# Multi-modal Stereognosis*


Luiz M. G. Gonçalves[1,2], Roderic A. Grupen[1], and Antonio A. F. Oliveira[2]

[1]Laboratory for Perceptual Robotics - Dept of Computer Science
University of Massachusetts (UMASS), Amherst MA 01003 USA
(lmarcos,grupen)@cs.umass.edu

[2]Laboratory for Computer Graphics - COPPE/Systems
Federal University of Rio de Janeiro (UFRJ), CP 68511, Rio de Janeiro, RJ 21945-970
oliveira@lcg.ufrj.br


## 1 Introduction

In this work, vision and touch (artificial) senses are integrated in a cooperative active system. Multi-modal sensory information acquired on-line is used by a robotic agent to perform real-time tasks involving categorization of objects. The visual-touch system proposed is able to foveate (verge) the eyes onto an object, to move the arms to touch an object, and to choose another object by shifting its focus of attention. Also, the system can detect changes occurred on a previously visited region. We propose a hybrid computational system in which an associative memory remembers the visual and tactile signature of objects for recognition and reinforcement learning tasks (Q-learning) are formulated to learn active sensing policies. Several vision architectures have been proposed recently. Kosslyn's architecture [6], based on results in experimental neuro-psychology, seems to be technically possible with special hardware. Despite practical issues it is intuitively attractive to attach computational mechanisms to that architecture. Ferrell [3] uses registered, multi-modal, topographically organized maps of the sensory-motor space to orient a robot (COG) towards environmental stimuli. In our approach we also use topographically organized (visual and arm) maps for feature extraction. But, instead of a pixel (patch) representation we have regions of interest (ROIs) segmented in those maps, as suggested in Kosslyn [6]. At each time, only one region is selected to be processed by the high level processes. This selective aspect provides a substantial reduction in the amount of computational efforts necessary for feature extraction.

## 2 System Control Architecture

We have developed a control architecture (Figure 1) based on recent control theory approaches [1, 2, 4] and motivated by biology and developmental psychology [7, 5, 6, 8]. Basically, input from each sensory system is organized into spatially-indexed areas (ROIs) constituting a perceptual (visual and arm) buffer. The attention window allows the access to the information provided by a single ROI. A pre-attentive mechanism computes activation values (involuntary sensory cues) for the ROIs. The attention-shift mechanism is responsible for selecting the most active ROI and effectively for focusing the attention window to it. After the eyes/arms controllers get converged onto a chosen ROI, features like intensity, texture, shape, color, spatial location, orientation, size, and weight, are extracted only from a ROI currently in the attention window and then matched into a long-term address by the associative memory. If one pattern activation (object representation that has properties closer to those of a perceived object) is over a threshold the object is identified, otherwise more information shall be provided. In this situation an improving or modeling action is suggested. That action can be an eye movement enhancing any particular characteristic or an arm movement touching, moving, or grasping the object. If after all information gathering actions there is no match for an input, a new or unknown object has been discovered. The new information must be stored in the long term memory and the associative memory updated with the corresponding new pattern representation (automatic supervised learning is activated).

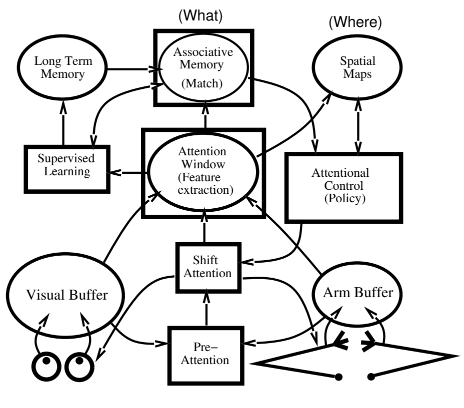


Figure 1: System Control Architecture.

## 3 The Simulation Environment

The simulated robot "Roger-the-Crab" (Figure 2) has five controllers (neck, two eyes and two arms) integrated in a single platform. A Phong illumination model with a Gaussian noise process calculate intensity values for each pixel in Roger's retinas. Also, a compensation for gravity and other ambient effects and basic robot kinematic and dynamic equations are determined to be used by the servo controllers of an arm or an eye. Independent controllers run concurrently for each eye and each arm. A back-propagation


*This work is supported by NSF under IRI-9503687, IRI-9704530, and CDA-9703217, and by CNPQ/Brazil



To appear in the Proceedings of III International Conference on Autonomous Agents (Agents '99). Seattle WA, May of 1999.


©1999 ACM.

neural network implements the associative memory. A discrete polar coordinate system is used for the eye/arm maps.

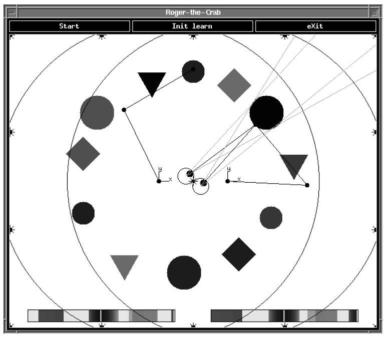


Figure 2: Roger the Crab Interface.

## 4 A Control-loop using Q-learning

To better orchestrate the system, besides a simple straightforward strategy, we have generated a behavioral policy using Q-learning [9]. Figure 3 shows a finite state machine in which the Q-learning is performed. Each state is relative to a property of the selected ROI, like *novelty*, *identity*, and *mapping-status*. *Identity* is set to 1 if a match is detected in the associative memory and *novelty* is set to 1 if no match occurs after all possible trials (a new object). The last state (*mapping-status*), is set to 1, for a ROI, when that ROI becomes the focus of attention. The actions refers to the physical movements of the arms/eyes, and to a map-*updating* operation. This last operation is performed after an object is classified as identified or new. The physical actions are: a complete attention-*shift*, a visual-*modeling*, and arm-*modeling*. Rewards are given to the map-*updating* actions determining the transitions from state 100 to 101 and from 010 to 011. The result of the Q-learning is a Q-table which encodes all possible states and the best action to be chosen from each one.

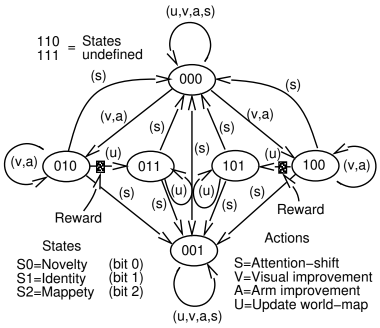


Figure 3: State space and actions.

## 5 Experimental Results

In the experimental tests, the two control policies have been applied to a task of surveillance and the results have been acceptable. Figure 4 shows the performance of the policy generated by the Q-learning in comparison with the simple strategy. Table 1 shows that the Q-learning approach performed more attentional shifts, fewer eyes/arms modeling actions, more positive identifications, and fewer new-object detections than the simple approach. As rewards are given for the map-*updating* action, even performing fewer new-object detections the Q-learning approach performed better overall. We have also performed other experiments running until completion (until all regions were visited). In those experiments all objects were visited (looked at by Roger), identified and mapped by the policies determined by both approaches. After all regions of interest are visited, the eyes remain in vigilant state.

| Appr | Att Shift | Eye/Arm modeling | Posit Ident | New Obj | Map Update |
|---|---|---|---|---|---|
| (1) | 72 | 76 | 29 | 27 | 56 |
| (2) | 70 | 82 | 20 | 32 | 52 |

Table 1: Numbers obtained after 204 control-loops using the Q-learning(1) and simple(2) approaches: attention shifts performed, modeling actions, objects positively identified, new objects detected, and objects settled in the maps.

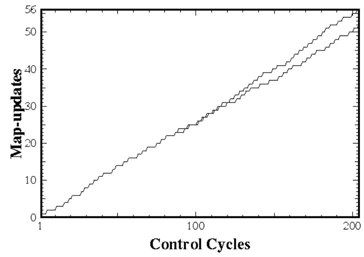


Figure 4: Performances. Upper line is Q-learning and lower line is simple approach.

## 6 Conclusion and Future Work

We have built an useful control mechanism which was successfully used by a multi-modal sensory system in the execution of surveillance tasks. The complexity of such an integrated system has been partially solved by (serial and selectively) processing regions of interest in the images. The multi-modal system is currently being implemented on a hardware platform composed of a Stereo-head, Datacube Image Processing architecture, two robotic arms with hands, and a host interface computer. The next step is to test this learning architecture on a real environment and to measure its performance in a variety of tasks. We believe that the top-down aspects of attention can also be explored by using the same system architecture.

## References


[1] E. Araujo and R. A. Grupen. Learning control composition in a complex environment. In *Proceedings of Int. Conf. on Simulation of Adaptive Behavior (SAB'96)*, Cape Cod, MA, September 1996.

[2] J. A. Coelho and R. A. Grupen. A control basis for learning multifingered grasps. *Journal of Robotic Systems*, 14(7):545–557, 1997.

[3] C. Ferrel. Orientation behavior using registered topographic maps. Tr, Massachusetts Institute of technology, Cambridge, MA, 1998.

[4] M. Huber and R. A. Grupen. Grupen. a feedback control structure for on-line learning tasks. *Robotic and Autonomous Systems*, 22(3-4):303–315, December 1997.

[5] B. Julesz. *Foundations of Cyclopean Perception.* University of Chicago Press, Chicago, IL, 1971.

[6] S. Kosslyn. *Image and Brain. The Resolution of the Imagery Debate.* The MIT Press, Cambridge, MA, 1994.

[7] D. Marr. *Vision – A Computational Investigation into the Human Representation and Processing of Visual Information.* The MIT Press, Cambridge, MA, 1982.

[8] A. Streri. *Seeing, Reaching, and touching.* The MIT Press, Cambridge, MA, 1993.

[9] R. S. Sutton and A. G. Barto. *Reinforcement Learning: an Introduction.* The MIT Press, Cambridge, MA, 1998.