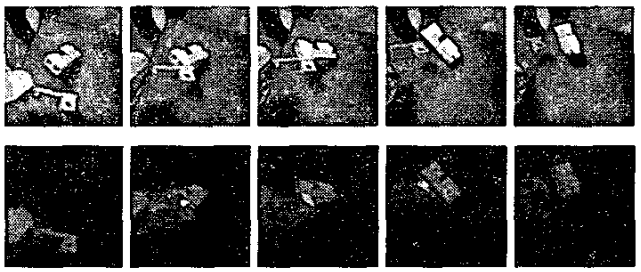

Figure 4: *Top row: sequence of images of the robotic arm approaching, impacting and releasing and object. This exploratory mechanism enables the extraction of objects' shape. Notice that two human legs are also moving in the background. Bottom row: Objects tracked over the sequence. An implicit contact occurs at the 2^{nd} frame from the left, upon impact of the robot's arm with the object. Another event (an explicit release) occurs on the 5^{th} frame.*

Figure 4 shows the detection of such an event for an object poked by the robot. Object segmentations for both the robot's grip and the poked object are shown in Figure 5.

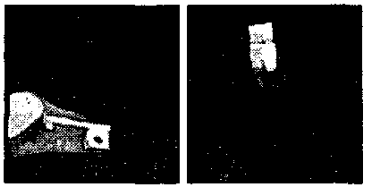

Figure 5: *Object segmentations for the robot's arm and the poked object, respectively. Most of the experiments were performed under strong shadows that move with the objects and actuators. But since a large corpora of segmentations is obtained for each object (one segmentation is extracted per each frame if object is moving), the average reduces effects from shadows and noise.*

Implicit Release. An object and the actor's actuator, or two objects assembled together, might be moving rigidly attached to each other. But without further information, the robot perceives them as a single object. Current work is being developed to filter the actor actuator using the learned actuator segmentations - see human and robot actuator segmentations in Figures 5 and 7, respectively. An event occurs upon the object release by the actuator, either by throwing or dropping it. This event may also result from a disassembling operation between two objects. For such cases, hierarchical information concerning object structure is available to an object recognition scheme.

The conditions required for the detection of this class of events are:

- ▷ the motion flow of the assembled region separates into disjoint regions
- ▷ the initial velocity of the ensemble is large
- ▷ the posterior velocities of the separated entities (objects and actuators are entities) are both large
- ▷ separation is maintained over two consecutive frames

The occurrence of this event is illustrated in Figure 6, which shows the separation event for the robot grip and an infant toy moving initially together.

Explicit Contact. This event is triggered by two moving entities coming into contact, such as two objects crashing into each other or an actuator slapping or grabbing (see Figure 7) a moving object:

- ▷ the two entities' templates overlap for two consecutive frames
- ▷ the a priori velocities of both moving entities are large (flow history kept for 0.5 seconds)

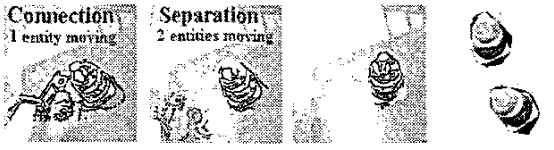

Figure 6: *Sequence showing the robot releasing an object (edges superimposed to image, which was made brighter for visualization). An explicit release event occurs at the 2^{nd} frame from the left. Initially (first frame), both the robot's grip and the object are moving together as a single object. The last frame shows two object segmentations sampled randomly from the large set of acquired segmentations. The actuator shadow is perfectly visible on the 2^{nd} frame.*

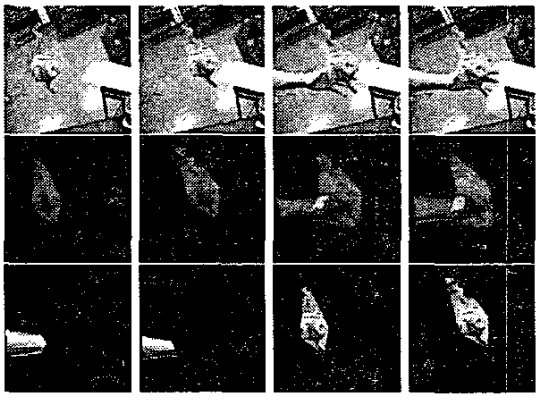

Figure 7: *Top row: a pendular object oscillating is grabbed by a human, and both stop moving after. Middle row: tracked regions - the contact region is also signaled. Bottom row: The two frames on the left present human arm segmentations, while the two on the right show two of the object segmentations.*

Explicit Release. The detection of this event is similar to the previous one, differing on the trigger condition: two moving entities loosing contact. The algorithmic conditions require that the entities' templates become disjoint over two consecutive frames, and the velocity of one of the entities change abruptly. A detection of such an event is shown in Figure 4.

3.2 Frequency Domain Events

The periodic detection is applied at multiple scales. Indeed, for objects oscillating during a short period of time, the movement might not appear periodic at a coarser scale, but appear as such at a finer scale. If a strong periodicity is not found at a larger scale, the window size is halved and the procedure is repeated again for each half.

Initially, a grid of points homogeneously sampled

from the image are initialized in the moving region and tracked thereafter over a time interval of approximately 2 seconds (65 frames). The motion trajectory for each point over this time interval is determined using the Lucas-Kanade pyramidal algorithm. Their trajectory is evaluated using a Fast Fourier transform, and tested for a strong periodicity.

Periodicity is estimated from a periodogram determined for all signals from the energy of the FFTs over the spectrum of frequencies. These periodograms are processed by a collection of narrow bandwidth bandpass filters. Periodicity is found if, compared to the maximum filter output, all remaining outputs are negligible. Figure 8 shows both the FFTs of signals filtered out and periodic signals.

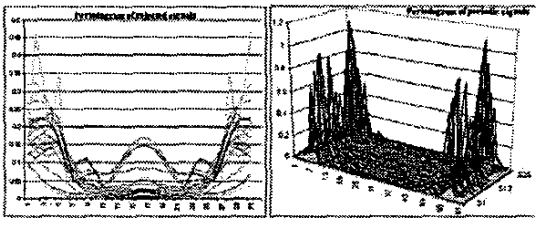

Figure 8: *The left graph shows the FFTs of discarded points, while the right image displays the FFTs for the set of points on the saw.*

Figure 9 shows the detection of a periodic event for the task of sawing. A large corpora of templates for the saw is obtained from the object segmentation procedure.

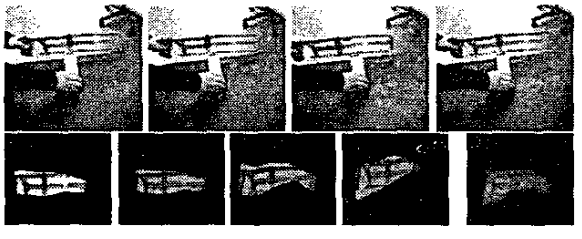

Figure 9: *The top row shows a sequence of images for a human sawing. Bottom row: First frame shows one of the more than 400 segmentations obtained in less than a minute of interactions. The* 2^{nd}, 3^{rd} *and* 4^{th} *frames show segmentations averaged over 64 templates. Last frame shows the average template for all the segmentations. This object is particularly hard to segment. Although in some cases shadows appear, and in other cases the object is not complete, in average the right shape is obtained.*

As knowledge concerning objects is build up, the task of stimulus selection will be simplified. In case an object is detected in the frequency domain, its stored location and appearance are then used by the spatial domain detector for locating that object.

4 Robot Tasking

A framework novelty is the fast production of a large corpora of data, which is essential for learning. The robot ability to categorize objects grows developmentally as a human teacher introduces the robot to new objects, and the robot explores them by itself.

4.1 Learning from Demonstration

A human teacher can communicate with the robot through a collection of frequency and spatial events that it can cause, with the goal of presenting objects or tasks to the robot. Figure 10 illustrates the task of hammering a nail (frame 4 shows an explicit contact, while an explicit release event occurs in frame 5.) The algorithm also detected hammer oscillations during task execution.

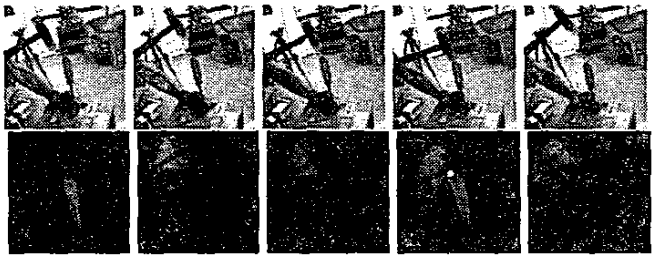

Figure 10: *Human hammering a nail with a hammer. Besides the hammer and the nail, the arm that secures the nail also moves, but does not create an event.*

4.2 Identification of Tasks using Hybrid State Machines

A task is defined as a collection of events on objects occurred while an actuator is visible or while significant parts of a scene are moving. Tasks are described by continuous states (such as a hammer oscillating or a hammer moving connected to a nail), and discrete transitions among them.

An hybrid state machine provides an adequate representation for tasks. Figure 11 shows a diagram of the hybrid state machine obtained for the task of hammering a nail in Figure 10. Tasks involving different objects or events are represented by different machines.

Hybrid Markov Chains. Future work will address state machines with stochastic discrete transitions, instead of deterministic ones. The probability of occurrence of one discrete transition will be determined from a collection of data obtained from executing the same task several times.

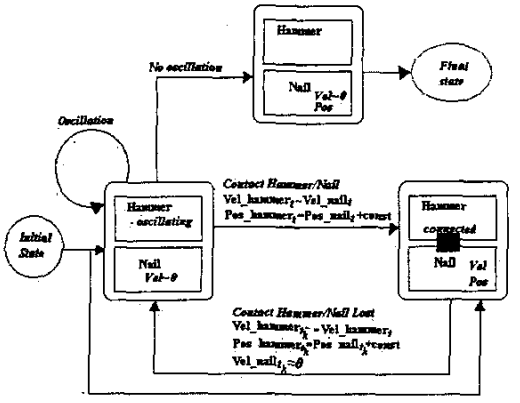

Figure 11: *Hammering task as an hybrid state machine. The task goal is represented as the final state.*

4.3 Robotic Manipulation Of Objects

Simple actions such as poking, grabbing or throwing can be learned using Section 4.1 framework (as well as more complex actions). Thereafter, the robot can use this knowledge to act on objects by itself.

Experimental results were presented throughout the paper for the poking of objects by the robot (at the moment, the robot's grip is not yet set up to grab objects). Even for such a crude and simple action such as poking, reliable object segmentations are possible for both the robotic arm's grip and the object poked, as shown in Figure 5.

Interactions among objects - Perceiving mass relations. The previous experiments included an actor. But events, as supported by Figure 12, are also detected between objects. This is especially useful to determine ensemble properties and mass relations among objects. Mass relations may be perceived as suggested in [6], according to intuitive physics. Indeed, once determined the center of mass' velocities, the perceived mass relations (upon a perspective transformation) are obtained from the momentum conservation law (ignoring friction). Therefore, the mass of the car is perceived as much larger than the ball's mass - and indeed the ball has a smaller mass. Figure 13 shows the two object segmentations at the contact instant.

5 Object Segmentation through Human - Robot Interactions

Focus was placed on a fundamental problem in computer vision - *Object Segmentation* - which was dealt with by detecting and interpreting natural human/robot task behavior such as waving, shaking,

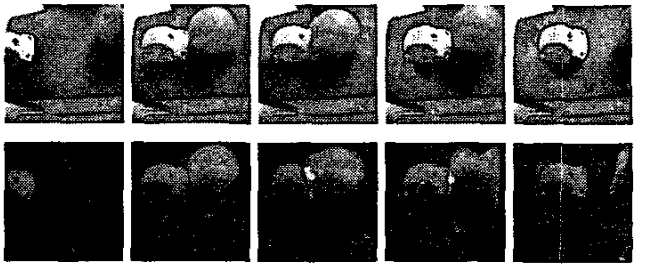

Figure 12: *A car impacts into a ball. This experiment only includes objects, no actor's actuator is present.*

Figure 13: *Segmentations for both the car and the ball at the contact instant.*

poking, grabbing/dropping or throwing objects. Object segmentation on unstructured, non-static, noisy and low resolution images is indeed a hard problem, as exemplified by Figure 14:

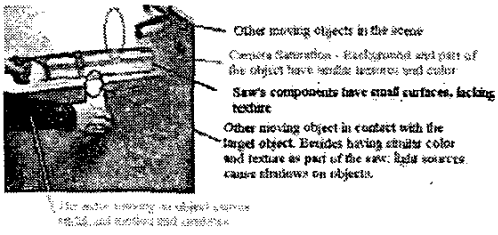

Figure 14: *The results presented in this paper were obtained with varying light conditions. The environment was not manipulated to improve natural occurring elements, such as shadows or light saturation from a light source. All the experiments were taken while a human or the robot were performing an activity.*

- ▷ object may have similar color/texture as background
- ▷ multiple objects moving simultaneously in a scene
- ▷ robustness to luminosity variations
- ▷ Real-time, fast segmentations
- ▷ Low resolution images (128×128 images for real-time)

A nearly real-time algorithm was developed for object segmentation. After the detection of an event, clusters of points moving coherently are available. The algorithm covers these points by n overlapping windows. Each image region is covered by four of such windows. A convex polygon algorithm is used to approximate all the moving points in each window, while windows with less than a minimum number of pixels are eliminated (this removes outliers). The union

Figure 15: *Segmentations for the objects in the hammering task. A failed segmentation is also shown.*

of all such convex polygons produces a collection of non-convex polygons. This algorithm is much faster than the minimum cut algorithm [9], and provides segmentation of similar quality to the active min-cut approach in [5]. A random sample of segmentations are presented from the hammering task in Figure 15. Figure 16 provides additional results.

The human ability to segment objects is not general-purpose, improving with experience. Object segmentation is a key ability on top of which other capabilities, such as object/function recognition are being developed. The framework here described is currently being further extended to estimate object kinematics/dynamics within the goal of task identification.

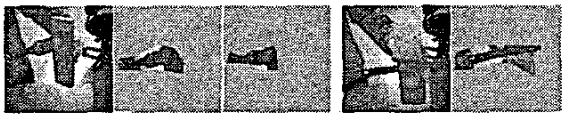

Figure 16: *Object segmentations. A large set of segmentations was obtained from several tasks.*

6 Conclusions and Discussion

The number of visual segmentation techniques is vast. Methods for characterizing the shape of an object through tactile information are also subject of extensive research, such as shape from probing or pushing. Although has been known for long that motor strategies can aid vision [4], work on active vision has focused mainly on moving cameras. An embodied approach to vision that exploits the robot's own body or a human body to actively act on the environment was described. In addition, a framework was developed for the identification of actions on real world objects, from movements (performed by a human, a robot or another object) with a strong periodic or spatial content. Objects were segmented from scenes where real-time tasks were being performed (e.g. hammering).

The robot is able to rapidly segment a wide variety of objects from low-resolution images, with varying conditions of luminosity, and multiple moving artifacts in the scene. A novelty of the approach is the detection of events at multiple time scales for a better compromise between frequency and spatial resolution. Future work will address the integration of the robot's control mechanisms for the execution of more complex tasks.

In this paper we introduced both humans and the robot into the learning loop to facilitate robot perception. The experiments carried out underline the importance of learning to interpret actions. There is a lot to be gained when introducing a human teacher in the loop that guides the robot's learning process. This knowledge is grounded as the robot acts by itself.

Acknowledgments Project funded by DARPA as part of the "Natural Tasking of Robots Based on Human Interaction Cues" under contract number DABT 63-00-C-10102, and by the Nippon Telegraph and Telephone Corporation as part of the NTT/MIT Collaboration Agreement. Author supported by Portuguese grant PRAXIS XXI BD/15851/98. Thanks to Paul Fitzpatrick for valuable discussions, and setting-up the robot's poking control.

References

[1] A. Arsenio, P. Fitzpatrick, C. Kemp and G. Metta, *The Whole World in Your Hand: Active and Interactive Segmentation.* Accepted for publication at the 2^{nd} conf. on Epigenetic Robotics, 2003

[2] A. Arsenio, *Boosting vision through embodiment and situatedness.* MIT AILab Research Abstracts, 2002

[3] A. Arsenio, *Designing Neural Oscillator Networks for Rhythmic Control.* From Animals to Animats, MIT-Press, 2000

[4] D. Ballard, *Animate vision.* In Artificial Intelligence, 48(1), 57, 1986

[5] P. Fitzpatrick. *From First Contact to Close Encounters: A developmentally deep perceptual system for a humanoid robot.* PhD thesis, MIT, 2003

[6] S. Palmer, *Vision Science.* MIT Press, 1999.

[7] D. Perrett, A. Mistlin, M. Harries and A. Chitty, *Understanding the visual appearance and consequence of hand action.* In Vision and action: the control of grasping, pages 163-180. Ablex, NJ. 1990

[8] S. Schaal and C. Atkeson, *Robot Juggling: Implementation of Memory-Based Learning.* In IEEE Control Systems, vol. 14, no.1, 57-71, 1994

[9] J. Shi and J. Malik, *Normalized cuts and image segmentation.* In IEEE Trans. on Pattern Analysis and Machine Intelligence, 22:888-905, 2000

[10] M. Williamson, *Robot Arm Control: Exploiting Natural Dynamics,* Ph.D. Thesis, MIT, 1999