To appear in XI International Conference on Computer Graphics and Image Processing, Rio de janeiro, RJ, Brazil.

A Control Architecture for Multi-modal Sensory Integration*

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]Laboratório de Computação Gráfica - COPPE Sistemas
Universidade Federal do Rio de Janeiro (UFRJ), CP 68511, Rio de Janeiro, RJ 21945-970
(lmarcos,oliveira)@lcg.ufrj.br

Abstract. This work describes the architecture of an integrated multi-modal sensory (vision and touch) computational system. We propose to use an approach based on robotics control theory that is motivated by biology and developmental psychology, in order to integrate the haptic and visual information processing. We show some results carried out in simulation and discuss the implementation of this system using a platform consisting on an articulated stereo-head and an arm, which is currently under development.

Keywords. Cognition, multi-modal stereognosis, attention, seeing, reaching, grasping

1 Introduction

In this work, we propose the integration of vision and touch in a behaviorally cooperative active system. The multi-modal sensory information is used by a simulated robotic agent to perform real-time tasks. We also discuss some background and current research relating vision and touch sensing systems. Vision is undoubtedly the most powerful and useful but also the more complex sensory system. A natural way to relate both systems is to assume touch subordinate to vision, with the arms and hands acting based on visual information. But, in some cases touch is more important than vision, since tactile information can disambiguate visual information. In this work, the relative importance of touch and vision are assumed to be highly context dependent. They work in parallel, providing (ambiguous or complementary) information to a decision system, responsible for providing adaptive responses (actions) to the environmental stimuli.

A useful visual-touch system must be able to foveate (verge) the eyes onto an object, to subsequently move the arms to reach and grasp an object and to choose another object once the current object is identified, by shifting its focus of attention. To validate such a system, we have defined the basic task of surveilance in which a robot learns how to construct an incremental map of its environment, dealing with new or known instances of objects. In order to perform this surveilance task, the system must also recognize and identify the objects present in its environment. As described in [18] the time required for biological systems to recognize objects in a single eye fixation is about 20 milliseconds per object.

Following this introdution, we will discuss related work. Then, a proposed system for forming consistent spatial models from vision and touch is presented and a simulation environment "Roger the Crab" is described. Finally, results are presented and discussed, and performance evaluated.

2 Vision, Reaching and Grasping (Related Work)

We can find a wide range of articles and text-books in the literature about stereo reconstruction and its application in Computer Vision or Robotics (e.g. [1, 2, 3, 7, 8, 9, 10, 20, 21, 22, 23, 29]. A number of researchers have followed the Marr [1] and Marr/Poggio [7] paradigm. They propose a model that computes depth maps using correlation measures. It is quite difficult to design algorithms for real-time applications using standard (even pipeline) architectures using this paradigm. A fast stereo algorithm using correlation measurements is described in the work of Huber and Kortenkamp [5]. Using pipeline array processing, they achieve thrughput up to five frames per second. An area sign correlation algorithm for the stereo matching was proposed by Nishihara [3, 4] and is derived from Marr and Poggio's work. With this approach, a robot [5] can pursue a slowly moving human. This is impressive performance, but in normal situations and for normal velocities we need faster algorithms for the tracking and pursuit tasks. Neuro-physiological studies [30] show evidence that in some tasks biological systems achieve up to a dozen times that throughput. To achieve better throughput, Sanger [24] offers an approach based on biological models of the disparity computation. This

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

approach employs Gabor filters, local approximations of Fourier transformations, as the basis for disparity computation (see Freeman and Adelson [25] for details on steerable filters). This model uses the fact that the displacement of a function generates a proportional phase shift in its Fourier transform. The binocular disparity at each location is therefore proportional to the phase difference of the corresponding left and right image patches [23]. Recent research [20, 21, 22] proposes a new approach for the neural encoding of binocular disparity. The neurophysiological data supports two models for the selective disparity of simple and complex cells in the primary visual cortex [22]. These involve binocular combinations of monocular receptive fields that are shifted in retinal position (position shift model) or in phase (phase shift model) between the eyes. The results show that these models are a reasonable computational approximation for the biological disparity computation, but in practice they are as expensive as the one used in [5]. This is due to the amount of parallel computation necessary to represent the disparity sensitive neurons and also to carry out the (phase and position shifting) Gabor filter calculations.

Kosslyn [18], based on results of experimental neuro-psychology, suggests a vision system that seems to be technically possible with special hardware. Kosslyn does not specify algorithms nor does he provide mathematical definitions for the object properties, the coordinate systems, and spatial maps to be used in his system. Moreover, Kosslyn suggests that disparity, for use in stereo reconstruction, is computed in some unspecified way by low-level processes, referring to Marr's 2.5 D sketch [1, 7]. But, despite practical difficulties, this system seems to be closer to the biological system than others. It uses some well known structures and mechanisms, like the attention window, associative memory and attention-shift. It is intuitively attractive to attach computational mechanisms to Kosslyn's descriptive specification (Kosslyn architecture is "descriptive", not computational, biological, or neuro-physiological).

Robotic grasping has been studied extensively [11, 12, 13, 14]. In the work described in [14] by Coelho and Grupen, tactile and proprioceptive information are used to provide a robust strategy for grasping. The developed tool seems to be useful not only for grasping solutions but also for solving other complex tasks. Based on previous grasping experience, a simple model of expected performance is derived. At a given time, based on the actual state and on previous experience, the controller can select the most effective policy for the context on-line. This method can be applied to more general robot control tasks as well as recognition and attention.

Reaching is another problem in which biological experiments have inspired computational solutions. Most work has tried to reproduce (or imitate) infants abilities to reach for a presented stimulus. The work described in [19] discusses the relations between touch and vision in infants. A neonate can see (even track) and reach after the 5th day of life. From birth to 3 months, oral, tactile, and proprioceptive information is related to visual information. A baby with prior tactile experience with an object can recognize the object visually. The opposite does not occur until 4 or 5 months. The psychological evidences of this work show that reaching in neonates is a balistic movement, stimulated (triggered) by any visual or auditory perception, without any kind of control during the approach, while in babies from 4 to 5 months reaching can be smoothly guided, even changing the trajectory during the approach. This suggests reaching as an inborn ability, that develops (learning) according to motor development. In [33, 34, 35] a theoretical work has produced a mathematical model for the development of reaching. The model suggests that infants are constantly learning about the current capabilities of their motor systems and adapting reaching strategies to accord with their current level of motor control. This suggests that reaching may not be an inborn ability, but can be learned, in a feedback learning paradigm. Note that in both approaches the develompment of reaching might uses a learning approach.

The conclusion is that, no matter the approach, it is better for the systems (vision, proprioception including taction, and probably audition) to learn in an integrated frame-work than as independent subsystems. In this sense, they can operate as coordinated devices in the execution of a given task. For example, if the task is to reach for some visually perceived object, errors in both the stereo vergence and tactile response are feedback for the integrated system.

3 Vision and Touch Integrated System Architecture

A general functional and descriptive view of an eye-arm integrated system is presented in this section (figure 1 shows the basic architecture). Input from each sensory system, extracted from the current region of interest (attention window), is transformed to a set of descriptive features. These features are used as a pattern activation code in a central associative memory, which matches the properties to a long term memory address. The long term memory has stored (or will store for new objects) information about the environment/objects (facts and history). A pre-attentional and an attention-shift control mechanisms are necessary to change the focus of attention from one region to another. In the next subsections, each structure and subsystem will be detailed.

3.1 Spatially-Indexed Perceptual State

In this subsection we will describe how input for the integrated system is organized on spatially-indexed areas,

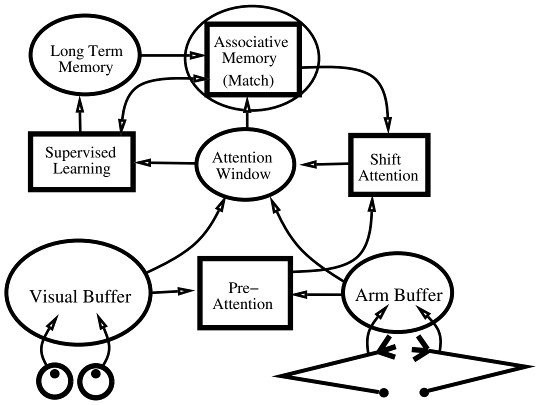

Figure 1: System Control Architecture

constituting the perceptual buffer. Information is grouped in such a way that topological/spatial indexes from multiple sensors are mapped to multiple features.

3.1.1 The Visual Buffer

Neuro-physiological studies reveals that the images produced by left and right retinas are primarily projected in areas V1, V2, V3, V3A, and V4 of the visual cortex, located in the posterior part of the brain. Researches [1, 7, 8, 18, 29] suggest that these retinotopically mapped areas constitute an structure called the "visual buffer". This structure seems to be like a multi-scale (pyramidal) map of the perceived scene, not necessarily in contiguous cortical areas. Information contained in one or more levels of resolution flows from the visual buffer to the high-level processing. The existence of this structure can be computationally justified by following Marr [1] and Nishihara's [4] approaches if we further assume that some small set of perceptual features are necessary and sufficient to build an effective and computationally efficient system. Most information will be ambiguous and unnecessary. The useful information is related to the object that is the current focus of attention. In this sense, regions of related percepts can be grouped and organized together. Information about object affordances, shape, texture, color, spatial location, orientation, and size can be extracted easily and perceptual systems can be orchestred efficiently.

3.1.2 The Arm Buffer

The arm buffer, similar to the visual buffer, contains arm and hand sensory (tactile and proprioceptive) information. Object properties like tactile texture, size, weight, softness, and spatial properties like position and orientation can be extracted from this structure. The arm buffer is located in biological systems in the somato-sensory region of the brain (a small area in comparison to the visual buffer).

3.2 Attentional Control

Two modalities of attention can occur: voluntary (or top-down attention), activated by the associative memory in identification tasks or by an attentional mechanism in other general attention tasks, and involuntary (or botton-up attention), activated by the pre-attention mechanism. Involuntary attention shift occurs when strongly activated stimuli (high color intensity, moving pattern, large visual angle, unexpected tactile stimulus, or stimulus not identified yet) require the attention window. A pre-attention mechanism is responsible for providing this kind of attention. In general attention tasks, voluntary attention is employed when the eyes must search for a stimulus that has properties most like those of a given model (also called top-down attention). Also, if we know that a stimulus will appear in a certain region, the pre-attention mechanism can set up the attention window in that region, waiting for the stimulus or it might be necessary to keep attention to an object for some reason (tracking, reading). In identification tasks, a voluntary attention shift occurs in the case of no identification of an object in the current trial and the presence of any activated pattern under the threshold in the associative memory. In this case, a zoom can enhance any particular characteristic, a guided search looks for a specific spatial hypothesis, a random search looks at regions of high interest, an, finally, an arm movement can improve the input properties.

3.2.1 Attention Window

The attention window is an structure that allows one to extract the information provided by one of the spatially organized regions of the perceptual buffer. Its position is not fixed inside the visual/arm buffer. On the other hand, the most useful visual information (e.g. depth) can be extracted if the focus of attention lies in the center of both visual fields. The size (also the shape) of the attention window can be dynamically changed, depending on the size (and shape) of the region of interest (ROI). The attention-shift mechanism selects a new processing region and effectively moves the attention window, disengaging it from the current region of interest and engaging it in the new one.

3.2.2 Pre-attention (perceptual cues)

The pre-attention mechanism operates at low level, computing the (involuntary) activation values for each region of interest (ROI). This allows the attention-shift mechanism to choose between the various activated stimuli, deciding where to direct the attention. Note that some ROI activation values are updated by the associative memory.

3.2.3 Attention-shift Mechanism

The eyes operate in saccadic movements to change the current focus of attention. Attention shifting has 2 components: one that actually shifts the body, head, eyes and/or attention window and another that at the same time primes the actual representation of any sought property, making it easier to encode. In [30], Julesz and Saarinen found that the time required for covert attention shift (moving the focus of attention without moving the eyes) in humans is about 30 to 50 milliseconds. The time required for a shift involving a saccadic eye movement is about 100 to 200 milliseconds [31].

Note that the arms can also operate on the attention shift. If an arm is reaching for an object and eventually touches some other object, the attention window could be redirected as result. In some cases, the eyes will look to the new stimulus in order to recognize or identify the object or to establish the associated state required to deal with the unexpected tactile event.

3.3 Associative Memory

Identification (addressed in this work) and other tasks require that multi-modal sensory input data must access the same representation of an object. The representation containing information like facts and history about environments/objects is accessed by an associative memory structure. The representation that has properties most like those of the current sensory input will become most activated. If this activation is over a threshold the object is identified, otherwise more information must be provided. The attention-shift mechanism will be required to do this work, changing the actual focus of attention. If after all "get more information" trials, there is no match for an input, a new or unknown object has been discovered and new information and facts must be stored (automatic supervised learning is activated).

A representation in associative memory specifies the address which corresponds to the pattern code produced by the properties. Representations of the actual size (invariant size, not retinal) and other properties are also stored. Note that the descriptive set of features converging to the same area in memory suggests that a neural network implementation would be appropriate. In fact, we have used a backpropagation neural network with a winner-take-all mechanism to choose the most activated pattern address.

A backpropagation network [16] is a fully connected feedforward multi-layered network. It maps the input vector $\vec{x}$ to an output vector $\vec{o}$ as determined by the weight vector $\vec{\omega}$. In order to train the BP network, sets of correct input and associated output are given to the training procedure. The errors between the desired (correct) output and the output computed by the network is then incrementally distributed to the connections affected by the input in a backward process. In general, the following equation is used for the training procedure.

$$\Delta\omega_{ij}(t+1) = \epsilon\delta_j o_i + \alpha\Delta\omega_{ij}(t),$$

where

$$o_i = (1 + e^{-\sum_{i=0}^{A}\omega_{ij}x_i})^{-1},$$

$$\delta_j = \begin{cases} o_j(1-o_j)(y_j - o_j), & \forall j = 1, ..., B \\ o_j(1-o_j)\sum_{k=1}^{B}\delta_k\omega_{jk}, & \forall j = 1, ..., A \end{cases}$$

3.4 Spatial maps

We have used a discrete angular coordinate system to perform the eye and arm maps. The ROIs are represented in configuration coordinates, which facilitates the movement (displacement) computation. A discrete polar map is defined for each eye. A ROI is represented as an integer interval $[a, b]$ in this polar map. In the same way, a two grid represents a discretized configuration space of the arm. We use two such maps for each arm, a boundary map representing all objects (or obstacles) and a potential map (used for path-planning). Angular coordinates are a natural choice for encoding space in an active perceptual system because they can be directly mapped to torque and velocity (motor-based coordinates) used by the positional derivative (PD) controllers and they also support path planning control.

4 The Simulation Environment

In order to computationally investigate the system presented in the previous section, we have implemented it on a simulation platform. The simulated robot "Roger-the-Crab" has 5 controllers (neck, eyes and arms) integrated in a single platform. Figure 2 shows Roger's environment.

The world can be constructed or modified by means of inserting, moving, or removing different types of objects. Currently we have circles, elipses, equilateral triangles, and squares. The object gray level, weight, and size can also be specified. We have implemented a Phong ilumination model and a Gaussian noise process to calculate light intensity for each image pixel using up to 16 punctual light sources and/or a sun light. A Laplacian of Gaussian filter, using 3 different kernel diameters, constitutes the low-level signal processing (simulating the known MAC-band effect that occurs in biological low-level vision). Also, compensation for gravity and other ambient effects, and the basic robot kinematic and dynamic equations (and their inverses) are calculated for use in the arm and eye servo controllers. This is performed at 50 Hz on the simulated clock, mapping the external world

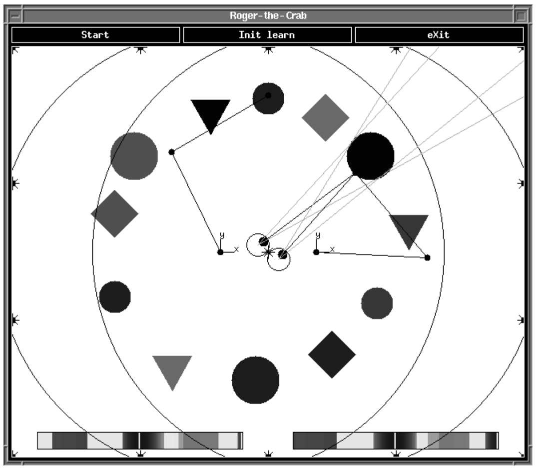

Figure 2: Roger-the-Crab

geometry into visual, proprioceptive and tactile information. We simulate the weight of an object by attributing an arbitrary value to each object mass and mapping it to torque and velocity (the proprioceptive information) necessary to lift the object. Independent controllers run concurrently for each eye and each arm. The coordination of all controllers is basically the definition of which concurrently controller will be run at a given time. Roger also has some operating mechanisms that are responsible for updating the current information and state space (consequently the world map). After a controller convergence, a set of high-level procedures will provide the changes in the state-space, based on the identification or not of an object.

4.1 Pre-attention and Attention-shift Operation

The pre-attention mechanism operates after convergence of the eyes/arms controllers. Its function is to extract edges delimiting the ROIs and to calculate the (involuntary) activation values (in get the attention window) for each ROI. These activation values are normalized functions $[0.0, 1.0]$ of retinal size, intensity average, visual motion, tactile response information, and mapping status. The mapping status (initially 1 for all ROIs) is a flag which is set to zero if a given ROI receives the focus of attention. These values are updated only for the ROIs that are currently in the visual/arm buffer.

The attention-shift mechanism also operates on converged states and is followed by the associative memory match. Its function is to calculate the most active ROI, using the activation values calculated by the pre-attention and updated by the associative memory match, to shift the current attention window to the winner ROI, and to make the neck and eyes controllers run until convergence (until get the attention window in the center of the visual field). The decision on where to put attention is made by a Winner-take-all mechanism. Note that the focus of attention can be determined by one of the eyes or arm. This eye or arm is referred to as dominant.

4.2 Converging the Eyes and Arms onto a ROI

The eye PD controllers, composed by the neck (pan) and by each eye (vergence), operate in a different manner than the arms. The dominant eye controller takes the current attention window center (determined by attention-shift mechanism), and calculates the angular displacement necessary to get there. Concurrently, the non dominant eye controller calculates the angular displacement necessary to maximize the correlation between the image centers and the neck controller (operating with a low gain) try to keep its gaze in the horopter. After each eye and neck controllers, the eye and neck servos, a PD position controller, run, updating the current angular velocity and torque. By maximizing the correlation between the eyes center in each step of motion, the eyes converge while moving to a new place, and as the neck PD controller gain is significantly lower than the eyes one, there is no backlash, resulting in smooth movement.

The arm movements are visually guided. The positional goal is the horopter. If one of the arm controllers should run, the path planner calculates a collision-free path from the current position to the goal based on current information contained in the arm (potential and boundary) maps. This path is calculated by solving a harmonic function (gradient descent strategy) applying a relaxation rule in the potential map based on the obstacle conditions specified in the boundary map. Note that a new object could not be represented in the boundary map yet. Then, if the arm bumps something, the controller can be turned off (converges), an arm bumper flag (activation value) can be set high, and the focus of attention can be shifted to this region allowing the eyes to identify the new object.

4.3 Extracting Object Properties (Features)

Once the eyes (or one of the arms) have converged in the center of the current attention window the 3D shape and other normalized properties can be extracted. This feature extraction is adaptive in its spatial resolution, depending on the visual size of the region of interest. In other words, this means that the same amount of information (stereo measurements, texture, intensity) is provided, independent of the visual size of the stimulus. Intensity is calculated as an average of pixel intensities. Shape is a variance vector (2nd order moments) of the stereo measurements on three levels of resolution (normalized disparity variance). This is not the best representation of shape, but can differentiate a small set of objects with different shapes. The size (extracted from the stereo measurements) is normalized between zero and an arbitrary maximum length, the same occurring with the weight (from the arms). Texture is also calculated as a normalized 3D vector of 2nd order moments of the 3 levels of resolution Laplacian of Gaussian responses.

Note that the horopter must be in the center of the attention window in order for the 3D shape reconstruction. In this fashion, the only policy the eye controllers have to learn is how to keep both eyes converged on the same object (tracking). This is accomplished simply by running the eye controllers (see previous subsection). Q-learning [15, 16, 17] can also be applied here to derive a vergence policy [36], in which the actions performed by the eye controllers leading to the convergence state are rewarded. Also, note that the spatial resolution can be determined as a function of the performance (precision and time of response) required to better execute a given task. A quick look might allow only a coarse level of disparity, intensity and texture computation, while a longer look might allow the computation at fine levels of resolution.

4.3.1 Matching the Object Properties

Once we have the shape, size, and location (all calculated from stereo), perceived luminance (intensity), visual texture, and the weight (from the arms), these are used as the activation pattern by the associative memory match, allowing the recognition and consequently identification of the current object. An incremental map of the environment is constructed (the pre-attentional maps), one object (or ROI) at a time. Once an object is identified it is incorporated to the map by setting its ROI "mapping status" to zero. This allows a shift of attention toward other ROI. If the eye-improvement action fails to lead to identification, the arm-improvement action can run again to improve the input. The features will be extracted again and a new match will occur. If the object remains unidentified, the associative memory can be updated with the new object properties, by means of invoking the supervised learning procedure. At the same time, the "map status" is set to zero. Note that in case of negative identification (in the presence of pattern activation), the mapping status activation value still remains high, requiring that the attention window keeps on the same ROI.

5 Experimental Results

We performed some experiments using different object types (circles, squares and triangles), with different intensities, sizes, and weights, and placed in various locations in Roger's environment. Table 1 lists some (arbitrarily given) properties and the corresponding sensory normalized values computed by the stereo reconstruction and other processes for one of the tested environments. These

Object	Intensity	Size	Weight
01	99	30	15
	0.83	0.45	0.72
02	79	30	10
	0.65	0.43	0.51
03	69	30	10
	0.56	0.44	0.47
04	59	30	5
	0.47	0.44	0.24

Table 1: World and perceived normalized feature values.

normalized values are directly used as input to the associative memory. The object types can be seen in Roger's environment in figure 2. The right arm and the neck and eye controllers have converged. In that moment, information about that sought object is being extracted and matched to a representation in the associative memory.

The BP network acts as expected, positively identifying objects if the activation is over a threshold. We have used a threshold computed as a weighted function of the minimum and maximum errors given by the training procedure. If an object is not identified only by visual improvement, the arm trial is accomplished by moving one of the arms and measuring the object weight. In case of no identification (activation is still under the threshold), the supervised learning procedure dynamically updates the network.

After all regions of interest are visited, the eyes remain in the vigilant state. We have defined another activation value (interest) that is set to zero when a ROI is firstly settled in the pre-attentional map. This value keeps increasing with the servo-clock until reset by an attentional visit. This forces the attention-shift mechanism to eventually choose that ROI as the focus of attention, allowing the detection of changes in the environment and putting the robot in a behaviorial active state. Without this mechanism, the eyes remain at rest until a change occurs in the world inside the visual-field. If the world changes, the system imediatelly puts its focus of attention in the region changed.

Another important expected result is that all objects were visited (looked at by Roger), identified (or not) and mapped. In addition, the regions of background (we have simulated a room with 4 walls, with low intensity values which are also projected in roger retinas) were also identified. This was expected, since in the current implementation Roger does not make any distinction between an object and a wall part. The objects were identified (due to the higher level of intensity) previously to the wall, winning in the attention-shift mechanism.

We tested Roger with a set of 36 different objects in the same environment, using 4 features to perform the memory match, with the objects having close feature values. All object representations could be inserted in the associative memory and then some other object instances randomically settled in the world could have the attention of Roger and could be correctly identified.

6 Conclusion and Future Work

By using the simulation platform we have demonstrated that the multi-modal sensory system described here is a useful tool in recognition and identification of objects. These are considered low-level tasks, and can be used by high level tasks like surveillance, path-planning, orientation, obstacle avoidance and motion. Avoiding the complete low-level image depth generation and decreasing the resolution of stereo and texture measures seems to be a good first step in the development of a feasible real-time stereo system. This architecture can be implemented in pipeline array IP processors. In these, the low-level processes can be sped up by using hardware implemented filters and good estimations can be provided for the disparity computations by using a multi-resolution buffer. In the simulation platform, the eyes' vergence was accomplished by only maximizing correlation measures. However, in the hardware platform the vergence mechanism can use focus [26] to approximately determine the vergence position and then finely verges by maximizing correlation. In addition, motion can be implemented by using motion estimation filters like the ones described in [27, 28].

The system successfully implemented on the simulation platform is currently being implemented on the hardware platform, basically composed of a Stereo-head, Datacube Memories/Image Processing devices, two robotic arms (with hands), and a host interface computer. The next step is to test this architecture in the real environment and to measure its performance in a variety of tasks. Another improvement that can be addressed is to assume an object as a region inside a set of well structured edges, instead of the simple definition of shape used in this work. This can improve the categorization of the objects, helping the associative memory match. The associative memory, in its turn can be improved by applying local training if a object is inserted. We currently retrain the whole network synaptic weights.

The immediate application that comes to mind is surveillance. Here, a map of the environment is incrementally and dynamically constructed. Questions like "who or what are in the lab?" or "is/was there an object at that position?" can be addressed by simple analysis of this dynamic map. As more advanced general tasks, robots can use the basic procedures developed in this work in order to learn how to navigate between different rooms in a building. The robots may also be able to learn facts, history, and other useful information about people in the building.

References

[1] D. MARR. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. MIT Press, Cambridge, MA, 1982.

[2] B. K. P. HORN. Robot Vision. MIT Press, Cambridge, MA, 1986.

[3] H. K. NISHIHARA. Practical Real-Time Imaging Stereo Matcher. Technical Report. MIT, AI Laboratory. 1984

[4] H. K. NISHIHARA. Minimal Meaningful Measurements Tools. Technical Report. Teleos Research, CA. 1991.

[5] E. HUBER and D. KORTENKAMP. Using Stereo Vision to Pursue Moving Agents with a Mobile Robot. Proc. of IEEE Conference on Robotics and Automation, 1995.

[6] A. PAPOULIS. Probability, Random Variables, and Stochastic Processes. MacGRAW-HILL, 1991.

[7] D. MARR and T. POGGIO. A Computational Theory of Human Stereo Vision. Proc. of the Royal Society of London, 204, pp 301-328, 1979.

[8] W. E. L. GRIMSON. From Images to Surfaces: A Computational Study of the Human Early Visual System. MIT Press, Cambridge, MA. 1981.

[9] D. H. BALLARD and C. M. BROWN. Computer Vision. Prentice-Hall. Englewood Cliffs, NJ, 1982.

[10] D. H. BALLARD. Animate Vision. Artificial Intelligence, 48:57-86,1991.

[11] J. JAMESON and L. LEIFER. Automatic Grasping, An Optimization Approach. IEEE Transactions on Systems, Man, and Cybernetics (SMC), 17:5. 1987.

[12] V. NGUYEN. Constructing Stable Grasps. International Journal of Robotics Res. 8,1:26-37. 1989.

[13] R. TOMOVIC, G. BEKEY, and W. KARPLUS. A Strategy for Grasp Synthesis with Multi-fingered Robot Hands. In Proc. of IEEE Conf. Robotics and Automation:83-89. Raleigh, NC, 1989.

[14] J. COELHO and R. GRUPEN. A Control Basis for Learning Multifingered Grasps, Journal of Robotic Systems 14(7):545-557, 1997.

[15] E. ARAUJO and R. GRUPEN. Learning Control Composition in a Complex Environment. Proceedings of Int. Conf. on Simulation of Adaptive Behavior (SAB'96). Cape Cod, MA, September, 1996.

[16] D. H. BALLARD. An Introduction to Natural Computation. The MIT Press, Cambridge, MA, 1997.

[17] C. J. C. H. WATKINS. Learning from Delayed Rewards. PhD thesis, King's College, Cambridge, UK, 1989.

[18] S. M. KOSSLYN. Image and Brain. The Resolution of the Imagery Debate. The MIT Press, Cambridge, MA, 1994.

[19] A. STRERI (Translated by T. POWELL and S. KINGERLEE). Seeing, Reaching, and touching. The MIT Press, Cambridge, MA, 1993.

[20] N. QIAN. Computing Stereo Disparity and Motion with Known Binocular Cell Properties. Neural Computation 6:390-404, 1994.

[21] N. QIAN and Y. ZHU. Physiological Computation of Binocular Disparity. Vision Research, 37:1811-1827, 1997.

[22] D. J. FLEET, H. WAGNER, and D. J. HEEGER. Neural Encoding of Binocular Disparity: Energy Models, Position Shifts and Phase Shifts. Technical Report, 1997.

[23] R. D. FREEMAN and I. OHZAWA. On Neurophysiological Organization of Binocular Vision. Vision research, 30:1661-1676. 1990.

[24] T. D. SANGER. Stereo Disparity Computation using Gabor Filters. Biological Cybernetics, 59:405-418. 1988.

[25] W. T. FREEMAN and E. H. ADELSON. The Design and use of Steerable Filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(9):891-906. 1991.

[26] A. HORII. The Focusing Mechanism in the KTH Head Eye System. Technical Report, Royal Institute of Technology, Stockholm, Sweden.

[27] S. SOATTO, R. FREZZA, and P. PERONA. Motion Estimation via Dynamic Vision. Technical Report, California Institute of Technology, Pasadena, CA, 1997.

[28] S. LEE and Y. KAY. A Kalman Filter for Accurate 3D Motion Estimation From a Sequence of Stereo Images. CVGIP Image Understanding, 54(2):244-258. 1991.

[29] B. JULESZ. Foundations of Cyclopean Perception. University of Chicago Press, Chicago, 1971.

[30] B. JULESZ and J. SAARINEN. The Speed of Attentional Shifts in the Visual Field. Proceedings of the National Academy of Sciences of USA, 88:1812-1814. 1991.

[31] B. FISCHER, H. WEBER, M. BISCALDI, F. AIPLE, P. OTTO, and V. STUHR. Separate Populations of Visually Guided Saccades in Humans: Reaction Times and Amplitudes. Exp-Brain-Res., 92:528-541, 1993.

[32] C. I. CONNOLLY and R. A. GRUPEN. On the Applications of Harmonic Functions to Robotics. Journal of Robotics Systems,10(7):931-946.

[33] R.K. CLIFTON, P. ROCHAT, D. J. ROBIN, and N. E. BERTHIER. Multimodal Perception in the Control of Infant Reaching. Journal of Experimental Psychology: Human Perception and Performance 20:876-886. 1994.

[34] N.E. BERTHIER. Learning to reach: A mathematical model. Developmental Psychology, (in press). 1996.

[35] N.E. BERTHIER. Infant reaching strategies: theoretical considerations. Infant Behavior and Development 17: 521. 1996.

[36] J. PIATER, K. RAMAMRITHAM, R. A. GRUPEN: Learning Real-Time Strategies for Binocular Vergence In preparation. University of Massachusetts, Amherst, MA.