MULTI-MODAL ATTENTION AND EVENT BINDING IN HUMANOID ROBOTS USING A SENSORY EGO-SPHERE

By

Kimberly A. Hambuchen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

May, 2004

Nashville, Tennessee

Approved:

Richard Alan Peters II

David. C. Noelle

Robert O. Ambrose

Kazuhiko Kawamura

D. Mitchell Wilkes

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor, Dr. Richard Alan Peters II. Without his encouragement and persistence, I would not have attained this level of education and expertise. Were it not for his complete faith in my intellectual abilities, this dissertation would never have been completed. For all of his work and assistance, I thank him greatly. To my parents, Ray and Donna, thank you for allowing me to be a professional student all of these years. I know I would not have lasted so long without the total support, both parentally and financially, they have given me. I would also like to thank my committee members. Dr. David Noelle has been outstanding in his guidance of my writing and I will eternally be in debt to him for his efforts. I thank him for going much further with my work than was ever expected. Dr. Robert Ambrose has been a great outside committee member and has made me feel completely welcomed in the NASA community. To Dr. Kazuhiko Kawamura, thank you for giving me chances to do so many things in the lab. Your confidence in my abilities to do almost anything has been much appreciated. And finally, thanks to Dr. Mitchell Wilkes. He has proven to be a calming committee member and great source of humor through it all. I thank him for all he has done for me over the past years.

Next I would like to thank past and present members of the Intelligent Robotics Laboratory and of the Cognitive Robotics Laboratory. They have all helped me tremendously over the years. To Tamara Rogers, I thank you for every last bit. You know what I am referring to – everything. Without Tamara, this dissertation may not have been completed. To Anthony Alford, Steve Northrup and Mark Cambron, thank you for all of your help and patience over the years. To Bugra Koku, Jian Peng and Carlotta Johnson, you guys have been so much fun to work with and I hope to find people like you in all areas of my life.

Finally, I would like to thank the Graduate Student Researchers Program at NASA for giving me a three year fellowship to conduct my research (Grant # NGT0-52). This work was also supported over the years by the following grants from DARPA and NASA: DARPA SMDC Grant #DASG60-01-1-0001, DARPA-MARS2020 NASA Subcontract Grant #NAG9-1446 and NASA-JSC Grant #NAG9-1515.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Robots today have advanced and complex sensors that can detect activity in the environment and in their own bodies. These robots also have an array of complex processing routines that can transform the detected sensory activity into useful information. From localization of a sound source to recognition of faces, these processing routines can create a vast assortment of sensory information for a robot to use. In most robots, however, the outputs of these sensory processors are more useful to robot developers than to the robots themselves. Such robots lack the basic cognitive skills that interpret the meanings of these stimuli. The processes of finding appropriate sensory stimuli, understanding what these stimuli signify and enabling a robot to use the stimuli appropriately so as to learn from its surroundings are still in their beginning phases.

This dissertation presents a solution to some of the problems associated with the processing and analysis of sensory information by robots. In particular, it describes a software mechanism for sensory **event binding**, a process whereby the responses of different sensors to a single event (external or internal) are recognized as such. It also describes software for **attentional processing**, a process that selects among all available sensory information that which is most important at the current time. These procedures are defined with respect to a **Sensory Ego-Sphere** (SES), a software structure for a robot that serves, among other things, as a short-term memory for a robot [Peters et al., 2001].

Robotic sensing requires sensors, physical devices that transduce energy into numerical signals, which can be grouped into two categories. **Exteroceptive** sensors respond to stimuli in the robot's external environment while **proprioceptive** sensors measure some aspect of the current internal state of the robot. The former category includes cameras, microphones and laser range sensors while the latter includes force-torque sensors and strain gauges. Sensors and their responses alone provide no fundamental value for a robot. The sensory signals must be processed by extracting specific information relevant to the robot's existence so as to structure the individual signals appropriately. Thus, the output signals of individual sensors serve as input to various **sensory processing modules** (SPM) that detect specific patterns in the signals and/or filter out irrelevant information.

'Event' is used frequently in this work to describe several different phenomena. When used with respect to the robot's environment, it refers to a single object or to the beginning, ending, or momentary pause of a temporally extended incident that generates or alters the energy transduced by one or more of the robot's sensors. 'Event' is used similarly with respect to the robot's internal workings. The robot's SPMs are typically designed to detect signal patterns or dynamics in the more narrowly focused streams of information that they extract from raw sensor output. **Sensory event** denotes the structured sensory output of an SPM. A **source** is the physical event that generates the stimuli leading to these sensory events. For example, a person can be a source that generates stimuli detected as sensory events by motion processors, face detectors and sound localizers.

Problem Area

A central problem addressed by this dissertation is that of relating the numerical time-series outputs of the various SPMs to each other and to their proper sources. That is, which events detected by the SPMs belong together by virtue of having been produced by the same source? Which should be bound together because they occurred in response to an action of the robot? Which should be ignored as irrelevant or spurious? How can the robot ignore these irrelevant or spurious stimuli without missing those that indicate danger or opportunity? The first two questions are answered through event binding which detects **spatio-temporal coincidence** of stimuli. Spatio-temporal coincident stimuli are those that contact the robot's sensors at about the same time and/or originate those from approximately the same region of space. The second two questions are answered through attentional processing which operates as a function of sensory events and the robot's tasks. Both attention and event binding rely on **short-term memory** (STM). Different SPMs have different **latencies**, which are the times required for a SPM to produce an output from its input. Thus, the temporal binding of different sensory events requires temporary storage of events which are detected more quickly than others. Moreover, the detection of a source may involve an accumulation of different sensory events over time, leading to an attentional focus. A Sensory Ego-Sphere that provides STM indexed by space and time is therefore used as the supporting structure for the goals of this work. The proposed system is unique in that it provides a single structure for event-binding and directing attention over egocentric space in a manner appropriate for the sensory systems of fixed-base humanoid robots. Figure 1 shows the architecture of the system which seeks to answer the questions presented above.

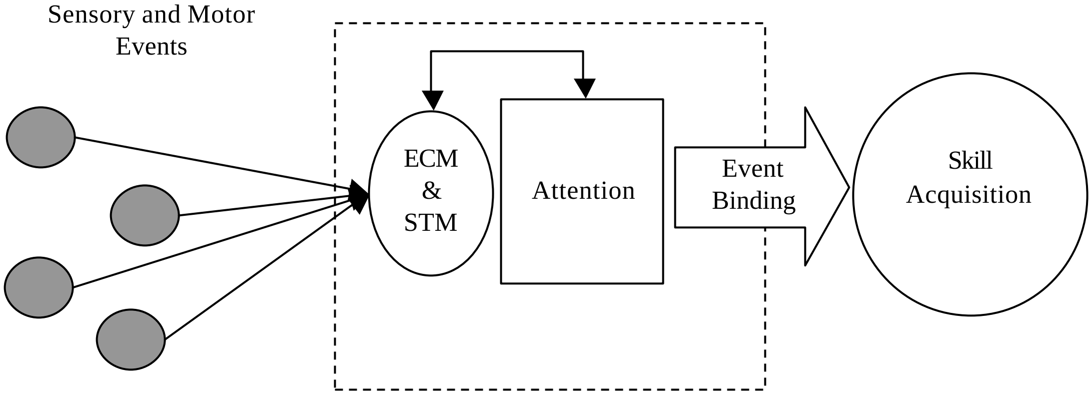


Figure 1: System Architecture

In this system, **egocentric mapping** (ECM) generates a spatial map of the robot's environment. ECM is used to represent sensory events from the robot's perspective so that events that occur in the same direction, as observed by the robot, are mapped to the same SES location. STM stores these events with their temporal properties (e.g. time of occurrence, processing latency). Attention selects the spatial location that is most important to the robot at the current time while event binding selects spatio-temporal coincidental events and sends them to skill acquisition areas of the robot.

Objectives

The objectives of this work can be stated as the development and testing of computational structures for spatio-temporal coincidence detection in sensory information and for salience-driven focus of attention.

A **salient** location is that in which events that are of high importance to the robot at the current time are detected. Events related to a current task are salient because they may help a robot to complete this task. An accumulation of events is salient because multiple activities in one area may indicate an immediate situation which the robot needs to resolve. For example, if the robot's current task is to locate a drill for grasping, then any event that relates to a drill (e.g. object recognition of the drill) is salient. At the same time, if a group of people are waving and talking to the robot, the detection of many faces, sounds and movements in one area is salient. The attention network presented in this dissertation determines which of these and other areas is the most salient. The event binding mechanism decides which events co-occurred to create this salience, that is which events are spatially and temporally coincident.

The robots used to test the work presented in this dissertation are ISAC, Vanderbilt University's humanoid, and Robonaut, the DARPA/NASA humanoid. For both of these robots, sensors exist to transform energy from activity in the environment (exteroceptive) or in the robot's body (proprioceptive) into signals. The sensors send their signals to SPMs. If a SPM detects an event, the event is sent to the SES where it is mapped egocentrically and temporally. The attention system determines the most salient location in the robot's environment and the event binding mechanism selects which sensory events from this area are spatially and temporally coincident.

ISAC and Robonaut should be able to direct their attention to areas in their environments which are the most salient. The assumption made in this work is that areas of high salience may require immediate action or further inspection. The location of the most salient area is important because the robots have limited physical resources that cannot be distributed among different stimuli. For example, the cameras cannot center on two objects at once, therefore the robot needs to know which object is the most salient to perform further inspection or take action upon the object. A hand cannot grasp more than one object at a time. The robot should be able to determine which object is most salient so that the hand may grasp the important object.

A location may be salient due to activity detected simultaneously by several different sensory modalities. Therefore, the attention network that operates on the environmental information should be multi-modal. Moreover, the robot should associate sensory events that co-occur and originate from the same source. For example, if the robot wants to place a screwdriver in a tool box, the attention network should guide the robot initially to the screwdriver and to the tool box. However, if the robot cannot reach the screwdriver because its arm is blocked by a table, the robot will never attain its goal. In this case, the most salient area of the environment is that where the table is blocking the robot's arm. The robot will have detected that the arm is not moving, a heavy force is acting upon the arm and a table exists in this location. If these events are bound together, they can be associated. With several such instances, the robot could learn that these co-occurrences indicate the presence of an obstacle.

Before this work, neither ISAC nor Robonaut had the ability to determine the most important sensory events or to group a collection of events that emanated from a single source. This ability was needed to support other research in skill acquisition. Therefore, the ultimate goal of this work is to create a link from the detection of events to the use of that information by later processing stages in the robots.

Significance of Work

Rather than allocating a separate attention network to each sensing modality, the sensor outputs from different modalities are combined together in one data structure, the Sensory Ego-Sphere (SES). The significance of this approach lies in using the SES as the central location for mapping and storing events, assigning salience to the events and selecting coincident events. The SES, described in Chapter 3, functions as a short-term memory that maps sensory events in an egocentric manner. It can store events from sensors with different resolutions. The attention network assigns salience to events stored in the SES. Salience is based on the **incidence** and **task-relevance** of an event as well as whether or not the event is **habitual**. The SES was originally designed to be a short-term memory. The salience of an event registered onto the SES decays over time. Due to resolution differences between sensors and sensor error, the salience assigned to a sensory event is spread around the point where the event was registered onto the SES. The spreading causes salience to build up in areas on the SES that contain many events. This means that an area on the SES might not house an event, but may still have high salience due to the spreading of salience from nearby events. After the attention network selects the area with the most salience, the event binder determines which events co-occurred spatially and temporally, the former by associating nearby events within the salient region and the latter by selecting those from the region that occurred within a short time interval. Other computational modules within the robot can obtain from the SES the data associated with the sensory events so bound.

The attention and event-binding mechanisms of this system both perform at levels that are essentially equivalent or better than those exhibited by common and/or alternative approaches. The attention network successfully locates the areas of high salience as determined by incidence, task-relevance and habituation of sensory and motor events when sensors report events within their resolutions. The event binder also successfully selects events that did occur together in the case of single event sources and multiple event sources. It is shown later in this dissertation that when compared to others' methods for finding salient events (e.g. focus of attention [Lang et al., 2003; Déniz et al., 2003]), the system described here performs as well as or better. The binding of co-occurring events performs well when compared to probabilistic measures used to determine co-occurrence.

Paper Organization

Chapter 2 presents a brief description of the Sensory Ego-Sphere and previous robotic methods developed in ECM, STM, sensor integration and attention. Chapter 3 presents the robotic platforms on which the methods in this dissertation are tested. This chapter also presents a functional description of the SES. Chapter 4 presents the methods used to develop the attention and event binding software structures and the motivations behind these methods. Chapter 5 presents the experiments designed to test the software structures and their results. This chapter also contains evaluations of the structures' performances. Conclusions and future work are detailed in Chapter 6. The Appendix provides a detailed explanation of the Sensory Ego-Sphere structure.

CHAPTER II

BACKGROUND MATERIAL AND PREVIOUS WORK

This chapter briefly revisits the problems undertaken in this dissertation, provides a succinct description of the Sensory Ego-Sphere and evaluates previous work in related areas. The related areas described are egocentric mapping, short-term memory, sensor integration and attention. Evaluations of methods used in each area on robotics systems are made. The results of these evaluations are contrasted against the methods presented in this dissertation (Chapter 4).

Problem Statement

The problems that this dissertation seeks to solve are that of spatio-temporal coincidence detection of sensory events and of attending to salient sensory events, both by a humanoid robot. The solution presented in this dissertation uses egocentric mapping and short-term memory to facilitate the event binding and attention system. The Sensory Ego-Sphere is the unified mechanism upon which the preceding systems are applied.

Sensory Ego-Sphere: A Brief Overview

The Sensory Ego-Sphere (SES) is a software object that serves as a mediator between the sensing and cognition of a robot [Peters et al., 2003]. The SES can function as a short-term memory and can facilitate attention as well as detection of co-occurring sensory events. It operates asynchronously as a data structure in a parallel, distributed control system that includes independent, parallel SPMs. The SES was inspired by Albus's egosphere [Albus, 1991].

Ideally, the SES can be visualized as a spherical shell centered on the coordinate origin of the robot. This situation provides an egocentric representation for the robot, thereby facilitating egocentric mapping. Each point on the shell is a locally connected memory unit with a temporal decay to provide short-term memory. A SPM sends an event to the SES to be attached to a point on the shell. The detected location of the event is projected onto the shell to find the attachment point. The SES attaches the event at the point closest to this projection, along with the time of registration and any other information the SPM may collect about the event. Specifically, the distance to the event is only stored if the SPM computes the 3-dimensional position of the event. However, the actual elevation and azimuth angles at which the event is detected are stored so that the robot may return to the event's exact location. In the case of visual events, the verge angles are also stored so that the depth may be calculated at a later date if needed. System components of the robot that use sensory data may read from the SES. Therefore, information can flow to and from the SES.

Structure of the SES

In its actual implementation on a robot, the SES is a database with associated computational routines. The records in the database are connected as nodes in a graph isomorphic to a regular tessellation of a sphere centered on the coordinate frame of the robot. In particular, the topological structure of the SES is that of a geodesic dome, defined as the "triangulation of a Platonic solid or other polyhedron to produce a close approximation to a sphere or hemisphere" [Weisstein, 1999]. Each vertex on the dome contains a pointer to a distinct data record. Thus, the SES is a sparse map of the world that contains pointers to events that have been detected recently by the robot's SPMs. As the robot operates within its environment, external and internal events stimulate the robot's sensors. Upon detection of an event, the associated SPM writes its output to the SES. The event is stored at the node that is closest to the direction in which the event occurred.

Since the robot's SPMs are independent and concurrent, multiple sensors stimulated by the same source will register their events onto the SES within a time interval determined by SPM latencies. If the source is directional, the different modules will register their events at the same location on the SES. Hence, given parallel, independent processing modules, events from different sensory modalities coming from similar directions at similar times will register close to each other on the SES.

The structure of the SES exists in both a theoretical geometric form and in a practical implementation form. The idealized geometric structure is presented in the Appendix. The practical structure of the SES exists as a geodesic dome interface, a database and communication managers, all of which are described in the Sensory Ego-Sphere section of Chapter 3.

Previous Work

This section presents previous methods developed for egocentric mapping, short-term memory, sensor integration and attention in robots. The goal in reviewing these methods is to determine what other methods exist and if any of these methods address the problems presented in this dissertation. Do any other methods of ECM use a unified structure? Do any other methods combine the use of ECM with a STM? Do any other methods use ECM and STM to facilitate sensor integration and attention?

It will be shown that, although mechanisms and methods exist for ECM and STM, none combines all of the functionality needed for further event binding and attention. Along with a method to egocentrically map sensory events, the mechanism needs to have a topological manner of linking spatially co-occurring events. The mechanism should also store sensory events individually in a short-term fashion so that the attention network can operate on these events. Finally, the mechanism should be able to handle sensory information from multiple modalities and at multiple resolutions.

Egocentric Mapping

An **egocentric** reference frame represents the environment in the perspective of the observer looking out into the world [Klatzky, 1997]. Egocentric frames are thought by some scientists as being input to **allocentric** reference frames, which represent locations external to an observer and are independent of the observer's position [Klatzky, 1997]. Typically, egocentric representations use polar coordinates to index locations with respect to an origin at the center of the observer's body [1][Klatzky, 1997].

The main objective for using egocentric mapping in all of the methods reviewed below is to represent sensory information that the robot detected in a manner that was innate to the robot's sensory systems. Most of the systems use only visual sensory information in their representations or have multiple sensory systems that report in the same coordinate frame. Active vision systems exist in spherical coordinates for all of these robots. The same is true for both ISAC and Robonaut – the natural coordinate frame in which both of these robots operate is spherical. Therefore, egocentric mapping of detected events is preferred because it can be easily used in a spherical coordinate system.

[1] A human's body center is taken to be a collection of the head and torso [Arbib et al., 1998]. A humanoid's body center can be defined anywhere, with the most efficient centers being centered at the head, torso, or base (depending on what type of movement the humanoid is capable of).

Most egocentric mapping methods used in the robotics field are for navigation and localization of mobile robots or as a way of maintaining local representations of a robot's environment. The latter method is how ECM is used in this dissertation. A representation of the robot's local environment is needed so that the robot can attend to sensory events within its workspace and so that the robot can detect which local sensory events are spatially coincident. Since the robots used in this dissertation are stationary (i.e. cannot move their bodies from one point in 3D space to another), only local representations of the environment are needed.

The Ulm Sparrows Robo-Cup team used a two-layer spatial representation for adaptive modeling of a soccer-playing robot's environment [Sablatnög et al., 1999]. The lower layer is an egocentric representation maps the robot's local environment, considering data found in both the robot's current field of view and data that no longer appears in the robot's field of view. The representation consists of multiple maps of features of the environment (e.g. distance of soccer ball, field landmarks, position of other robots). This method allows a robot to know where sensory stimuli are that cannot be seen at a given moment so that later processing can be performed on these stimuli. The objective of this approach is to provide a robot with immediate and adaptive mapping abilities for quick selection of low-level behaviors. Initially, the approach in this system is like that used in the SES in that all known objects (not just those in the field of view) are mapped egocentrically. However, no topological links are supplied between features in the maps. Topology in the ECM is required for event binding and helps to facilitate attention. Without the topology, coincident features cannot be integrated. Also, salience cannot be applied to areas without topology.

Kraetzschmar et al. adapt the reference system described in [Sablatnög et al., 1999] to create a hybrid approach for spatial representation in their applications to service robots

[Kraetzschmar et al. 2000]. They use egocentric, allocentric, region and topological maps to represent their robot's work environment. The authors use occupancy grids to denote locations in which objects are located. The topological map denotes each occupied area and links adjacent areas for path-planning and navigation. This representation uses the egocentric view to make a local map while the topological linking of objects discovered in the egocentric view is done in a 2D planar space while the overall objective of the system is that of path-planning and navigation. Unlike that of Sablatnög et al., this system includes the topology of features detected in the environment and can link features together in two dimensions. However, the occupancy grid approach in this system must be sampled to generate a map of the robot's local space. This approach requires computations to create a map of the area as opposed to the immediate availability of a local area map intrinsic to the SES.

One method that has a structure similar to that of the SES is the bubble model developed by Soyer et al. [Soyer et al., 2000]. The authors describe their bubble model as an egocentric spatio-temporal visual memory. The objective of this memory is to integrate different visual features of the robot's environment in a spatio-temporal manner, although no further processing is done with this information. Mathematically, the bubble is a 3-dimensional structure with a deformable surface. Control points on the bubble coincide with potential fixation points of the robot, both of which are simply pairs of pan and tilt angles from the robot's visual system. When a visual feature is detected, the strength of the feature relative to its processing routine is used to deform the surface of the bubble at the specific fixation point. For example, edge detection is used to find bars on a window. The strength of a detected edge is used to deform the bubble surface at the robot's different fixation points, or in the direction of the detected edge. The 'bubble function' represents the deformed surface. This function uses the image produced by the robot's camera and the strength measure to produce a surface output. By applying the bubble function at a given fixation point, the surface of the bubble is inflated relative to the strength of the detected visual features. This approach, while computationally and temporally very expensive, can integrate different visual features. Deformations formed by a specific set of visual features can define the co-occurrence of those features. The bubble model also seems useful in that multiple resolutions of sensory information can be used to deform the bubble. However, the information about visual features is stored in an array in mathematical form. No method exists to select information about one specific feature that may exist in a deformation of several features. Also, no topological connections exist to spread attentive measures across the bubble, although a measure of salience could be included in the deformation.

Brill et al. use markers to represent a robot's local space in a dynamic 3D environment [Brill et al., 1995]. The markers store task-relevant objects found in a robot's surroundings in egocentric space. The objective of the marker model is to maintain a representation of the robot's 3D environment. The markers are data structures that store what an object is and where it is located in the robot's environment. A main advantage of using the marker system is that once an object has been detected, the robot always knows where the object is even if it is occluded or not in the robot's field of view. However, when the robot moves, all markers must be updated. The new locations that the markers store must be estimated by transforming the previous locations in the new coordinate frame. This method is computationally wasteful in that every marker is updated upon every new position of the robot, rather than updating only as needed. Also, no topology is built into this model. Therefore, markers cannot link objects together without many computations due to the 3D coordinates in which the markers reside.

Fitzgerald developed the EgoMap to maintain an egocentric short-term memory for visual attention and tracking on MIT's humanoid, Cog [Fitzgerald, 2003]. Similar to the SES, the EgoMap is a spherical shell centered at the robot's head. It stores, however, only directional information about detected objects. On the shell, there exists a two-dimensional grid that is composed of bins. These bins are indexed spatially in dimensions that are similar to longitude and latitude, but the exact mathematics of the system is not described. The purpose of the EgoMap is to allow Cog to redirect its gaze to objects that are not in its current field of view. While the exact method of egocentric registration is not described in the paper, the representation does provide at least directional STM like the SES. The EgoMap does not have any topology and is not used for purposes other than storing short-term sensory information in an egocentric manner. No cognitive abilities of Cog are enhanced by the EgoMap.

The egocentric representation most similar to the SES is Albus's egosphere [Albus, 1991; Albus and Meystel, 2001]. The egosphere is defined as "a spherical coordinate system with the self (ego) at the origin." Albus adopts Klatzky's definition of egocentric space [Klatzky, 1997] and transforms it into the egosphere. In the egosphere, each location in the world occurs at a specific azimuth and elevation, i.e. polar coordinates. Albus proposes using several egospheres to represent different aspects of the world. For example, he describes a sensor egosphere which sits at the origin of any sensor allocated to a robot (e.g. camera, sonar). Other examples are the head egosphere and the body egosphere.

Short-Term Memory

Short-term memory (STM) is needed for storage of sensory data and the integration of sensory data over time for further processing [Albus and Meystel, 2001]. Such a memory is considered “short-term” because the time interval over which data is stored is approximately equal to the planning scope at which tasks are implemented [Albus and Meystel, 2001]. That is, the robot should not hold all sensory information it detects for arbitrarily long periods of time. The robot’s STM should retain information relevant to the current task situation. For the work in this dissertation, the sensory data need only reside in memory for as long as it is useful to the current task or is transcribed by some other computational process (e.g. a learning mechanism).

The main reason that STM is considered in this work is that measures the attention network associates with events in short term memory need to decay over time. If the values of sensory events are not decreased over time, the robot has a much higher chance of attending to information that may no longer be true. The less time an irrelevant item sits in memory, the less chance it has for becoming a focus of attention.

STM plays a role in both attention and event binding in this work. In binding sensory events, STM must store and track the events. Events that occur together may not be detected at the same time due to different latencies of the SPMs (i.e. each SPM has its own processing latency, some longer than others). This time discrepancy can be resolved in STM. STM is needed for attention to monitor events involved in the focus of attention and to keep those events at hand for further processing. Below are descriptions of STMs used in other robotic systems.

Artificial neural networks (ANN) are used in many learning and navigation tasks for mobile robots. ANNs function as short-term stores since past sensory data are stored into the network. ANNs function to partition the sensor space so that sensor inputs correlate with specific regions of the sensor space, that is the network can predict future sensory measurements by using past sensory measurements. This is useful for robots that operate in a constant and/or static environment. However, for attention, salience needs to be assigned to sensory data which in turn is ranked by its saliency. The SES is a short-term repository for sensor input that can perform this saliency rank. Other computational modules in the robot can access the data structured by the SES. These other modules may well include ANNs. Moreover, non-SPM routines can interpret data on the SES. The SES is more appropriate for the work developed in this dissertation than ANNs because the SES can act as an interface between the robot's cognition and sensing. ANNs provide cognition, not an interface. Examples of ANNs used as STM in robotic systems are Hidden state or Hidden Markov models [Baldi and Chauvin, 1993; McCallum, 1996; Drescher, 1991; Littman, 1993; Chrisman, 1992] and recurrent neural networks [Floreano and Mondada, 1996; Meeden, 1996; Nolfi and Floreano, 2000; Ziemke, 1999; Ziemke and Thieme, 2002].

Kayama et al. describe their implementation of a visual short-term memory for a robot [Kayama et al., 1998]. The purpose of the STM on the robot is to recall objects the robot has discovered in its environment. The authors create a panoramic mosaic of visual descriptions taken from images of the robot's environment. As images of the environment are snapped, color segmentation is performed and segmented areas are labeled as regions. Each known region is tagged with spatial and temporal information. This tag information is stored in a database. The spatial information characterizes the regions (i.e. segment's size, shape, color) while the temporal information describes the history and variation of a region. A network is formed to describe the topology of regions in adjacent images. The information produced by the STM is solely used for memorization of the environment, though. No other skills of the robot use the information provided by the STM. The authors demonstrate their STM with the humanoid robot Saika. In the demonstrations, the robot is allowed to make a panoramic mosaic of its environment. The authors then moved, removed or occluded an object in the environment. Saika was then asked to find the original object. In the cases where the object was not removed from the environment, Saika found the object. When the object was removed, Saika began to scan the environment again. This system is very similar to the SES - it is a topographical map that stores short-term sensory information about the environment in an egocentric manner. An advantage that the SES does have over this system is that the topology between objects already exists in the structure of the SES; no mathematical computations need to be formed to link objects together because the nodes which store events are already linked. This type of topology is important for future attention and event binding calculations.

Soyer et al. use their bubble model, described in the egocentric representation section, as a short-term memory, also. If a robot detects a visual feature, a bubble is formed at the current fixation point of the robot and the quantitative measurement of the feature is used to deform the bubble in the direction of the feature. When the robot returns to the location at a later time, it can recall the stored sensory information. The information in the fixation point is logistically stored as a 2D array in computer memory. Once again, the computational complexity of this method is much greater than that of the SES, whether it is in the ECM or the STM.

The egocentric markers developed by Brill et al. and described in the ECM background serve as a short-term memory, also [Brill et al., 1995]. The markers are egocentric memory data structures that are capable of maintaining sensory information that is either occluded or not in the robot’s current field of view. The markers store the sensory information’s relevance to the current task and the location of the stimuli in the robot’s egocentric frame. The locations of the markers are updated in short-term memory by either dead-reckoning, measurement of acceleration or optical flow methods. The advantage of this system is notably in the ability of the robot to have access to occluded or unseen information and in the methods of updating STM for mobile robots. Again, though, the markers have no direct method of linking objects to each other so as to bind sensory information, which is a necessity in this dissertation's work for attention and event binding.

Cañas and Garcia-Alegre describe their use of occupancy grids to generate and maintain a representation of the robot's local environment [Cañas and Garcia-Alegre, 1999). As the robot navigates through its environment, a decision function integrates sonar readings and determines whether or not grid cells are occupied. Grid cells are segmented according to whether or not they are occupied. The main purpose of this STM is to represent obstacles to the robot in a more abstract manner than that of actual sensor data readings. For stationary robots, the occupancy grid method is a reasonable alternative to the SES, although it is not shown if the occupancy grid would be useful in binding coincident sensory events and in performing saliency-based attention. That is not to say that these functions could not be applied on an occupancy grid. Indeed, the SES can be viewed as a kind of occupancy grid, only one that is indexed by directional coordinates rather than Cartesian coordinates.

Sensor Integration

Sensor integration is the combining of sensory information detected at roughly the same time from different sensors to form a percept [Masumoto et al., 1994; Dudai, 2002]. It is also described as the combining of stimuli that emanate from different sensory modalities in multiple spatial locations [Maravita et al., 2002]. Most sensor integration work in the robotics field has focused on mobile robots and how to navigate or localize robots by fusing the few sensor signals available to the robots. Kam et al. group methods into low-level fusion and high-level [Kam et al., 1997].

Low-level fusion processes usually send their output to a map-making algorithm or a path-planner. Kalman filters are widely used for low-level sensor fusion when the statistics of the system and sensors are known, with the objective of fusion being to create an environment model for the robot. In centralized architectures, Kalman filters are applied to a range of sensory data for use as environment models and position monitors for robots. Cox used Kalman filters in his robot Blanche to fuse incoming sensory data [Cox, 1991]. Hong and Wang used Kalman filters to fuse noisy and fuzzy sensory data [Hong and Wang, 1994]. Others have used Kalman filters for visual mapping and automatic guided vehicles [Ayache and Faugeras, 1988; Borthwick et al., 1994].

In systems that process different sensory modalities with different techniques or that depend on patterns found in maps, Kalman filters prove to be very difficult to apply [Kam et al. 1997]. Instead, some developers use rule-based methods to fuse sensory data [Flynn, 1988]. Rule-based methods are simple to implement but do not generalize to different environments very well [Kam et al., 1997]. Geometric and topological maps have become a popular method of fusing sensory data from many different modalities. Blanche, Cox's robot, uses egocentric, topological maps in combination with Kalman filtering techniques [Cox et al, 1991]. The SES is an example of a topological map that also has an idealized geometric structure. However, the purpose of event binding in the context of this dissertation is to integrate co-occurring sensory events for a robot's higher-level processes.

High-level fusion processes usually integrate their output directly into the control processes of the robot architecture. These types of fusion processes are similar to the event binding process described in this dissertation. The integration of co-occurring events on the SES for skill acquisition, as described in this dissertation, can be described as a high-level fusion process. Masumoto et al. describe their hierarchical model that uses high-level sensory-motor fusion for intentional sensing [Masumoto et al, 2003]. The model uses 'processing units' to provide autonomous control to the robot. A processing unit consists of a recognition module, a motor module, and a sensory-motor fusion module. The recognition module receives low-level sensory signals from multiple sensors, converts the signal into an event and sends the event to a "higher layer" and to the sensory-motor fusion module. The sensory-motor fusion module receives input from the "higher layer" in the form of sensory goals and from the recognition module. The role of the sensory-motor fusion module is to predict changes in the sensory environment that are produced from actions the robot takes in its surroundings. The sensory-motor fusion module sends its predictions to the recognition module while motor commands from the "higher layer" are sent to the motor module. The motor module then converts these "higher layer" motor commands to low-level commands for the robot. Basically, this system contains SPMs that convert sensory stimuli into events. These events are fused with goals of the system to determine what motor commands to perform next and to predict the changes that might result from the motor actions. The system developed by Masumoto et al. is similar to the attention and event binding developed in this dissertation in that sensory events are combined with information about the robot's goals to create an output. However, this system does not seek to drive motor commands and predict the changes the motor actions might have on the robot's surroundings. The objective of the system is to provide output for skill acquisition processes. This approach is useful in detecting co-occurring sensory events with respect to the robot's task. It is also useful in predicting the environmental changes caused by the robot's actions. The method does not afford for storage of sensory events or for mapping of these events. The events detected can only occur within the robot's field of view.

Tremblay and Cutkosky describe their sensor fusion approach for dexterous manipulation [Tremblay and Cutkosky, 1995]. The goal of their work is to use sensors and context to reliably detect events in the dexterous manipulation of objects. The task in this research is decomposed into phases or episodes, each of which is associated with possible events. Each of these events has a set of sensor-based and context-based features associated with it. The sensor-based features consist of fingertip position error and filtered force. The context-based features consist of desired acceleration, force-velocity dot product and desired fingertip velocity. Events consist of fingertip contact, finger acceleration and unknown disturbances. As features are detected by the tactile sensors, they are given a confidence value assigned by a confidence distribution function. The overall confidence for an event is the weighted sum of its associated features' confidences, with the weights being assigned *a priori*. The objective of Tremblay's and Cutkosky's fusion, though, is event detection for dexterous manipulation. The processing of the system is hard-wired for the specific sensors on a robotic hand.

Lang et al. described their adaptation of the anchoring process to fuse together coincident sensory events [Lang et al., 2003; Coradeschi and Saffioti, 2001]. The authors developed what they call 'multi-modal anchoring' to identify and track people in the environment. The authors define multi-modal anchoring as the process of linking symbolic representations of objects in the world (i.e. "person") with the sensory representations of these objects (e.g. detected face). These representations are bound together to create an anchor. The connections are dynamic so as to allow tracking of multiple objects. Every time new sensory data is found, it is anchored to a new symbol. (Although this system is developed on a mobile robot, the robot does not move during any of the experiments used to test the multi-modal anchoring so that multiple views of a single object do not need to be considered.) Once a symbol is established, the sensory events that describe that symbol are bound with the symbol until the events no longer exist. In this system, the only symbol that the robot can anchor to representations is ‘person’. The only representations available for the robot are face detection, leg detection using sonar and sound source localization. The multi-modal anchoring system was shown to be highly successful at tracking multiple people both in a laboratory situation and in a crowded conference hall. The anchoring process is skillful at binding together co-occurring sensory events, although only three types of sensory events are detectable for the specific robot and there is only one possible type of anchor. The main drawback of this system is that the symbol and events must be known to be bound. An anchor cannot be created for an unknown symbol or using unknown/unexpected events. This is crucial to the system developed in this dissertation because the events that co-occur are not presumed to be known or to co-occur from a known source.

Attention

The main purpose of attention in this thesis is ‘selection-for-action’ as described by Balkenius [Balkenius, 2001] with the ‘action’ being skill acquisition. The selection process locates sensory events (selection) that provide input for skill acquisition (action) on both Robonaut and ISAC. The attention systems reviewed below include uni-modal attention systems and multi-modal attention systems. Some of these systems include goal-driven control while others use only bottom-up salience to drive the direction of attention.

Uni-Modal Attention Systems

Cave developed FeatureGate, a model of top-down and bottom-up influences on visual attention [Cave, 1999]. FeatureGate uses visual images as input to the system and selects the region that is most different from its neighbors ("pops-out") and most closely matches the target. FeatureGate was developed for use on ISAC by Driscoll [Driscoll et al, 1998]. Wolfe's Guided Search model also used top-down and bottom-up influences in visual attention [Wolfe, 1994]. Both FeatureGate and Guided Search compute a focus of attention in images by selecting features that stand out from neighboring features (bottom-up) and by selecting features that match a visual target (top-down). The image pixel that is most different from its neighbors and most closely matches the target wins the focus of attention. Both of these visual attention systems have influenced the attention network developed in this dissertation. The attention network uses both bottom-up and top-down information to drive the focus of attention; however, in this attention network, the attention network seeks areas whose salience values pop-out (i.e. multiple events appearing in one area) and areas whose events are relevant to a goal. While the systems developed by Cave and Wolfe are applicable only to visual images, they also serve as background for attention systems presented later in this section.

Itti et al. model salience in visual scenes also to direct attention to pop-out regions [Itti et al., 1998]. The model is similar to FeatureGate and Guided Search; however, this version of the attention system only searches for pop-out areas. Goal or target information is not used in this system. It has, however, been adapted to perform goal-directed visual attention, both by the authors and by others. Navalpakkam and Itti developed a goal-oriented attention model for extraction of task-relevant objects in a scene [Navalpakkam and Itti, 2002]. The authors expanded the original visual attention system with a topographic task-relevance map that encodes the relevance of every visual location to a robot's current task. The authors' motivation is to save computational complexity by tracking only events/objects that have an expected relevance to the current task. Their architecture consists of four parts: a visual brain, working memory, long term memory and an agent. The visual brain maintains a salience map, a task-relevance map and an attention guidance map. The salience map contains a salience value for each point in the input scene. The salience values are determined by Itti's visual attention system described above [Itti, et al., 1998]. The task-relevance map holds the relevance of each point in the input scene. The task-relevance is determined when one of these image points becomes the fixation point of the robot, which is the location to which the robot is attending. The relevance of that point is compared to the contents of the working memory and assigned a relevance value. The attention guidance map is simply the product of the salience map and the task-relevance map. The working memory maintains what visual objects are expected to be relevant to the current task. The long term memory holds knowledge about the real world and about abstract objects. The agent dispatches information between the visual brain, working memory and long-term memory. The authors are adamant that the agent is not a homunculus; it is simply an information relay. The main advantage of this system is that the search space is pruned before relevance is determined. This is efficient compared to systems that scan the entire visual space and then assign relevance to areas. Also, the use of separate salience and task-relevance measures is comparable to the design of this dissertation's attention network. The disadvantage of the system is that directed shifts of attention cannot be made – salience is based on what is known about the task and how visual features in the environment relate to the task. Therefore, if the system were to be adapted to include multiple sensor modalities, only task-relevant events could drive focus of attention.

Breazeal presents her attention system for Kismet using context to focus attention in a visual space [Breazeal, 1999; Breazeal, 2002]. The objective of this attentional system is to direct the limited computational resources of the robot and to select among the appropriate behaviors to allow Kismet to act in a complex social manner. The bottom-up component of the attention system is modeled after Wolfe's visual attention model and uses salience-based maps like Itti et al. [Wolfe, 1994; Itti et al., 1998]. The top-down influences are controlled by Kismet's motivation and behavior system. The bottom-up features enhance areas in the visual space in which people could be. The top-down motivations drive the social desires of the robot. These two sets of influences are combined with a habituation map to determine the most salient location in Kismet's environment. The system does not provide for items outside of the robot's field of view to drive attention, although with the use of both foveal cameras and wide-angle cameras this ability may not be needed. The author does note that an ego-centered salience map would allow the robot to attend to areas not in its field of view. This could be rectified by a structure like the Sensory Ego-Sphere. Also, the author does not use multiple sensory modalities to drive attention. In the next section, though, a multi-modal system is described that adapts Breazeal's attention network to include outputs from two sensory modalities.

Balkenius and Hulth developed an attention system as selection-for-action by controlling attention with bottom-up and top-down processing methods [Balkenius and Hulth, 1999]. Their attention system is a filter that decides how incoming sensory stimuli should be processed, with the system's goal being to exclusively let through stimuli that are currently relevant to the robot or to a target source. The attention system filters sensory input based on the features of a target source or on the spatial location of a target source. In this way, feature or spatial cues can direct attention to a desired object. The output location of the attention system is then used as input for the robot's next action. This system is similar to that described by Albus and Meystel [Albus and Meystel, 2001] in that the most relevant area is that which is near the focus of attention. This creates quick response times. While this system could be useful in solving the problem of directing a robot's attention to important areas while ignoring irrelevant information, the system does not function on a structure. Without a structure to hold sensory information, salient events cannot be bound together. The system provides no method for integrating spatio-temporal coincident events once a focus of attention is found.

Multi-Modal Attention Systems

Albus and Meystel describe attention as "a mechanism for allocating sensors and focusing computational resources on particular regions of time and space" [Albus and Meystel, 2001]. The authors explain it as part of their hierarchical control system developed for intelligent systems: Real-time Control System (RCS) [Albus and Meystel, 2001]. They claim that a hierarchical architecture can facilitate the focus of attention by allowing the higher levels to decide what sensory data is important while the lower levels use the information passed down to actually focus attention. To determine which information is important, the authors use a top-down and bottom-up approach. Behavioral goals produced by the higher levels of the RCS tell the system what is important. This accounts for the top-down influences in the system. Unexpected, unexplained and unusual events detected by sensory processors let the system know that the world model is incomplete or needs to be updated. This accounts for the bottom-up influences in the system. As in other systems, top-down influences are revealed in behavioral goals while bottom-up influences present themselves in salient aspects of the environment. In relation to the egosphere, the relevance of data on the egosphere is inversely proportional to both their spatial distance from the origin of the sphere and their temporal distance form the current time of the egosphere. However, some information that is far away is also very important. Behavior goals influence the focus of attention on events that may not seem relevant in a spatio-temporal distance relation but are relevant for that particular goal. The similarities of the approach of this system to that developed in this dissertation are that 1) both bottom-up and top-down influences are used to direct attention, 2) the sensory information used in directing attention lies on the egosphere and 3) events closest to the current time have the most relevance. However, appropriating relevance to sensory information based on spatial distance from the origin of the egosphere is not entirely appropriate for this system. For example, the closing of a lab door should be a salient event because it may indicate a person has entered the room. Using the egosphere, the events signifying the closing of the door would receive less salience than the detection of a tool lying in front of the robot.

Two other systems described below include the possibility of using multiple sensory modalities to drive attention without actually using multiple sensors. The first is an overt visual attention system developed by Vijayakumar et al. [Vijayakumar et al., 2001]. The objective of this system is similar to the overall aim of this dissertation in that the attention system provides input to control systems of a humanoid robot. The visual attention is driven by a bottom-up saliency map and suppresses areas which contain irrelevant inputs. The authors also suggest that the system can be made multi-modal by weighting the inputs from different sensors in the saliency map. However, no attempt at this was made for comparison to the attention network in this dissertation. Gonçalves also developed an attention system that can be controlled in a bottom-up or top-down manner [Gonçalves, 2001]. The system is only tested using task-relevant features and does not actually include any bottom-up information, though. These two attention systems could be adapted to include sensory events from multiple modalities. The disadvantage with these two systems is that neither system retains the sensory information for separate event binding. The sensory events are not stored for easy recall so that determining which events co-occurred both spatially and temporally is not possible.

Déniz et al. adapt Breazeal's attention system to include sound information and to be guided by high-level modules [Déniz et al, 2003]. The types of high-level modules are not specified, but the authors note that the modules represent task-driven process. Therefore, it is assumed that attention can be shifted to areas of high task-relevance away from high salience due to other sensory stimuli. The authors admit, though, that the high-level modules are only part of the design and have not yet been implemented in the system. Since no top-down influence can be provided, the only difference between this system and Breazeal's system is that sound information can modulate attention. In an experiment described in Chapter 4, the sound data is provided as a cue for the robot to attend to the visual object closest to the sound. The system does illustrate that sensory events can illicit a cueing effect for other events (e.g. a sound can direct attention to a nearby object). This is a beneficial function when visual events that the robot should attend to are out of the robot's field of view. The system developed in this dissertation is tested in an experiment similar to that described by Déniz et al. If a data structure like the SES was included in this system, it would be very useful in solving the problems presented in this dissertation. The system did not influence the attention network in this dissertation, though, because the first publication of the system occurred after the proposal of the SES-based system.

Finally, the most intriguing attention system reviewed was developed by Lang et al. [Lang et al., 2003]. The authors developed a multi-modal attention system for a mobile robot that is only used to attend to people. The authors use a multi-modal anchoring process to anchor face detection, leg-detection and sound localization outputs together. Attention is focused on the person that the robot decides is addressing it. Attention remains focused on that person until he or she has stopped talking for more than two seconds or until the anchor for that person can no longer be sustained. Multiple people can be tracked in Lang's system and attention is shown to be quite accurate, as is discussed in Chapter 5. The results show that Lang's system is quite successful in attending to a person who is speaking. But, the purpose of Lang's attention system is to identify persons-of-interest (POI) and maintain attention on the POI until another POI is found. The purpose of the research in this dissertation is to identify areas of interest in the robot's environment and pass on the sensory events that contributed to the interest. Lang's system can currently only attend to a known object using the anchoring method. The system in this dissertation does not assume that an event source is a known object. The SES-based system needs to bind together any events that may have occurred as a result of a sensory source. Lang's system is dependent on knowing the events that sensory sources produce and binding these sources before attention is allocated.

Summary

Some of the reviewed methods for ECM and STM have suitable features for facilitating event binding and attention. However, most methods exclude topology between detected features or events. This is a required element of the SES. Without the topological links between sensory events, those that co-occurred cannot be bound together spatially. Also, the attention network needs topological links so that salience spread in an area may accumulate due to multiple events detected in that area. As for handling multiple sensory modalities and multiple resolutions, none of the mechanisms explicitly handle both of these issues. However, some may be extended to have these capabilities. Since the SES previously existed on both ISAC and Robonaut and it has all of the functionality needed for binding spatio-temporal coincident events and for saliency-based direction of attention, the SES remains the unified mechanism upon which these processes are developed.

Most of the methods of sensor integration that exist have the objective of combing streams of sensory data for path-planning and navigation of mobile robots. The few methods found that serve a purpose similar to that described in this dissertation were not sufficient for the objectives in this work. Lang's multi-modal anchoring operates very well when the objects receiving anchors are known. However, none of the methods store all detected sensory events originating from (possibly) unknown objects for the detection of spatial and temporal coincidence. Some of the attention systems could have been applied in this dissertation. However, those that could function on the SES or any single data structure were not documented until after the SES-based system was designed and implement.

In summary, no unified mechanism has been found that can implement ECM and STM as desired for the purpose of facilitating attention and event binding. No sensor integration method has been found that can select spatially and temporally coincident sensory events from unknown sources. No attention system was clearly developed before the proposal of the SES-based system that could operate on a data structure and select a focus of attention based on both the appearance of sensory events and their relation to a robot's tasks. Therefore, the system introduced in Chapter 1 and described in Chapter 4 serves as the solution to the following questions.
Which events detected by a robot's SPMs belong together by virtue of having been produced by the same source or in response to a single action of the robot? Which events should be ignored as irrelevant or spurious without missing those that indicate danger or opportunity?

CHAPTER III

SYSTEM PLATFORM

This chapter presents the platforms on which the methods in this dissertation are developed and evaluated. The two robots used to test the methods for attention and event binding are ISAC, Vanderbilt University's cognitive humanoid, and Robonaut, the DARPA/NASA humanoid. These two robotic platforms are presented first. A functional description of the Sensory Ego-Sphere and its implementation on both ISAC and Robonaut are then given.

Robotic Platforms

ISAC (Intelligent Soft Arm Control) is a research-oriented humanoid robot developed at Vanderbilt University [Kawamura et al., 2002, 2001]. ISAC has two 6 degrees-of-freedom (DOF) arms that are controlled pneumatically by McKibben artificial muscles [Klute et al., 1999]. ISAC has two hands that operate under a hybrid pneumatic-electric power system [Christopher, 1999]. ISAC also has an active vision system operating on two 2-DOF pan-tilt units [Srikaew, 2000]. Microphones on either side of the robot enable sound localization [Liu, 2001]. An array of five infrared motion sensors under the pan-tilt units enables infrared detection of motion [Sekmen, 2001]. All of ISAC's software modules use the Intelligent Machine Architecture (IMA) [Pack, 1997]. IMA provides distributed computing across multiple processors. IMA permits multiple SPMs to operate in parallel so that sensory events from different modalities can be detected simultaneously. ISAC's controllers, SPMs and the SES operate on four Pentium XEON processors and two Pentium 4 processors. Figure 2 shows the humanoid ISAC.

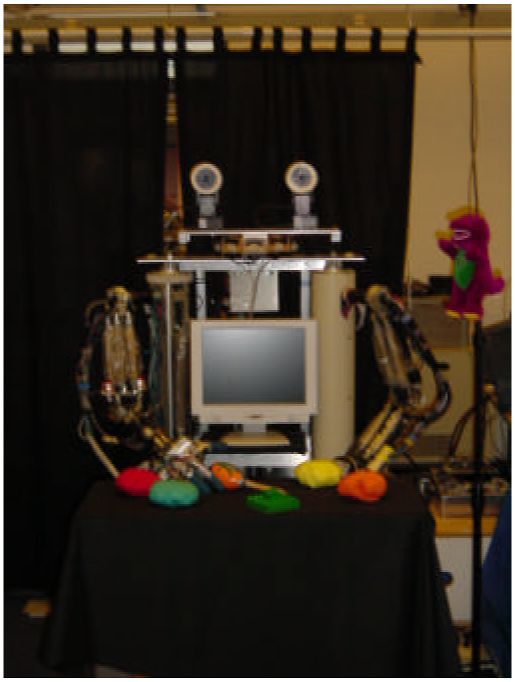


Figure 2: The Humanoid Robot, ISAC

Robonaut is NASA's humanoid robot that will eventually serve as an astronaut assistant and perform extravehicular activity duties on the International Space Station [Ambrose et al., 2001]. The humanoid is currently attached to a fixed point for research purposes. It can perform articulated motion within its frame and has over 50 DOF to do so. Robonaut has two five-finger hands that are used for dexterous manipulation, with each of the hands having 19 DOF [Diftler et al., 2003]. A range of image processing routines serve Robonaut's visual system; however, only visual object recognition and tracking are used in this dissertation [Bluethmann, 2003].

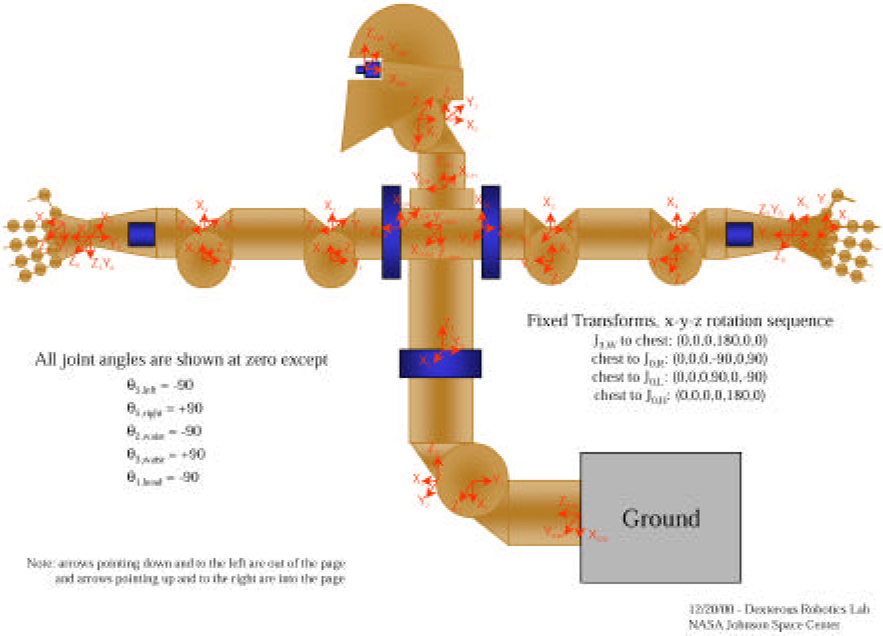

All joint angles are shown at zero except
$\theta_{5,left} = -90$
$\theta_{5,right} = +90$
$\theta_{2,waist} = -90$
$\theta_{3,waist} = +90$
$\theta_{1,head} = -90$
Fixed Transforms, x-y-z rotation sequence
$J_{2,W}$ to chest: (0,0,0,180,0,0)
chest to $J_{0,R}$: (0,0,0,-90,0,90)
chest to $J_{0,L}$: (0,0,0,90,0,-90)
chest to $J_{0,H}$: (0,0,0,0,180,0)
Ground
Note: arrows pointing down and to the left are out of the page
and arrows pointing up and to the right are into the page
12/20/00 - Dexterous Robotics Lab
NASA Johnson Space Center

Figure 3: Robonaut

Sensory Ego-Sphere

Geodesic Dome Topology

A geodesic dome serves as an implicit structure for the SES since it is a quasi-uniform triangular tessellation of a sphere into a polyhedron [Edmonson, 1986; Urner, 1991]. Stewart defines it as “the optimal solution to the problem of how to cover a sphere with the least number of partially overlapping circles of the same radius” [Stewart, 1991]. The triangles connect at vertices forming twelve pentagons and a variable number of hexagons. The pentagons are evenly distributed so that the node at the center of one is connected to the centers of five others by *N* vertices, where *N* is the frequency of the dome. The number of vertices, *V*, in the polyhedron as a function of the frequency is given in Equation 1.

$$V = 10 * N^2 + 2 \tag{3.1}$$

To form a sphere, the vertices of the polyhedron are equalized from the center. Figure 4 illustrates the progression of a polyhedron from a frequency of 1 to a geodesic dome with a frequency of 4.

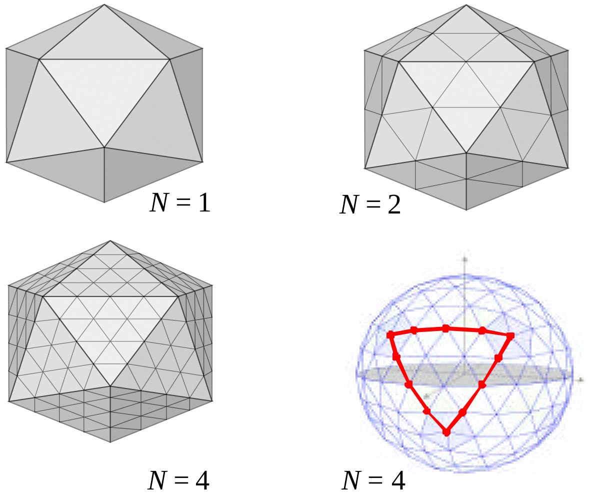


Figure 4: Tessellated Polyhedrons and a Geodesic Dome

A dome with frequency of one is an icosahedron which has 12 vertices, each of which connects with 5 neighbors. A dome with frequency of two is constructed from the icosahedron by placing a new vertex at the midpoint of each edge and connecting each new vertex with the four nearest neighbors to which it is not already connected. This subdivides each triangular face into four new triangles. Globally, the process adds a set of hexagons to the construction; of the 42 vertices in the result, the 30 new ones are connected to six neighbors while the twelve original vertices remain connected to five. For the SES, a vertex neighborhood is defined as the five or six neighboring vertices around the central vertex. A neighborhood of one around a vertex results in 5 or 6 vertices. A neighborhood of two results in the original neighbor vertices and all of their neighbor vertices. The neighborhood definition helps to facilitate attention and event binding methods described in Chapter 4. To be useful as a sensory data structure, the tessellation frequency must be determined by the resolution of the various sensors on the robot.

Database

From an information processing standpoint, the SES is a multiply-linked table of pointers to data structures. Ideally, there are 6 or 7 pointers for each vertex on the dome, one to each of its 5 or 6 nearest neighbors and one to a variable length list, which can be contained within multiple database tables. The non-neighbor list items are pointers to tagged-format data structures, each of which is a database record that contains an alphanumeric tag followed by a time stamp, the event's spatial location and a terminated list of other pointers. The pointers and the list reside in memory for fast access whereas the data structures exist in a standard database. Each tag indicates the modality and type of the event. The corresponding time stamp indicates when the data was registered onto the SES. The spatial location indicates the actual direction of the source of the detected event. The pointers associated with the tag hold the locations of other records pertinent to the data type such as the sensory data itself or any function specifications associated with it that may be provided by the SPMs (e.g. the name of a recognized face, the confidence with which an object recognizer detected its target). The number of tags and their types on any vertex of the dome are completely variable.

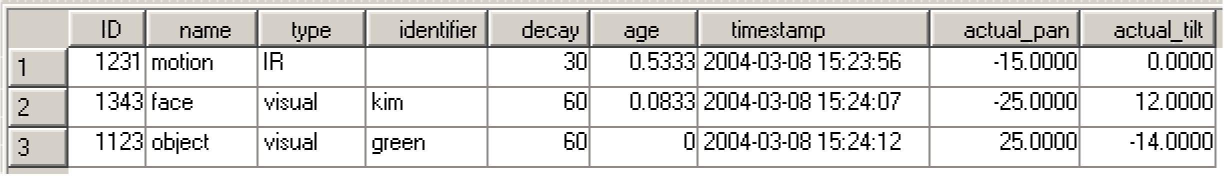
For both ISAC and Robonaut, a MySQL server provides the database. The actual database for ISAC's SES contains tables: one for nodes on the geodesic sphere, one for registered event and their information, one for the last events registered by each SPM (used for habituation), one for the robot's tasks (used for assigning task-relevance to events), one for current salience of each event per node and one for current salience per node (to determine the focus of attention). Table 1 lists the database tables for ISAC and brief descriptions of the information these tables hold.

Table 1: SES Database Tables and Descriptions

Table Name	**Table Contents**
Nodes	Indices, angles, neighbors for each node
SES	Registered events and their tags
History	Last registered event from each SPM
Activation	Incidence, task-relevance, habituation per node contributed by each registered event
Task-Relevance	Current tasks and their descriptions
Attention	Total salience per node

The actual database for Robonaut contains the above tables and two extra tables: a table for the robot's actual position at time of registration, and a table for extra data attributes. These last two tables in Robonaut's database were requested by Robonaut developers at NASA's Johnson Space Center; they are not pertinent to the work in this dissertation.

The node positions table contains the (azimuth, elevation) angle pairs for each vertex on the sphere, an associated integer index pair per vertex, a node identification number, and all neighbor node identification numbers. The integer index pair assigns an i-index to a node to identify its elevation location and a j-index to identify its azimuth location. This pair facilitates efficient neighbor-node finding while the node identification number allows for simpler queries. The table for registration data contains the following tags about registered data as fields in the table: name, type, identifier, actual azimuth and elevation angles, time of registration, age of the event and an age limit. Figure 5 shows a sample of records from the registered data table.



	ID	name	type	identifier	decay	age	timestamp	actual_pan	actual_tilt
1	1231	motion	IR		30	0.5333	2004-03-08 15:23:56	-15.0000	0.0000
2	1343	face	visual	kim	60	0.0833	2004-03-08 15:24:07	-25.0000	12.0000
3	1123	object	visual	green	60	0	2004-03-08 15:24:12	25.0000	-14.0000

Figure 5: SES Database Sample

In Figure 5, the name is the SPM that detected the event while the type is the sensor that sends input to the associated SPM. This separates events that report similar information from different sensors (e.g. visual motion from infrared motion) and that report different information from the same sensor (e.g. visual green object detection from visual face recognition). The identifier serves as an extra descriptor of the event. These three tags are established by the SPM that sends the event to the SES. The actual angles refer to the original direction at which the event was detected, so that the robot can return to the original location of the event. The time of registration serves as the event’s timestamp while the age denotes how old the data is, respective to its timestamp and decay. The decay value is the amount of time in seconds that the event should remain on the SES. The all data tables are linked to the nodes’ table by the node identification number of the node that receives data registration. Each field may or may not be assigned a value.

Communication Managers

SPMs write information to the SES through a software agent called the SES Manager which in turn interfaces to the database. For ISAC, this manager is a Visual Basic 6.0 application that communicates to ISAC's other components as an IMA software module. The SES Manager on Robonaut is a Visual C++ 6.0 application that communicates with other system components via an information stream controlled by NDDS software.

The SES Manager provides all current functionality of the SES, with the exception of purging and decaying of records which is handled by a Decay Manager. Requests are sent to the SES Manager which in turn either registers events onto the SES or retrieves events from the SES. When an SPM requests registration of an event, the SES Manager collects all provided information about the event, creates a record in the database and marks it with a time stamp. The direction of the event, in relation to the SES coordinate frame, is used to locate the closest vertex for storage of data. To determine the vertex closest to the original direction, distances between the azimuth and elevation of the event's detected location and vertex angles are computed. The maximum distance between vertices is used as a bound for this maximum distance. The maximum distance[2] between vertices is dependent on the tessellation (e.g. for a tessellation of 14, the maximum distance between vertices is ~5.92°. In this situation, all vertices whose angles are within about 6° of the event's detected location are selected and a difference is taken for each vertex angle pair.) The vertex with the smallest distance measurement is selected as the registration node. If the registering sensor is in a different coordinate frame than the robot's SES, articulated motion transformations are performed between coordinate systems. (This

[2] The distance is computed as Euclidean distance in azimuth-elevation space. Spherical distance was not considered before testing of this system; spherical distance is mentioned in future work and will be applied to future versions of the SES.

transformation is described in the Appendix.) Once the transformed angle pair is computed, the closest vertex on the SES is found and the data is registered at that node. Figure 6 illustrates the projection of an object onto the SES and the vertex onto which the object is projected. In this figure, the sensor that detected the event and the SES exist in the same coordinate frame.

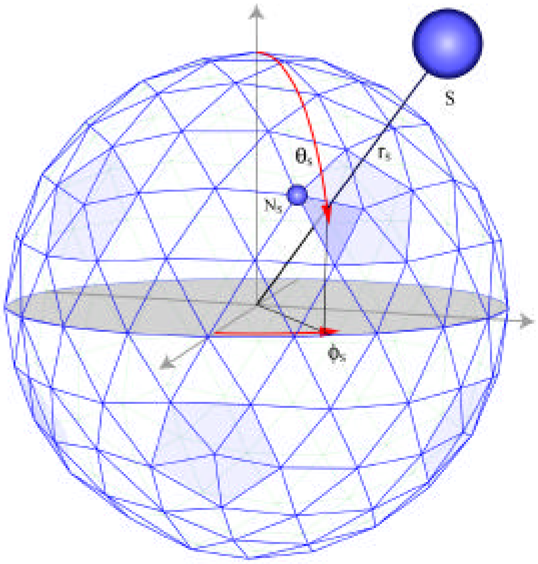


Figure 6: Projection of an Object onto the SES

A system component can requests retrieval of data using any data tag in the database tables (e.g. data name, data location, data age, etc.) The SES Manager queries the database using the specified tag. All retrieved data is returned to the requesting agent. If the request is of a location, the SES Manager finds the vertex closest to the desired location. If the requesting agent specifies a neighborhood size of one to include in the search, all data registered at the central node and its neighbor nodes is returned to the agent. If no neighborhood is specified, all data at the closest node is returned to the agent. Since the vertices on the geodesic dome serve as nodes in a graph, the fixed number of nodes keeps the search time fixed as the amount of data on the sphere increases.

A decay manager exists to purge old data from the database. The decay manager uses the data's temporal decay limit to determine when an event has expired. The decay manager also decays the salience of events as those events age. When the decay manager finds that the timestamp of an event has expired, the record and all information (including that contributed by the attention network) concerning that event is purged from the SES database.

If a robot moves with respect to its fixed frame, a set of equations must be applied to transform sensory events from their sensors' coordinate frames to the coordinate frame of the SES, if it is different. Although Robonaut is fixed to a frame, it can move its body so that the coordinate frames of the sensors and joints move in space, also. ISAC, however, cannot alter the coordinate frames of its sensors. The camera coordinate system defines the SES coordinate system; all other sensors (i.e. hand sensors, arm sensors, IR sensors and sound sensors) report their events within the camera coordinate frame. For Robonaut, the SES coordinate frame is centered at the robot's base while sensory events are detected in the head coordinate frame, the hand coordinate frames and the arm coordinate frames.

Figure 7 shows the different coordinate frames. The mathematical solutions to this problem are also given in the Appendix.

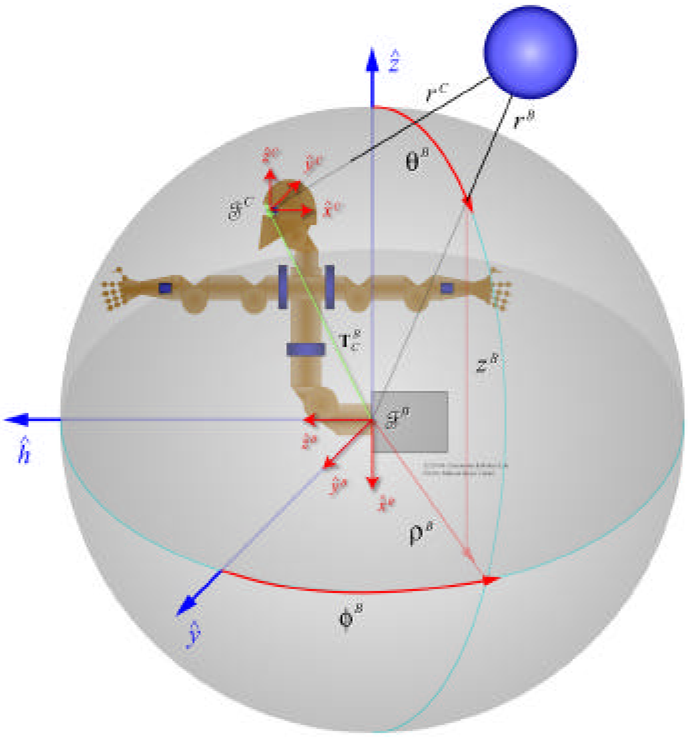


Figure 7: Coordinate Frames of Robonaut's SES and Active Vision System

CHAPTER IV

METHODS AND MOTIVATIONS

This chapter presents the methods used to develop the attention network and event binding mechanism and the motivations behind these methods. The aims of the methods described in this chapter are to locate the most salient area in a humanoid robot's environment and to transform sensory events that produce this salience into a collection of co-occurring events. Salience may be caused by stimuli coincidence (e.g. loud noise combined with sudden movement) or by stimuli relating to a current task or goal (e.g. detection of a face when the goal is to greet people). The purpose is fulfilled by combining an attention network with an event binding mechanism. Both of these components are detailed in this section.

First, the approach and reasoning for the attention network are presented. The attention network section describes methods of applying salience to SES nodes based on incidence, task-relevance and habitualness of events. Then, the event binding mechanism is detailed. The chapter concludes with a description of the information flow from SPMs to the output of the event binding mechanism.

Information Flow

The flow of information in this system is illustrated in Figure 8.

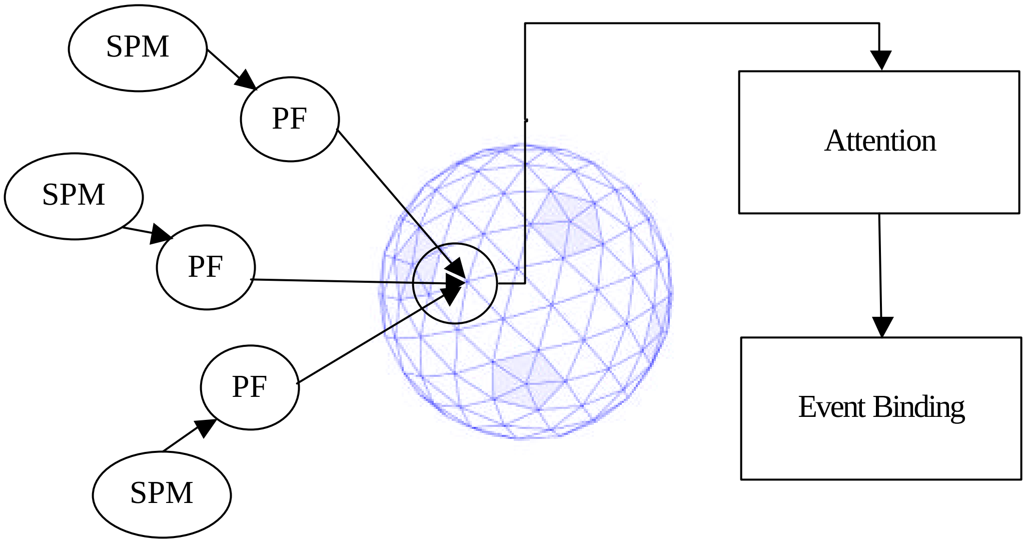


Figure 8: Information Flow from SPMs to Event Binding

Sensory information begins as stimuli in the robot’s environment. Sensors detect these stimuli and send their signals through SPMs attached to the sensors. SPMs structure the sensory signals to indicate if an event has occurred. When an event is detected, the SPM sends this event to an attached pre-filter (PF). The PF examines the event to determine if it is relevant to the current task or is habitual. An event is denoted as habitual by the PF if it occurs with a regular time period and in the same location. Examples of habitual events are motion or sound detected continuously in the same area. The pre-filter then sends the event detected, with its associated task-relevance or habituation information, to the SES for registration. When the SES registers the event, salience is spread to neighboring nodes. This salience is a combination of event incidence, task relevance and habituation values. While the robot functions in its workspace, the attention network scans the SES to find the location of highest salience. If the highest salience value is zero, then the robot is told that no salient area exists. This could occur if no people are in the robot’s environment, the robot is not interacting with its surroundings or few SPMs are running.

Otherwise, the location with the highest salience is sent to the event binder where co-occurring events are bound and sent out to higher-level processing areas of the robot.

Attention Network

The purpose of the attention network is to determine the most salient area of the robot's environment, whether it is related to the robot's task or is unexpected. The attention network scans the SES to find areas high in both incidence and in task-relevance and selects the area that meets this goal. The attention network only considers SES nodes that have either incidence or task-relevance; therefore, if no events are occurring in the robot's environment or only habitual events are occurring, then the attention network does not find salience and no winner is selected. Incidence, task-relevance and habituation of events are assigned upon SES registration.

The attention network in this system evaluates incoming events on three criteria: 1) incidence in the environment, 2) relation to current tasks and 3) habitual occurrence. The incidence measure awards salience to events simply for occurring. The purpose of this measure is to inform the attention network that an event has occurred at a specific location. Without this measure, unexpected events do not receive any salience. For example, the robot may want to find a red object for grasping but cannot reach the object because its arm is blocked by a table. In this case, sensors on the arm report events to the SES (e.g. arm movement stopped, hand proximity sensors high, hand tactile sensors high) but if these events are not related to the goal, they would not receive salience. Therefore, the incidence measure assigns salience to any event. Also, if multiple events are registered onto the same SES nodes, the incidence values of each of these events will accumulate thereby increasing salience in the area. The task-relevance measure awards salience to events that are relevant to the robot's current tasks. This measure brings to the forefront events that aid the robot in completing desired tasks or reaching desired goals. The habituation measure decreases salience of events that occur in the same location at regular time intervals. This measure keeps habitual events from achieving the highest salience continuously. The resulting salience is the final value for the area at which the event is registered.

The robots used to test this work have some sensors with low resolutions and some sensors that occasionally or frequently report errors. Therefore, the salience assigned to events is not relegated to only the SES registration nodes. Salience is spread from an event's registration node to neighboring nodes. This spread compensates for both sensor error and low-resolution sensors. Radial basis functions are used to spread incidence and task-relevance from the node at which an event is registered to neighboring nodes. Habituation is applied equally to all nodes receiving salience from an event. Habituation is not spread; instead, it is factored into the activation at all nodes receiving salience form an event.

Incidence

An incidence value is awarded to events upon registration onto the SES. The incidence of an event grants salience to activity in the environment. The incidence value is necessary to adjust direction of attention to unexpected events that do not relate to any of the robot's current tasks but may help the robot to discover new skills or avoid danger.

When a SPM registers an event onto the SES, the node N_j that is closest to the event's angular location (f_j, q_j) receives the event. Incidence is then spread along a fixed number of edges to neighboring nodes. The number of edges is determined through experiments for spatial binding in Chapter 4. Equations 1 and 2 show the calculations used to determine incidence of a neighbor node (k) with respect to its central node (j) at the time of registration (t_0).

$$I(j,k,t_0) = \frac{1}{E_{j,k}} \exp(-\frac{1}{a_I} D_{j,k}{}^2) \tag{4.1}$$

$$D_{j,k} = \sqrt{(f_j - f_k)^2 + (q_j - q_k)^2} \tag{4.2}$$

In Equation 2, $E_{j,k}$ represents the number of edges in the shortest path between the registration node, N_j, and the node receiving incidence, N_k; a_I is the incidence factor while $D_{j,k}$ is the Euclidean distance between the angular locations of the nodes on the SES, given in Equation 3. The incidence factor remains fixed for each sensor and mainly exists to inflate the values. When $j=k$, $D_{j,k} = 0$ and $I(j,k,t_0) = 1$. Therefore, the incidence factor is chosen to normalize the values from zero to one and to increase the values spread to immediate neighbor nodes.

Some SPMs used in testing on ISAC report their events in only the azimuth direction, f. Therefore, no elevation angle q is available for these events. To allow for overlap of one-dimensional events with two-dimensional events that may co-occur, incidence is assigned along a range of q values. When a one-dimensional SPM posts an event to the SES, the SES finds all nodes N_j that are closest to (f_j, q_j) for $-45° \le q_j \le 10°$. These particular values are chosen because of the height of the robot and the functional range of the robot's pan-tilt units controlling the cameras. Incidence is then spread from each of these nodes N_j according to Equation 2. Because of this multiple-node spread, a single 1-dimensional event can contribute incidence multiple times to a single node. To account for this, the incidence factors for 1-dimensional events are set to be equal to half of the incidence factor for 2-dimensional events. It should be noted that the event is only registered onto the SES once, at $q_j = 0°$.

As registered data age, their incidence values are decayed. Only the incidence is decayed, rather than the entire salience, because multiple factors affect the salience. While incidence may decay with time, task-relevance only changes with the shifting of tasks or goals by the robot. Two methods of decaying incidence are tested on the system. The first method is a linear decrease shown in Equation 4. The second method uses an exponential decay, shown in Equation 5. Both depend on the age of the data as given in Equation 6.

$$I(e,j,t) = I(e,j,t_0)*(1 - Age(e,t)) \tag{4.3}$$

$$I(e,j,t) = I(e,j,t_0)*\exp(-Age(e,t)) \tag{4.4}$$

$$Age(e,t) = \frac{t - t_0}{L_e} \tag{4.5}$$

In Equations 4 and 5, the incidence for event *e* at node N_j for time *t* is decayed using the incidence of that node at time t_0, which is the time of registration for event *e*. When an event has reached full maturity on the SES, its age is one. Using the linear decrease in Equation 4, the incidence from the event is zero by the time the data has expired. However, the exponential decay in Equation 5 allows the event to contribute incidence up to the time the event expires. When an event expires, the event and all salience it contributed to the SES are purged from the database. Both methods are tested to determine which produces more accurate results.

Task-relevance

A task-relevance value is awarded to events upon registration onto the SES or at the time a task is established. The task-relevance measure grants salience to events that may aid the robot in completing current tasks or goals. Like incidence, task-relevance is assigned using a radial basis function, shown in Equation 7.

$$TR(j\ ,k) = \frac{1}{E_{j,k}} \exp(-\frac{1}{a_{TR}} D_{j,k}{}^{2}) \tag{4.6}$$

Equation 7 is similar to Equation 2 except that the task-relevance factor a_{TR} is used. The task-relevance factor is determined through experiments to find what value allows task-related events to overcome other events as the most salient.

Task-relevance is determined upon registration of data onto the SES. The pre-filter attached to the SPMs reporting data determines if the output event is relevant to any current tasks. An event is deemed relevant to a task if any of its data tags match tasks or goals as described by the robot. Data tags are defined by the PF that sends events to the SES. The data tags are defined by their modality types and their SPMs. Examples of data tags for an unknown human face are 'face', 'person' and 'stranger'. The face detection and recognition SPMs establish these data tags [Qiu, 1997]. Examples of data tags for a blue Duplo block are 'object' and 'blue', which are established by the color segmenting SPM [Srikaew, 2000]. Currently, neither ISAC nor Robonaut have methods of determining their own tasks autonomously; therefore, the tasks and goals described by the robot are user-defined. Examples of these tasks are 'grasp drill', 'look at person' and 'find green object'. In these examples, if any of the events posted on the SES match drill, person or green object, then the event is denoted as task-relevant. For example, if the robot is looking for green objects, all output from a green color-segmenter is given task-relevance. Once this relevance is determined, the data is sent through the pre-filter to the SES with its task-relevance factor. This factor is then applied to Equation 6 and task-relevance and incidence are spread to neighboring nodes. The range of values the task-relevance factor can take is empirically determined.

Task-relevance may also be determined when a new task is established. In this case, all data that match on any of the task-related indices receive relevance. When a task is established in the robot, it is retained in the SES database forever. The only attribute of the task that changes is its value. Therefore, when an event is registered, it may match a task that is not current but resides in the database. The event is denoted as being task-relevant for that particular task but receives zero task-relevance. If the task becomes current and receives a value, then all events in the SES that match the new current task are tagged as such. The nodes to which task-relevance should be spread are already tagged so that only a value has to be assigned.

Habituation

Any events that are registered onto the SES at the same node location on a regular time interval are habituated in the attention network. In this work, the time interval is determined by the developer; however, in future work, it is desired that the attention network learns what timing makes an event habitual. The habituation mechanism uses a time decay to calculate the habituation value for an event e at time step S_t. The time step S_t is incremented every time the event is registered at the same node within the same time period. The formula is given in Equation 8.

$$H(e,t) = \exp(-b_H S_t) \tag{4.7}$$

In this equation, b_H is the habituation rate and is determined through experimental testing. The salience from a habitual event is multiplied by this habituation value. A habituation value of one indicates that all salience remains while a habituation value close to zero indicates that almost no salience remains.

The following rules govern when an event ceases to be habitual [Balkenius, 2000]. Some dimension of the event has changed (e.g. location, confidence value, name of recognized face).

A period of time has passed in which the habitual event did not occur (this time period is that within which the event is considered habitual).

The event is novel (e.g. IR motion is habitual but motion detected by vision is new).

The pre-filter attached to a sensor's SPM determines an event's habituation value. If an event re-occurs within a certain amount of time, the time step for that event is incremented ($S_t = S_t + 1$) and the habituation value for that event is calculated. This factor is applied to the final salience of each node that received incidence from event *e*. An event's time step can be reset to zero if the event does not reoccur within the specified time frame, the event matches a new task or goal or the event occurs at a new location. If a habitual event does not occur for one time step, then it is may not be habitual and should receive more salience (e.g. if the event is habitual motion and motion is not detected for a few seconds, this area may require attention). In the testing of this work, the habituation can be turned on or off thereby allowing the developer to determine when habitual events can or cannot be detected. However, in future work, habituation should not be applied to an event until it has repeated a specified number of times.

Attention Winner

The attention network scans the SES on a regular time interval to find the node with the highest salience. Equations 9 and 10 show the formulas used to calculate the total salience of all nodes for all events.

$$S(j,e_n) = (I(j,e_n) + TR(j,e_n)) * H(j,e_n) \qquad (4.8)$$

$$S(j) = S(j,e_1) + S(j,e_2) + \ldots + S(j,e_n) \qquad (4.9)$$

First, the salience for node *j* due to event e_n is calculated from the incidence, task-relevance and habituation of that event (Equation 9). Then, the salience at node *j* due to all events contributing salience at that node is calculated (Equation 10). The node *j* that receives the highest salience value is selected as the winner or focus of attention (FOA). This node is then sent to the event binding mechanism.

Event binding

The objective of event binding is to group co-occurring events that originated from the same source. Event binding occurs spatially and temporally. Two assumptions are made in the method of event binding presented: 1) events that originate from the same source are more likely to occur in the same location (e.g. a bat hitting a ball produces a sound, motion and detection of a ball and a bat in the same location) and 2) events originating from the same source are likely to be detected at about the same time.

The incidence measurement used in the attention network serves to determine spatially-connected events. Whenever a winning node is chosen by the attention network, the event binding mechanism selects all events that contributed incidence values to that node. This process is referred to in the rest of this document as spatial binding. The timestamps of these events are then compared against each other; this process is denoted as temporal binding.

During temporal binding, the timestamps are inserted into an array in descending order. The differences between successive timestamps are taken. If all of these timestamp differences are smaller than a limit, the differences between every other timestamp are taken. If all of these timestamps are smaller than a limit, the differences between every third timestamp are taken. This continues until either only one difference is less than the time limit or none of the differences is less than the time limit. If one difference is smaller, it is selected as the winning difference. If no differences are smaller, then the smallest difference from the previous difference array is selected as the winner. All events whose timestamps are included in the winning difference are selected and bound together as co-occurring. The time limit for co-occurring events is determined by the latency of both the attention algorithm and the rate at which the algorithm scans the SES. The latency of the attention algorithm is assumed to be negligible so that the frequency with which the attention network seeks high salience can be adjusted as desired by developers for the different robots.

Although other methods of determining temporally co-occurring events were examined, this method proved to be the most efficient in the software used. Other methods performed the same functions but they generated longer processing times and were computationally inefficient. Below is an example of the time-matching algorithm.

$$\overline{TS} = \begin{bmatrix} 14:31:22 \\ 14:31:22 \\ \mathbf{14:31:26} \\ \mathbf{14:31:27} \\ \mathbf{14:31:27} \\ 14:31:32 \end{bmatrix} \qquad \overline{D1} = \begin{bmatrix} 0 \\ 4 \\ \mathbf{1} \\ \mathbf{0} \\ 5 \end{bmatrix} \qquad \overline{D2} = \begin{bmatrix} 4 \\ 4 \\ \mathbf{1} \\ 5 \end{bmatrix}$$

Figure 9: Temporal Binding example

$\overline{TS}$ is the array of timestamps in descending order. $\overline{D1}$ is the array of differences between successive timestamps. In this example, the time limit between co-occurring events is three seconds. Since three values in $\overline{D1}$ are less than this limit, a second difference array is calculated. $\overline{D2}$ is the difference between every other timestamp in $\overline{TS}$. Since only one value is less than the limit of three, that value is chosen as the winner. The location of this value (1) is the third component of $\overline{D2}$. This corresponds to the difference between the third, fourth and fifth components of$\overline{TS}$. Therefore, the events that correspond to these timestamps are bound together and sent out as co-occurring events. Winning times and differences are highlighted.

CHAPTER V

SYSTEM PERFORMANCE AND EVALUATION

This chapter describes experiments designed to test the methods described in Chapter 4, the performance of the methods within these experiments and a discussion of the overall system performance. First, sensor and SES specifications are presented. Information concerning the sensors' resolutions and SES tessellation is needed to specify many of the variables in the equations given in Chapter 4. Results from event binding and attention experiments on ISAC are presented next. Results from experiments on Robonaut conclude the chapter.

ISAC: Sensors, Sensory Processing Modules and Sensory Ego-Sphere

On the ISAC platform, some sensors report to only one SPM while others report to multiple SPMs. Table 2 lists the sensors used throughout testing and their resolutions.

Table 2: ISAC – Sensors and their Resolutions

Sensor	**Resolution (degrees)**
Infrared	15
Microphones	15
Cameras	0.3
Hand Proximity/Tactile	5

Table 3 lists the SPMs used, the time latencies of their processing routines and the standard SES age limits L_e associated with each SPM.

Short-Term Memory Variables

Next, the time latencies and the time limits for the types of events posted onto the SES were determined. The time latency for an event is the time from detection of the stimulus at a sensor to the registration of the event onto the SES. This measurement is used to adjust the timestamps of events when they are posted onto the SES. If a SPM expends two seconds processing data, the event from the SPM will not be registered onto the SES at the same time other co-occurring events are registered. Therefore, the time latency from that SPM is subtracted from the time of registration for the event. The time limit for an event, L_e, is the time that it remains on the SES. The time latencies are specific to SPMs (i.e. face detection, IR motion detection, and color segmentation) while the time limits are specific to data types (i.e. person, object, motion). Table 2 in Chapter 4 lists these variables.

Table 3: ISAC - Sensory Processing Modules

Sensory processing module	**Time latency (sec)**	**SES Age limit, L_e (sec)**
Infrared motion detection	0.2	30
Sound localization	0.2	30
Face detection/recognition	0.8	60
Visual motion detection	0.5	30
Color segmentation	0.5	60
End effector localization	0.5	60

The time limits for some of the SPMs may seem too large for the type of information they output. For example, motion can be very quick and fleeting. By the first pass of the decay manager, the motion may indeed be gone. If the motion remains on the SES for a small interval of time (e.g. 1 second), the event contributes very little salience to itself or its neighboring nodes. This will occur because after the decay manager makes one pass to decrease data ages and salience values, the salience value of the event will be very low. In this case, motion would not be a focus of attention for the robot.

Since the cameras' resolutions are very fine yet the IR and sound sensors' resolutions are very coarse, a tessellation of $N = 14$ was selected for ISAC's SES. This value gives a 4.092° to 5.92° difference between vertices on the SES. With this tessellation, the resolutions of the IR and sound sensors are each about three edges of the SES.

Table 4 lists the sources used in experiments to test attention and event binding on ISAC. The table also lists the events that can be detected by ISAC from the sources. Not all sources' events are detected all of the time. However, all source events can be reliably detected under controlled conditions. Presentation of sources refers to when the source appears to the robot as opposed to detection of sources. Therefore, simultaneous presentation of multiple sources does not mean simultaneous detection of those sources. It should be noted that any object can produce end-effector sensory events either when placed in ISAC's hand or when ISAC reaches for the object.

Table 4: ISAC - Sources and their Sensory Events

Source	Possible sensor events
Rattle	Visual motion, color segmentation, IR, sound
Person	Face detection, IR, sound
Colored block	Color segmentation
Colored ball	Color segmentation
Talking Barney doll	Color segmentation, visual motion, IR, sound
Door to room	Visual motion, IR, sound

ISAC: Event binding

Event binding was tested to determine if the methods described in Chapter 4 could select events that actually originated from the same source. Experiments were divided into two phases: spatial binding and temporal binding. Event binding experiments were done first because spatial binding is dependent on the incidence measure and its variables. The spatial binding experiments determine an appropriate neighborhood size for spreading incidence for each sensory modality. The robotic platform for the experiments in this section was ISAC.

Spatial Binding

Spatial binding using the incidence formula in Equation 2 was performed to determine if events emanating from a single source can be detected as co-occurring in the spatial domain. Spatial binding was tested first to determine the maximum number of edges along which incidence and task-relevance should be spread for each sensor. Two experiments were run: one to collect the number of trials in which all events originating from a single source were correctly bound and one to collect the number of trials in which events all originating from separate sources were falsely bound as co-occurring. The first set of trials consisted of presenting a single source (the rattle) to the robot in a controlled environment[3]. The source produced three sensory events detected by IR motion detection, sound localization and orange color segmentation. The registration locations of each event were collected in 57 trials. In every trial, all three events were detected by the robot. Only 8 of these 57 trials contained events from sensors that reported correctly within their resolutions. (The sensor errors occurred in the azimuth direction (f_j) .The elevation direction, q_j, was not considered because the IR and sound sensors do not report output in that direction.) In 19 other trials, two sensors reported correctly within their resolutions while a third sensor reported within twice its resolution (this sensor was either the IR or sound sensor). For use in evaluations of the spatial binding, the angle of the detected event reported incorrectly was altered by 15° to mimic what would have been a correct result. This group consisted of 8 trials. Therefore, 16 trials were used to test if sensory events that co-occur were bound correctly in the spatial domain (Group A).

The second set of trials consisted of presenting three separate sources (the rattle, three separate presentations at different locations) to the robot in a controlled environment. The difference between sources ranged from 15° to 30°. This ensures that some events from trials in this group are detected as co-occurring, since 15° is the resolution for both IR and sound sensors. The rattle was used in the second set of trials so that the types of events produced in both experiments were the same. Each source produced a sensory event detected by IR motion detection, sound localization or orange color segmentation. The registration locations of each

[3] In a controlled environment, no spurious events can be detected. The only activity that can be detected by the robot are events from test sources.

event were collected in 16 trials (Group B). In every trial, all three events were detected by the robot.

In trial Group A, the visual event was taken as the actual location of the object, due to the fine resolutions of the cameras. The events from this group were then simulated in the spatial binding system so that different values could be used in the incidence equation (Equation 4.2). The events were registered onto the SES and incidence was spread from the events' registration nodes. For this and all other experiments, the incidence factor used was set to $a_I = 200$. The maximum number of edges along which incidence was spread was varied from $\max(E_{j,k}) = 1$ to $\max(E_{j,k}) = 3$. When all three events were registered onto the SES, the node with the highest salience was selected. All events that contributed to this salience were selected as co-occurring. For Group A, the trial was successful if all three events were selected as co-occurring. For Group B, the trial was successful if only one event was selected from the winning node (i.e. no events were selected as co-occurring). Table 5 shows the results from the spatial binding trials. The numbers represent the percentage of each groups' 16 trials in which the events listed were bound as co-occurring. Group A consists of trials in which the events actually co-occurred while Group B consists of trials in which the events did not co-occur. Therefore, for Group A, the percentages show correct hits and for Group B, they show false alarms.

Table 5: Spatial Binding Results

	Group A			**Group B**		
	IR, Vision	**Sound, Vision**	**IR, Sound Vision**	**IR, Vision**	**Sound, Vision**	**IR, Sound Vision**
$\max(E_{j,k}) = 1$	75%	62.5%	37.5	6.25%	12.5%	18.75%
$\max(E_{j,k}) = 2$	100%	93.75%	93.75%	31.25%	25%	50%
$\max(E_{j,k}) = 3$	100%	100%	100%	25%	12.5%	37.5%

A slight anomaly does exist in that the false alarm rates for Group B are lower when more incidence is spread. For trials in which the incidence was spread to $\max(E_{j,k}) = 2$**,** the false alarm rate is higher than those having $\max(E_{j,k}) = 3$**.** This may be due to the extra distance factor used in the incidence spread (Equation 4.1). If this is the case, it can be resolved by removing the factor of $E_{j,k}$ from the computation. This is a suggestion discussed in the Future Work section of Chapter 6.

A statistical analysis of the experimental results was performed to evaluate the successfulness of spatial binding. The probability that, given the locations of the detected events, the events all originated from the same source is used in the analysis. This analysis allows a ROC curve to be created using the probability that the co-occurring events actually co-occurred (true positives) and the probability that events that did not actually co-occur are found to be co-occurring (false positives). Since the visual events have the lowest resolution, their azimuth locations (f_v) are taken as the actual locations of the sources for Group A; therefore, the visual sensors' produced a 0% error. Means and standard deviations of sensor errors were taken from the entire group of data used in this experiment. The error of the IR sensors has mean $m_{IR} = 6.658°$ and standard deviation $s_{IR} = 3.996°$ while the error of the sound sensors has mean $m_{IR} = 7.894°$ and standard deviation $s_{IR} = 4.63°$. Because the sensors always have some error, the probabilities used in this analysis are the probability that, given the locations of the detected events for Group A, the IR and sound events occurred within a distance e of the visual event (Equation 5.2).

Let $\mathrm{p}(f_{IR_{\mathrm{act}}} | f_{IR_{\mathrm{det}}}, f_{V_{\mathrm{det}}})$ be the probability that given the detected locations of the IR and visual events, $f_{IR_{\mathrm{act}}}$ is the actual location of the IR event. Since the location of the detected visual event is assumed to be the location of the actual event at all times, $\mathrm{p}(f_{IR_{\mathrm{act}}} | f_{IR_{\mathrm{det}}}, f_{V_{\mathrm{det}}})$ becomes $\mathrm{p}(f_{IR_{\mathrm{act}}} | f_{IR_{\mathrm{det}}})$ and is defined in Equation 4.1.

$$\mathrm{p}(f_{IR_{\mathrm{act}}} | f_{IR_{\mathrm{det}}}) = \frac{\mathrm{p}(f_{IR_{\mathrm{det}}} | f_{IR_{\mathrm{act}}})\mathrm{p}(f_{IR_{\mathrm{act}}})}{\int \mathrm{p}(f_{IR_{\mathrm{det}}} | f_{IR_{\mathrm{act}}})\mathrm{p}(f_{IR_{\mathrm{act}}}) d f_{IR_{\mathrm{act}}}} \tag{5.1}$$

With uninformed priors, Equation 4.1 becomes $\mathrm{p}(f_{IR_{\mathrm{det}}} | f_{IR_{\mathrm{act}}})$ which is the probability density function (PDF) of the IR sensor error. Since the mean and standard deviation were calculated for the IR sensor, the PDF is known and normal. The probability that the IR event actually occurred within a range e of the visual event can be defined a $\mathrm{P}(-e \leq f_{IR_{\mathrm{act}}} - f_{V_{\mathrm{act}}} \leq e | f_{IR_{\mathrm{det}}}, f_{V_{\mathrm{det}}})$. Since the IR sensor PDF is assumed to be normal, the mean and standard deviation for this difference probability are $m_{diff} = m_{IR} - m_V$ and $s_{diff} = s_{IR} + s_V$. Since the visual sensor error is an impulse, the mean and standard deviation for the probability become that for the IR sensor PDF. If the same math is applied to $\mathrm{p}(f_{S_{\mathrm{act}}} | f_{S_{\mathrm{det}}}, f_{V_{\mathrm{det}}})$ for the sound event, then the probability that both the detected IR event and the detected sound event co-occurred with the visual event can be found.

This becomes $\mathrm{P}(-e \le f_{IR_{\mathrm{act}}} - f_{V_{\mathrm{act}}}, f_{S_{\mathrm{act}}} - f_{V_{\mathrm{act}}} \le e \mid f_{IR_{\mathrm{det}}}, f_{S_{\mathrm{det}}}, f_{V_{\mathrm{det}}})$ and, assuming that all sensors are independent, can be found from Equation 5.2.

$$\begin{gathered}\mathrm{P}(-e \le f_{IR_{\mathrm{act}}} - f_{V_{\mathrm{act}}}, f_{S_{\mathrm{act}}} - f_{V_{\mathrm{act}}} \le e \mid f_{IR_{\mathrm{det}}}, f_{S_{\mathrm{det}}}, f_{V_{\mathrm{det}}}) = \\ \mathrm{P}(-e \le f_{IR_{\mathrm{act}}} - f_{V_{\mathrm{act}}} \le e \mid f_{IR_{\mathrm{det}}}, f_{V_{\mathrm{det}}})\mathrm{P}(-e \le f_{S_{\mathrm{act}}} - f_{V_{\mathrm{act}}} \le e \mid f_{S_{\mathrm{det}}}, f_{V_{\mathrm{det}}})\end{gathered} \tag{5.2}$$

Using the values of these probabilities for both Group A and for Group B, a ROC curve can be created to compare the actual results from both groups of data. In these curves, the probabilities from Equation 5.2 were taken using a range of $0 \le e \le 62$ (after 62, all probability values equaled one). Figure 10 shows the ROC curve for the co-occurrence of the visual and IR sensor events. The x-axis is the probability that the IR and visual events from Group B were falsely bound as co-occurring while the y-axis is the probability that the IR and visual events from Group A were correctly bound as co-occurring. On this graph, the actual results from the spatial binding are shown as red blocks. The line represents the ROC curve while the blocks show the results from spatial binding using a maximum edge for spreading incidence of $\max(E_{j,k}) = 1$, $\max(E_{j,k}) = 2$ and $\max(E_{j,k}) = 3$.

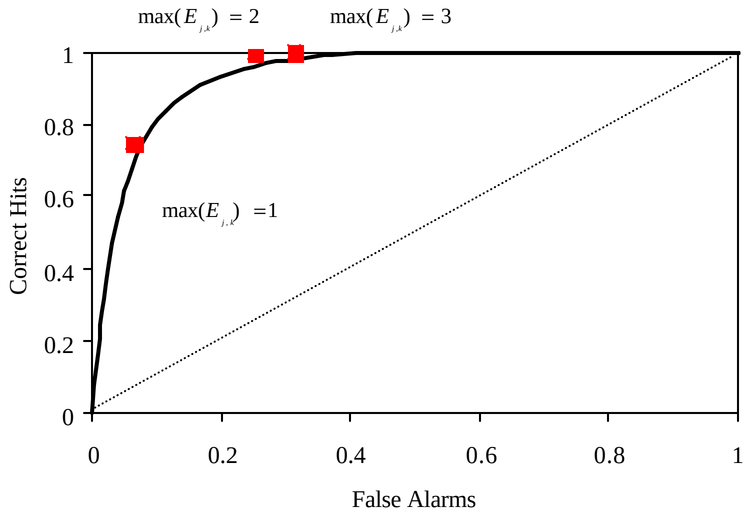


Figure 10: ROC Curve and Spatial Binding results for Visual and IR Events

For the spatial binding results, the points on the graph each represent the number of correctly bound events from Group A versus the number of incorrectly bound events in Group B. The spatial binding results fall almost exactly on the ROC curve, implicating that the spatial binding results perform as well as the probabilistic method. The slight deviations of the blocks from the ROC curve is do to the small number of samples used in the spatial binding and the numerical round-off that the statistics incur.

Figure 11 shows the ROC curve for the co-occurrence of the visual and sound sensor events while Figure 12 shows the ROC curve for the co-occurrence of the visual, IR and sound sensor events.

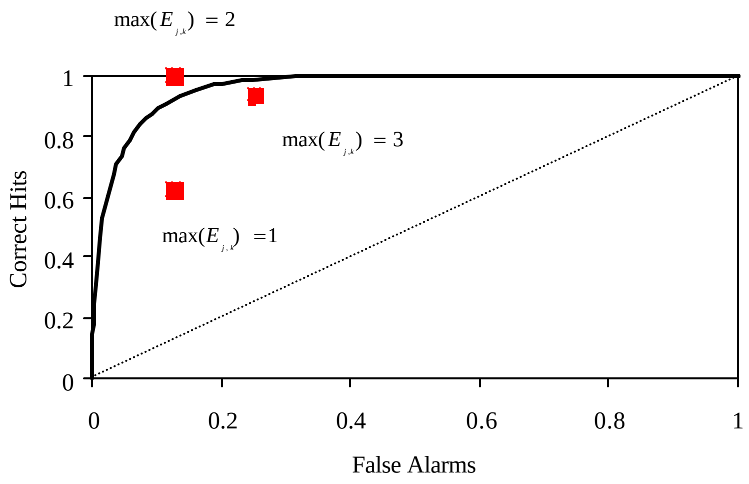


Figure 11: ROC Curve and Spatial Binding results for Visual and Sound Events

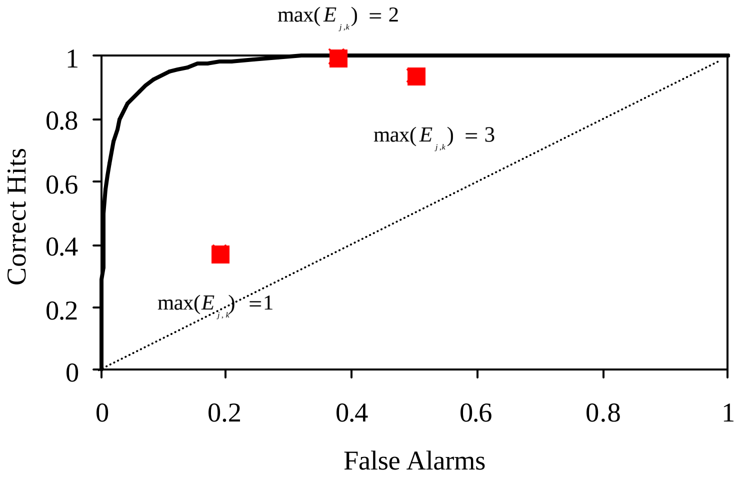


Figure 12: ROC Curve and Spatial Binding results for Visual, IR and Sound Events

For both Figure 11 and Figure 12, one point ($\max(E_{j,k}) = 1$) falls well below the ROC curve on both plots. (Since the co-occurrence of all three events is based on the co-occurrence of the visual and IR events and the co-occurrence of the visual and sound events, Figure 12 is affected by all perturbations in Figure 11.) The anomaly in Figure 11 may be explained by the poor performance of the sound sensors. The false alarms caused by the sound sensors and by the IR sensors in Group B were on separate trials, so that the combination of the two false alarms creates an even larger deviation on Figure 12. Once again, the slight deviations of the other two results from spatial binding $\max(E_{j,k}) = 2$ and $\max(E_{j,k}) = 3$) are do to the small number of samples used in the spatial binding and the numerical round-off that the statistics incur.

With the exception of the points discussed above due to the sound sensor, each of these graphs shows that the spatial binding mechanism produces results that are quite similar to those found through the probabilistic method. To use the probabilistic method for spatial binding in real time, the PDFs of the sensors must be known and updated and the area under these curves must be computed. While this may not be more computationally expensive than the spatial binding method put forth in this dissertation, the spatial binding affords more benefits in the long term: the attention network uses the same incidence measure for its computations. Therefore, spatial binding is selected as the method for determining spatial coincidence of sensory events. Since the resolutions of two of the three often used sensors in this work are 15°, a maximum neighborhood of $\max(E_{j,k}) = 2$ is selected as the default for the system. This allows a spread of about 10° on either side of an event, resulting in better performance of spatial binding. Also, the results show that statistically, co-occurring events are bound 93.75% of the time using this neighborhood value.

Temporal Binding

Next, temporal binding was tested to determine if events emanating from a single source can be detected as co-occurring in the spatial and temporal domains. This experiment was also performed to test if the event binding process would determine as co-occurring events that emanated from a different source (false positives). The experiment consisted of presenting two sources (Person A and Person B) to the robot in a controlled environment. Each source produced three sensory events detected by IR motion detection, sound localization and face recognition. The registration locations of each event and the time between presentations of the sources were collected during 12 trials. In every trial, all six events were detected by the robot.

In the first four trials, the spatial distance between the two sources was 5° while the temporal difference between presentations of the sources varied from 1 to 4 seconds. In the next four trials, the spatial distance between the two sources was 10° while the temporal difference between presentations of the sources varied from 1 to 4 seconds. In the last four trials, the spatial distance between the two sources was 15° while the temporal difference between presentations of the sources varied from 1 to 4 seconds.

The events collected from these three groups of data were then run through the entire event binding system. The events were registered onto the SES and incidence was spread from the events' registration nodes. The maximum number of edges along which incidence was spread was set at $\max(E_{j,k}) = 2$. When all three events were registered onto the SES, the node with the highest salience was selected. All events that contributed to this salience were selected and passed onto the temporal binding process. The temporal binding process then determined if any events were co-occurring. The time limit between co-occurring sources (T_B) was varied along the trials. Person A was always detected after Person B so that Person A always had the highest salience. (This occurs because newer events have their salience values decreased less than older events.) If all three events from Person A were selected as co-occurring without selecting events from Person B, then the trial was denoted as successful. If one event from Person B was selected as co-occurring along with the events from Person A, the trial was denoted as having falsely bound events. Table 6 shows the results from the 12 trials in which the time limit T_B was less than the difference between source detection times.

Table 6: Temporal Binding Results (T_B < temporal difference between source detections)

Distance between sources	**Successful trials**	**Trials with falsely bound events**
5°	50%	50%
10°	75%	25%
15°	100%	0%

These results show that the closer that sources occur in space, the more likely it is that events originating from separate sources will be bound as co-occurring. For trials in which the time limit between co-occurring events was equal to or greater than the temporal difference between source detections, events from separate sources were always bound as co-occurring. These results indicate poor performance of the event binding mechanism when the spatial and temporal limits are pushed. However, the visual SPMs of ISAC cannot actually detect two separate sources at identical times due to the control architecture of the robot. The pan-tilt units controlling the direction of the cameras can only be directed by one SPM at a time. Although two visual SPMs may detect events simultaneously, the time to shift control of the pan-tilt units from one SPM to another is ~2 seconds. On ISAC, for a visual event to be detected, the SPM must have control of the head so that it can center the target. Otherwise, a visual SPM cannot detect an event. This sets a time limit for temporal binding at two seconds for visual events. The IR and sound localization SPMs can report two separate events nearly simultaneously (see Table 2). However, the sensors reporting to these SPMs each have resolutions of 15°. This low resolution does not allow the binding mechanism to differentiate between two separate events less than 15° apart.

These results do indicate, though, that the method for binding co-occurring events should include different measures, for instance a contextual evaluation or a salience threshold. A contextual evaluation would compare the spatially bound events to determine if those events could originate from the same source. For example, if the spatially coincident events are IR motion and a recognized stationary table, the two events could not have originated from the same source. For the salience threshold, a minimum salience value that an event could contribute to coincidence would be established. Any spatially coincident events whose contributing salience values fall below the threshold would not be bound. These issues are further considered in the Future Work section of Chapter 5.

ISAC: Attention

Attention experiments were performed to determine if the network could locate the most salient area in the robot's environment under many different conditions. The most salient location was determined by the developer and the success of the attention network was contingent on whether a pre-determined location was selected as the most salient. Five separate experiments were run to test the network and to determine standard variable values. The first experiment established the task-relevance factor value at which task-relevant events overtake incidental events. The second experiment evaluated the effect different habituation factor values have on the total salience of habitual events. The last three experiments compare the performance of the attention network to three other methods of finding salient areas. In all experiments, the presentations of sources to the robot were controlled.

Task relevance versus Incidence

In the experiment to establish a task-relevance factor (a_{TR}) value, 5 trials were executed in which two sources were presented to the robot at the same time. Each source produced from one to three detectable events. The spatial distance between sources was always at least 15° while the temporal differences between presentations varied. When assigning both incidence and task-relevance to events, the maximum neighborhood edge distance was $\max(E_{j,k}) = 2$. These trials were performed to find the values at which areas containing task-relevant events become the most salient areas. For each trial, the task-relevance factor began at $a_{TR} = 1$. Task-relevance factors were applied per event.

Table 7: Task-Relevance vs. Incidence Results

Task-relevant event(s)	**Other event(s)**	a_{TR} **(per event)**
IR	Visual	1.4
Visual	IR, Sound	1.4
Visual (color)	Visual (face), IR, Sound	2.2
IR, Sound	Visual, Hand	1.3
IR, Sound (right)	Visual, IR, Sound (left)	1.4

Table 7 shows the results from the 5 trials. The task-relevant events came from one source and the other events came from the second source. In each trial, several episodes were executed. The task-relevance factor was altered for each episode by increasing or decreasing the value by 0.1 until a defined boundary was established. This boundary is the value at which the task-relevant event(s) overtook the other event(s) and is listed in the last column of Table 7. As the number of incidental events increased relative to the number of task-relevant events, the task-relevance factor increased. From these results, it is determined that for any task of minor importance, the task-relevance factor should be set to $a_{TR} = 2$ so that areas in which multiple non-task-related events occur around the same time could still be the most salient areas. For any task of greater importance, the task-relevance factor should be set to $a_{TR} = 3$. For any task that always takes precedence over any other events, the task-relevance factor should be set to $a_{TR} = 5$.

Habituation

Experiments were performed to determine appropriate habituation factor (b_H) values for use in the attention network. In these experiments, the habitual event was a hand waving at a constant rate in the same place.

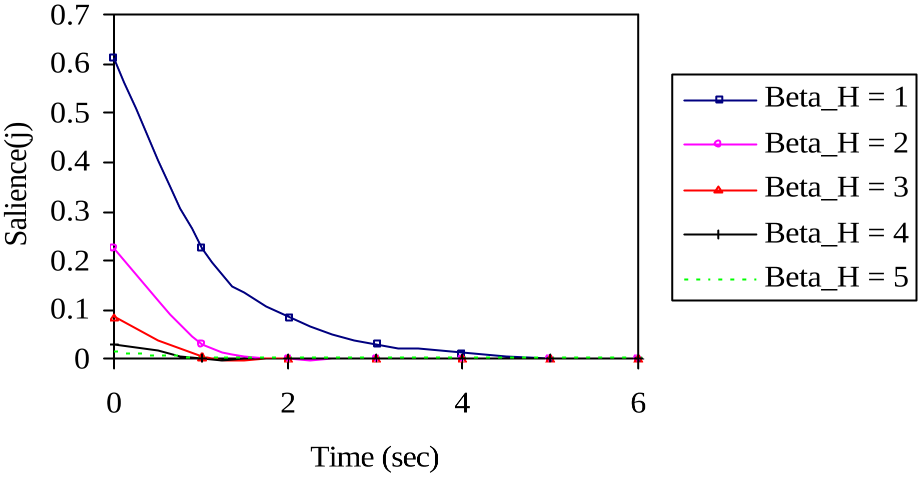


Figure 13: Decay of Salience due to Habituation

Figure 13 shows the decay of salience for a habitual IR event using different rates of decay. The graph shows that as the rate of decay increases, the salience of the habitual event decreases rapidly. For a habitual event to become the focus of attention at all, the rate of decay should be set to $b_H = 1$. Otherwise, the event may have completely habituated by the next pass of the attention network.

Focus of Attention

The next three experiments compared the attention network's success in selecting the most salient location against the successes of other methods of focusing attention. The first experiment compares the network in this dissertation against a simple search of the SES database. The second experiment compares this network’s performance against the multi-modal attention network developed by Sebastian Lang [Lang et al., 2003]. The third experiment compares the performance to the multi-modal attention network developed by Déniz et al. [Déniz et al., 2003]. In each experiment, the objective of both systems is to find the most salient location of the robot’s environment given the provided information (i.e. the current tasks). In the last two experiments, experimental conditions are set up to match those described by the compared methods’ authors.

Attention Network versus Database Search

In this experiment, 8 different sources were presented to the robot at separate times. Table 8 shows the sources used and the sensory events each source produced.

Table 8: FOA vs. Database Search: Sensory Sources Used

Source	**Events produced**
Green block	Green color recognition
Red block	Red color recognition
Blue block	Blue color recognition
Face A	Face detection
Person B	Face detection, IR motion detection, sound localization
Rattle	Visual motion detection, IR motion detection, sound localization
Door closing	IR motion detection, sound localization
Grasp	Hand proximity sensor high, arm velocity stopped, hand closed

The objective of this experiment is to determine if the attention network can correctly select the most salient location. The results of this experiment are then compared to a database search for salience to determine if the network performs better, the same or worse than the database search. For this experiment, the sources were presented successively at intervals of four seconds. The most salient area should be where the most recent event occurred, where event(s) related to the current task occurred or where a high number of incidental tasks occurred. For the attention network, the maximum number of edges used was $\max(E_{j,k}) = 2$ for all SPMs.

Figure 14 shows the shifts of attention (black line) for both the attention network in this work (**A**) and for the database search (**B**). In **A**, the method used to decay incidence of events was the linear decrease shown in Equation 3. The results from using the exponential decay from Equation 4 showed no difference from those in Figure 2. The linear decrease method was selected as the method used to decrease incidence because it required slightly less computational processing.

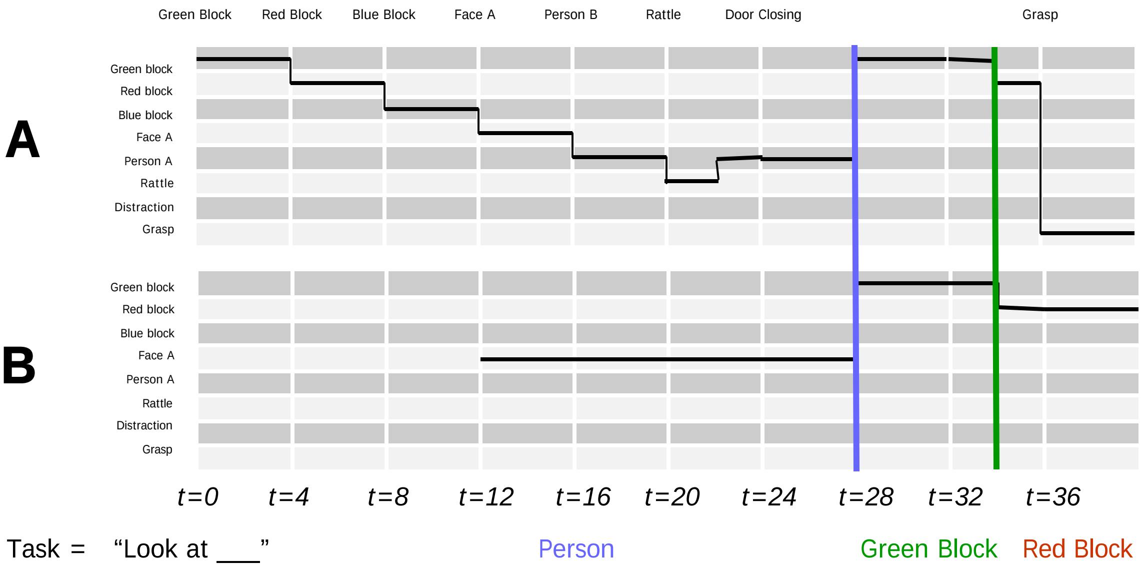


Figure 14: FOA using Attention Network (A) vs. FOA using Database Search (B)

The current task begins as “Look at person” and shifts to “Look at green block” at time $t = 28$ seconds. The current task then shifts to “Look at red block” at time $t = 34$ seconds. The sources to the left of the table list the areas that are the focus of attention while the sources above the table designate the source being presented to the robot. The task-relevance factor is set to $a_{TR} = 3$ for the ‘Look at person’ task and to $a_{TR} = 2$ for the ‘Look at green block’ and ‘Look at red block’ tasks. The different values of task relevance were assigned to mimic situations when the robot has multiple tasks with one task having more priority than the others. In this experiment, looking at the person was given the highest priority over looking at the blocks. For event binding, the time limit for detecting co-occurring events was $T_B = 3$ seconds.

In Figure 14, the shifts of attention for **A** follow the most recent source presented to the robot until events are detected that match a person (i.e. face detection and IR motion detection). This is a desired result because until person features are detected, there is no area matching the current task. (As a reminder, Table 4 shows the events that can be produced by a person.) Also, the sources up to this point produce only one event each; therefore, no area exists that has gathered multiple events. The next shift of attention occurs when a person is detected. In addition to face detection and IR motion detection, this source also produced sound. Since this source matches the current task and produced the most events, it became the focus of attention. The next shift of attention was caused by the rattle, which produced two events that are features of a person – IR motion detection and sound localization. The incidence values of the three events produced by the rattle overtook the salience of the person until the next pass of the attention network. The person then became the focus of attention, again. The next source was a door closing. Although this source produced two events that are features of a person, the incidence values generated by these two events were not enough to override the salience of the task-related events. The next shift of attention occurred with the change in current task from ‘Look at person’ to ‘Look at green block’. The focus of attention maintained at the green block, which was the object of the current task, until the current task was again altered to ‘Look at red block’. The red block then became the focus of attention until a grasp of the robot’s hand occurred. The incidence values from the three events produced by the grasp were enough to surmount the salience of the task-related event. This is a desired result for the system, however, because the grasp should require further action by the robot whereas the red block was not an immediate concern. In this situation, the robot is not sensitive to the grasp through task-relevance but, as per the robot’s developers, the robot should be sensitive to the grasp over other events. Lastly, all events from each focus of attention were bound together as co-occurring.

The database search could only use the information from the current task to determine what the focus of attention should have been. In the search, the current task with the highest value determines what is searched for in the SES database. When an event that matches any part of this task appears, it is selected as the focus of attention. This focus is sustained until either the event is purged from the database by the decay manager or until a different task becomes the highest priority. During the first current task (“Look at person”), the database search yielded no shift of attention until Face A appeared. Since features of a person were found, the search was completed even though Person B appeared. When the task changed the next two times, the database search found the correct focus of attention given the information that the search had (i.e. only task information).

The results from the attention network are compared to the database search results to show whether or not the task-relevance of attention is needed. The assumption is that if the robot knows what the current task is, then it can simply search its SES database to find the task-relevant events. When compared against the search method, the attention network performs no better or worse. Both methods found their foci of attention in negligible time; however, the database search did slightly less computational processing. Basically, the task-relevance measure in attention has no advantage over a database query for task-relevant events. However, the attention network method is preferred for the obvious reason: the attention network can find non-task-related foci of attention.

Parameters of the attention network were modified to fit the experiment in this section. The time period for detecting co-occurring events T_B was set to three seconds because it was known that the sources were being presented to the robot in constant intervals. Since the goal of this experiment was to evaluate the attention network's ability to find the most salient area, the event binding parameters were not a concern. However, in future use on the robot, the time period will have to be set at one value for constant use of the event binder. Solutions to this problem are discussed in the Future Work section of Chapter 5.

Also, the values of the task-relevance factors a_{TR} differed as a function of the current task. This was done because the robot can have multiple goals at one time with some goals being more important than others. This importance was reflected in the task-relevance factor; therefore, this modification was kept as a part of the system.

During this experiment, it was discovered that there was no need to spread incidence from one-dimensional sensors along a longitudinal axis. All motion and sound occurred around the 0° elevation of the SES; therefore, spreading incidence from these events along two edges from the 0° elevation point was sufficient. In the remainder of the experiments in this dissertation, all one-dimensional sensors are treated as two-dimensional sensors. That is, the

SPMs processing data from one dimensional sensors report events using two angles rather than one.

Attention Network versus Lang's Multi-Modal Attention System

In this experiment, conditions were set up to roughly match an experiment used to test Lang's multi-modal attention system for a mobile robot [Lang et al., 2003]. The objective of this experiment was to compare the success of the attention network in locating the most salient area with the results of Lang's attention system. Lang used his attention system to shift the focus of attention between different people. In this experiment, four people stood around the robot. Person 1 was at 45°, person 2 was at 0°, person 3 was at -30° and person 4 was at -60°. For that specific robot, 0° was straight ahead, as it is for ISAC. The experiment was set up for ISAC similarly. However, the data was collected from one person at a time and later simulated through the network as if the four people were present simultaneously. This was done because ISAC does not have the ability to continuously track multiple people but it was desired to have the experiment match the conditions of Lang's experiment.

Each person was detected at the given angular locations and at an elevation angle of $q = 0^\circ$. Each person spoke for 10 seconds individually so that the most salient location should be where the person currently speaking is. The maximum number of edges to which salience was spread was $\max(E_{j,k}) = 1$. Figure 15 shows the shifts of attention (black line) for both the attention network in this work (**A**) and Lang's attention system (**B**). The results of Lang's system are reproduced from [Lang et al., 2003].

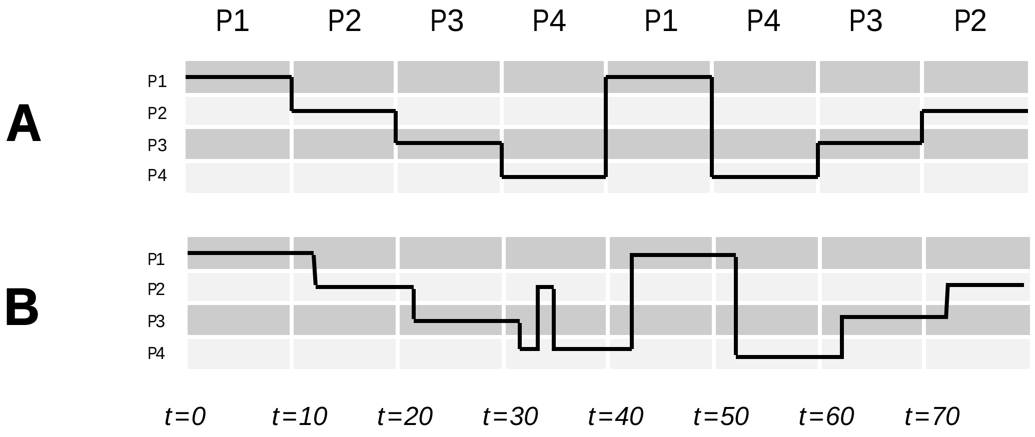


Figure 15: FOA using Attention Network (A) vs. Lang’s Attention System (B)

Lang’s system sought to anchor each person as a separate entity and to focus its attention on the person currently speaking. The figure shows that for all but one of the speaking intervals Lang’s system focused attention on the appropriate person and maintained that focus. The first time P4 spoke, the robot attended to P4, but then lost its anchor on P4 and attended to P2, who was the person standing directly in front of the robot. P4 eventually became the focus before he/she finished speaking. The authors state that the undesired shift of attention was due to the distraction of the person standing in front of the robot.

The SES-based attention network did not encounter any problems in detecting the person who was speaking as the most salient area. In each trial, the desired location was detected as being salient and the face detection event and the sound localization event were bound as co-occurring events. The SES-based attention system did not fail like Lang’s system because multiple sensory events always create a higher salience than that of a single event. Therefore, the detection of a person cannot override the detection of a person and detection of sound, unless the specific person relates to the current task. This experiment shows that using only incidence measures from detected sensory events and no task information, the desired focus of attention can still be found. The time limit between co-occurring events was set to a large amount of time (i.e. 180 minutes) because the faces were detected much earlier than the sounds were detected. This has proven to be a problem: the event binding mechanism cannot bind two events as co-occurring that were detected far apart temporally, even if the two events originated from the same source. The Future Work section of Chapter 5 discusses how this problem may be solved.

Attention network versus the Multi-Modal Attention System of Déniz et al.

In this experiment, conditions were set up to roughly match an experiment used to test the multi-modal attention system developed by Déniz et al. [Déniz et al., 2003]. The objective of this experiment was to compare the success of this dissertation's attention network in locating the most salient area to the performance of the system developed by Déniz. The experimental setup consisted of two objects, a person and a coat-rack. For ISAC, a person and a random object were used. Each object was detected visually. The objective of the Déniz experiment was to determine if sound events could cue the robot to look at the closest visual event. Therefore, the desired salient area for the attention network was the location of the sound and visual events. Figure 16 shows the shift of attention for both systems.

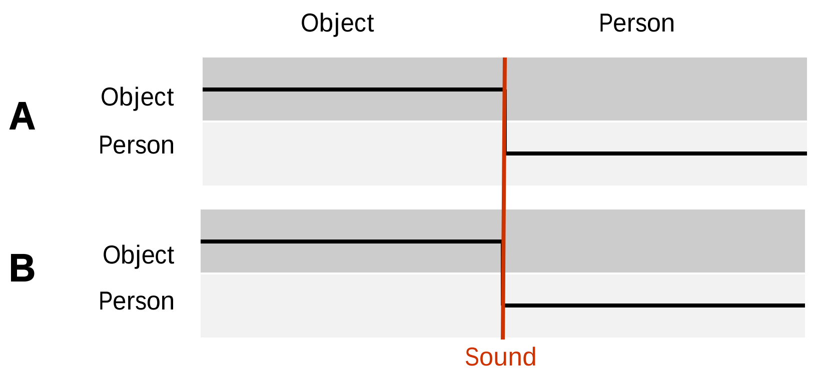


Figure 16: FOA using Attention Network (A) vs. the Attention System of Déniz et al. (B)

When the sound event is detected, both systems shifted attention to the person, which was the visual event closest to the sound event. This experiment shows that the attention system can perform as well as the Déniz system in cuing events from one modality using events from a different modality.

Robonaut: Attention and Event Binding

The attention network and event binding on Robonaut function like those developed on ISAC. SPMs exist to detect events and pre-filters serve to determine whether those events are task-relevant. Data was collected over several trials of reaching for and grasping a drill or wrench. One SPM detected objects via object recognition. Another reported the detection of an expected force on the arm via torque signals. A third SPM reported detection of the opening of the robot's hand. The SES for Robonaut also uses a tessellation of $N = 14$.

Experiments were run on two data sets. The first data set contained visual, arm and hand data taken directly from the robot during 6 different teleoperation trials. In these trials, the robot was teleoperated to grasp a wrench. The data was played back through RoboImitate which simulates the data stream output of Robonaut [Campbell, 2003]. RoboImitate is a software simulation that allows data to be collected during experiments and then played back later as if the data was originating from Robonaut. The first data set was played back through RoboImitate while the described SPMs processed the data stream to find sensory events. When shifts of attention were recorded along with the events bound as co-occurring. In each trial, Robonaut was teleoperated to perform five grasps. During these trials, Robonaut's right arm was teleoperated to reach for a wrench suspended in front of the robot. When the arm reached the wrench, Robonaut was teleoperated to grasp the wrench. Once the grasp was completed, the robot was teleoperated to open the hand and to move the arm back to a starting position. In all trials, the wrench did not move and the object tracker continuously reported the location of the wrench. Because of this, no habituation was used in these experiments.

Since the visual event was continuously reported, the FOA always occurred at the visual event which was the desired location. When the expected force SPM detected an event, the FOA again was at the desired location (location of the visual and force events) and the visual event and the force event were bound as co-occurring. When the hand SPM detected an event, the FOA was at the desired location (location of the visual event and the hand event) and the visual and hand events were bound as co-occurring.

The second data set contained data similar to the first set with an extra visual target. Before the data stream was played through RoboImitate, a drill was inserted into the stream to mimic Robonaut finding multiple visual objects. The location of the drill was more than 15° away from the wrench and did not occlude any sensory stimuli in the robot's workspace. This data set was used to determine the boundary of the task-relevance factor, as in the task-relevance experiments performed on ISAC. When the object tracking SPM detected the second visual event, the FOA shifted to this event when $a_{TR} \geq 1.4$.

The results from these experiments show promise for directing Robonaut to locations in its environment that need resources, whether for skill acquisition or for further processing. More experiments will be performed in the future to determine if the attention network and event binding mechanism are robust enough to operate on a humanoid robot having articulated motion with respect to a fixed frame.

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

Conclusions

This chapter provides a conclusion to the work presented in this dissertation and ideas on future work involving attention and event binding. The problems which this dissertation sought to solve were presented as:

Which events detected by the SPMs belong together by virtue of having been produced by the same source or occurring in response to an action of the robot?

How can the robot ignore irrelevant or spurious stimuli without missing those that indicate danger or opportunity?

These questions were answered by combining egocentric mapping and short-term memory to facilitate attention and event binding. The significance of this solution is that a unified mechanism, the Sensory Ego-Sphere, was used to as the structure upon which these four processes could function. The two robotic platforms on which the methods developed were tested were described. Other methods used for ECM and STM were presented and then the Sensory Ego-Sphere was detailed.

The attention network showed that it can reliably detect the most salient location of the robot's environment, whether the salience is produced by multiple unexpected events or by task-relevant events. When compared to other methods of finding an attentional focus, the network performed as well as or better than the other methods. The attention network as applied on the

SES has shown to be a good function for allowing a robot to allocate its limited physical resources to the most important area in its location.

The event binding mechanism showed that it can perform similarly to probabilistic methods developed. The results and evaluation show that spatial binding is accurate relative to the sampling of sensors used in this dissertation. Since probabilistic methods require large amounts of computation compared to the spread of incidence, the spatial binding is kept as a method of detecting spatially coincident events. The temporal aspect of event binding is sensitive to the time period within which co-occurring events can occur. This time period is a pre-defined interval that was modified for different experiments. The results of the temporal binding experiments demonstrate that the temporal aspect of the event binding may not be appropriate. Not only does the method rely on specifically pre-defined values, but it also does not allow for events that occurred far apart temporally to be detected as co-occurring (e.g. a face is detected at $t = 0$ seconds but the person begins talking at $t = 24$ seconds; these two events would not be detected as co-occurring even though they originate from the same source). Since it was assumed that events that originate from the same source are likely to be detected at the same time, this situation was not considered during testing. However, it appears that the temporal binding should be discarded in favor of another method that evaluates salience values or contextual elements of co-occurring events.

Overall, the system solved the problems put forth in this dissertation sufficiently well for the current research direction of both ISAC and Robonaut. However, the temporal binding algorithm should be altered or replaced entirely with another means of determining temporally coincident events. The spatial binding and the attention network performed well on both platforms, demonstrating that it can detect the most salient location of the robot’s environment (i.e. it can disregard task-irrelevant events without ignoring those that indicate an opportunity or danger). In summary, the SES has provided a solid platform for detecting spatially coincident sensory events and for directing a robot's attention based on salient areas in the environment. This salience was generated from random events in the robots' environments, from task-relevant events and from habitual events. The algorithms used for spatial binding and for attention will be permanently applied to the robots with possible further modifications that are mentioned in the next section.

Future Work

Several suggestions can be made for adaptations to the attention network and event binding mechanism so as to provide more accurate performance. Some of these suggestions are the result of poor system performance while others are inspired by afterthoughts on how to incorporate more functionality into the system.

First, the radial basis functions used to spread both incidence and task-relevance should be updated. The maximum number of edges to which salience is spread should be left out of the equation since the distance calculation handles assigning relative amounts of salience to neighboring nodes. Also, the distance measurement should change from Euclidean distance to spherical distance since the distance between vertices differs on different parts of the sphere. On ISAC, the Euclidean measurement does not affect the results; all activity on the SES occurs between about 10° and -40° and the salience is not spread beyond twice the maximum distance between vertices (~12°). However, the measurement should be altered for Robonaut since activity occurs all over the sphere. The new incidence equation would look like that shown in

Equation 1 while the distance between nodes would be a chord rather than a straight line in Euclidean space.

$$I(j,e,t_0) = \exp(-\frac{1}{a_I} D_{j,k}{}^2) \tag{6.1}$$

Next, the method of determining salience per node for a given event should be altered. Currently, if a task-relevant event is habitual, it will eventually lose all salience. Therefore, it is suggested that the habituation value only decreases the incidence at a node rather than both incidence and task-relevance. Equation 2 presents the new calculation.

$$S(j,e_n) = (I(j,e_n) * H(j,e_n)) + TR(j,e_n) \tag{6.2}$$

Also, habitual events may be important for the robot to attend to – if the robot is constantly running into a wall and producing multiple habitual events, then the robot needs to direct its resources to the area of the events. Therefore, it is suggested that the salience of co-occurring habitual events is not decreased but increased. The habituation value would simply be inverted in this case, as shown in Equation 2.

$$S(j,e_n) = (I(j,e_n) / H(j,e_n)) + TR(j,e_n) \tag{6.3}$$

Equation 5.2 should be used in the case of a single habitual event. If multiple habitual events have been found to be co-occurring by the event binding mechanism, then Equation 5.3 should be used instead. Another suggestion involving habitual events is to have the ending of a habitual event be a detected event itself. This would allow the cessation of a constant noise or constant motion to be detected as an event.

The most necessary adaptation to the system is to threshold the amount of salience that an event can contribute to the FOA and be considered co-occurring, as discussed in the Conclusions above. If an non-coincident event occurs close enough to other coincident events, it may still spread incidence to the node receiving the focus of attention. The salience contributed by this event may be very low, though (e.g. salience of 0.05) compared to the other coincident events (e.g. salience of 0.75 or greater). Further testing would need to be done to determine the appropriate value at which to threshold salience. In this situation, any event whose contributed salience falls below the threshold is not considered in event binding. Another suggestion is to add a context evaluator to the event binding mechanism. This evaluator would determine if co-occurring events could actually be co-occurring. For instance, if a green block and a face are detected in the same place around the same time, they could currently be bound as co-occurring. However, it is highly unlikely that a green block and a face would originate from the same source. A context evaluator could examine this and conclude that the events are not co-occurring.

Anther consideration is to assign SES age limits to events that are dependent on which events have been found to be co-occurring. A green ball detected with co-occurring motion would necessitate a small SES age limit (e.g. 30 seconds). A green ball detected without co-occurring movement would most likely be stationary, which suggests a larger SES age limit (e.g. 10 minutes). Also, when a current task is determined, events that are found to be task-relevant may be assigned a longer SES age limit than the standard limits. This would allow the robot more time in attending to task-relevant areas.

Finally, the data used to test Robonaut was calibrated due to errors from the visual object recognition SPM. To determine if event binding can overcome the calibration error from the visual stream, the attention network and event binding mechanism should be tested using the uncalibrated data.

APPENDIX

Idealized Geometric Structure of the Sensory Ego-Sphere

The idealized geometric structure was designed and developed by Peters [Peters et al., 2003]. Consider a binary set W defined on a 4-dimensional, Euclidean, space-time manifold $\mathrm{M} \equiv \mathrm{R}^3 \times \mathrm{R}$ and an associated indicator function,

$$\mathrm{W}:\mathrm{M} \rightarrow \{0,1\} \tag{7.1}$$

such that for any space-time point $\mathrm{p} \in \mathrm{M}$,

$$\mathrm{W}\{\mathrm{p}\} = \begin{cases} 1 \text{ if } \mathrm{p} \in \mathrm{W} \\ 0 \text{ if } \mathrm{p} \notin \mathrm{W} \end{cases} \tag{7.2}$$

W comprises the "world" – a set of geometrical objects within a void over time.

Let $\mathrm{M}_t = (\mathbb{R}^3, t)$, 3-space at time *t*, and consider W_t to be the "state of the world" at time *t*.

$\mathrm{W}_t\{\cdot\}$ is the object indicator restricted to time *t*.

At any given time, *t*, we designate one point in 3-space as the "ego-center", p_t.

$$\mathrm{p}_t = \begin{bmatrix} x_t & y_t & z_t \end{bmatrix}^T \in \mathrm{M}_t \tag{7.3}$$

The ego-center lies on a continuous curve

$$\tilde{\mathrm{p}} = \{ \tilde{\mathrm{p}}_x \mid -\infty < x < \infty \} \tag{7.4}$$

in space-time, M, where

$$\tilde{\mathrm{p}}_x = \begin{bmatrix} x_x & y_x & z_x & ? \end{bmatrix}^T . \tag{7.5}$$

For all *t*, there is an identity between p_t and $\tilde{\mathrm{p}}_x$:

$$\begin{bmatrix} \mathrm{p}_t^T \mid t \end{bmatrix}^T = \tilde{\mathrm{p}}_t . \tag{7.6}$$

That is, the egocenter moves through space over time so that at any given time, *t*, the egocenter lies at point $\mathrm{p}_t \in \mathrm{M}_t$. Over all time, the egocenter *is* the curve $\mathrm{p}\% \in \mathrm{M}_t$.

Let $F_t{}^p$ represent a 3-D rectangular coordinate frame that follows p*t*, the trajectory of the egocenter. Let $\overline{\mathrm{B}}_e\{\mathrm{p}_t\}$ be a depleted ball (a spherical shell) of radius $e > 0$ centered at p*t*; it likewise follows the trajectory (e is arbitrary, only f and q have meaning in SES context).

Let $1_t^p(r;q,f)$ represent the ray originating at p*t* having polar and azimuthal angles (q,f) with respect to $F_t{}^p$. The distance along the ray from its origin is represented by *r*. Finally, let $d_t{}^p(q,f)$ represent the particular distance, *r*, along the ray from the egocenter to the first object point in M_t. That is $d_t{}^p(q,f) = r$ such that $\mathrm{W}_t\left\{1_t^p(r;q,f)\right\} = 0$ for all $r < d_t{}^p(q,f)$ and $\mathrm{W}_t\left\{1_t^p(d_t{}^p(q,f),q,f)\right\} = 1$.

A Sensory Ego-Sphere, S_t^p is defined as the instantaneous projection of 3-space onto the spherical shell centered at p*t*,

$$S_t^p : \overline{\mathrm{B}}_e\{\mathrm{p}_t\} \to ¡ \tag{7.7}$$

such that

$$S_t^p(q,f) = d_t{}^p(q,f) \tag{7.8}$$

for $-\infty < t < \infty$, for $q \in [0,p]$, and for $f \in [0,2p]$. Thus, we define the SES mathematically as the set of radial distances from a designated point to the first encountered object points in space. Figure 17 shows the projection of an object onto the spherical shell, S_t^p.

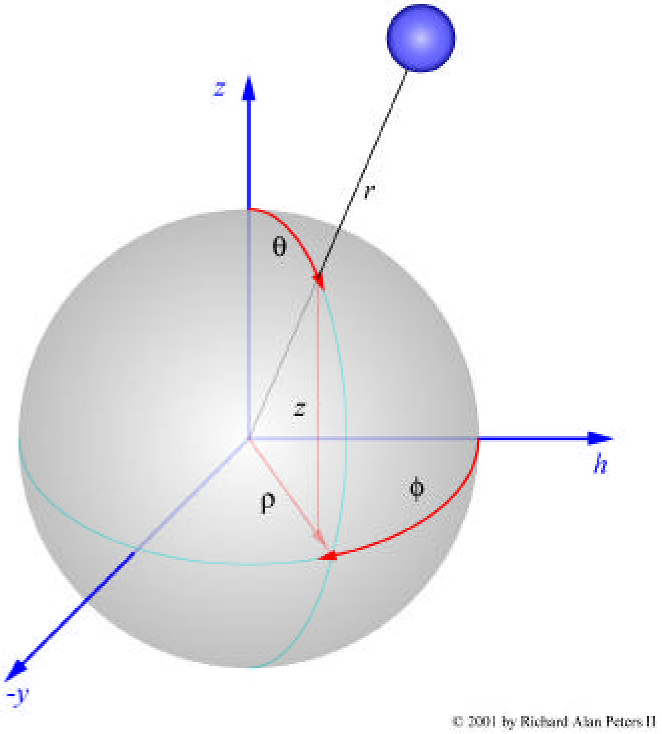


Figure 17. Projection of an Object onto the Sphere

Practically, we cannot define the SES with Euclidean density. Also, the geometric definition above implies that the structure is memoryless, whereas in fact it can be used as a memory structure. Moreover, we store on it much more than the distance to the first object in space.

SES for a Robot having Articulated Motion with respect to a Fixed-Base Frame

An articulated robot such as a humanoid has appendages and end-effectors that move with respect to its base frame. The proprioception of a dynamic body configuration is a spatially distributed sensory process that is a function of the robot kinematics. The physical contact of the robot's body with a surface can elicit a simultaneous response from its various sensors (e.g. force, torque, strain, tactile). The sensory events that result from any of these can be registered by projecting the instantaneous locations of the sensors and the joints on the SES. The projection is straightforward; the position, p_t^S, with respect to the base frame of a given sensor is written in spherical coordinates as

$$\mathrm{p}_t^S = \begin{bmatrix} r_t^S \\ q_t^S \\ f_t^S \end{bmatrix}. \tag{7.9}$$

Distance r_t^S is written at SES location (q^S, f^S) with a time stamp of t. Whereas sensory events that occur on the robot's body are easily projected to the SES through its kinematics, remote events detected by a directional sensor, such as camera platform are more problematic.

Consider a stationary object imaged at time $t = 0$ by a camera head whose frame, F_0^S, is rotated by A_S^B with respect to the base frame, F^B, and displaced from it by T_S^B.

If the displacement, r_0^S, of an object point from the camera frame is known, then the coordinates of that point with respect to the base frame are given by

$$\mathrm{r}_0^B = \Phi_S^B\left\{\mathrm{r}_0^S\right\} = A_S^B \mathrm{r}_0^S + T_S^B. \tag{7.10}$$

However, as is often the case, if the distance from the camera head to the object is unknown then all that is known is that the object lies on the ray

$$l^S(r^S) = \begin{bmatrix} r^S \\ q_0^S \\ f_0^S \end{bmatrix} \tag{7.11}$$

from the camera frame in direction (q_0^S, f_0^S). Scalar r^S is the distance along the ray from the origin of camera frame.

Let $\mathrm{r}^S = \begin{bmatrix} r^S & q_0^S & f_0^S \end{bmatrix}^T$ be any point on ray l^S written as a vector with respect to the camera frame, $F_0{}^S$. Vector r^B, the location of r^S with respect to the base frame, is given by (3.24). Let $\mathrm{l}^B(r^S)$ be the line segment from the origin of the base frame to the point on l^S a distance r^S from the origin of the camera frame. If we let r^S vary from 0 to ∞, then $\mathrm{l}^B(r^S)$ traces an arc of a great circle on the SES. The arc extends from the intersection of the SES with the ray through $\hat{\mathrm{T}}_{S,0}^B$ (the unit vector at F^B in the direction of $F_0{}^S$) to the intersection of the SES with the ray from the origin of F^B with direction (q_0^S, f_0^S) (i.e. the ray from F^B parallel to l^S).

To find the direction to the object from the base frame when the distance to the object is unknown, either a second camera must image it, or the first camera must be moved to a second position (that is not in the plane of r_0^S and $\hat{\mathrm{T}}_{S,0}^B$). The ray from the second camera in the direction of the object projects to an arc on a second great circle on the SES. The projection of the object on the SES is at the point of intersection of the two arcs. In fact, to compute the direction of the object with respect to the base frame, it is *not* necessary to compute the great circles and to find their point of intersection. A great circle is defined as the intersection of a spherical surface with a plane through its center. The arc traced by camera 0 is defined by the plane that contains unit vectors $\hat{\mathrm{r}}_0^S$ and $\hat{\mathrm{T}}_B^{S,0}$. Similarly, the arc traced by camera 1 is the intersection of the plane containing unit vectors $\hat{\mathrm{r}}_1^S$ and $\hat{\mathrm{T}}_B^{S,1}$. Ray $\mathrm{l}^B(r^0)$ from the origin of the base frame in the direction of the object is the intersection of these two planes. Now, the vector cross product

$$\hat{\mathrm{a}}_0 = \hat{\mathrm{r}}_0^S \times \hat{\mathrm{T}}_{\mathrm{B}}^{S,0} \tag{7.12}$$

is perpendicular to the first plane and

$$\hat{a}_1 = \hat{r}_1^S \times \hat{T}_B^{S,1} \tag{7.13}$$

is perpendicular to the second. This implies $1^B(r_0^S)$ is perpendicular to both $\hat{a}_0$ and $\hat{a}_1$. Therefore, $\hat{r}^B$, the unit vector at the base frame in the direction of the object, is given by

$$\hat{r}^B = \hat{a}_0 \times \hat{a}_1. \tag{7.14}$$

The articulated motion transformations were developed and designed by Peters [Peters et al., 2003]. Figure 18 illustrates the transformed projection of an object from Robonaut's camera coordinate frame to its base SES frame.

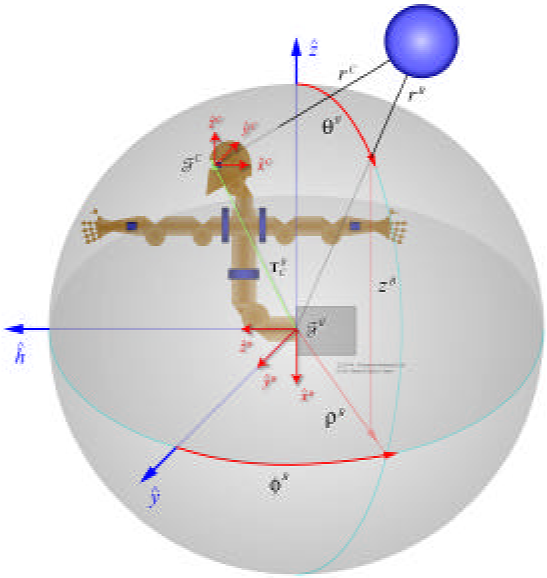


Figure 18. Transformation of an Object from Robonaut's Head Coordinate Frame to its SES

BIBLIOGRAPHY

Albus, James S. (1991) “Outline for a theory of intelligence”, *IEEE Transactions on Systems, Man and Cybernetics, 21(3)*, 473-509.

Albus, James S. and A. M. Meystel. (2001) *Engineering of Mind: An Introduction to the Science of Intelligent Systems*, Wiley and Sons.

Ambrose, Robert O., S. Askew, W. Bluethmann, and M. Diftler. (2001) “Humanoids Designed to do Work,” *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, pp. 173-80.

Arbib, Michael A., P. Érdi, and J. Szentágothai. (1998) *Neural Organization: Structure, Function, and Dynamics*, Cambridge: MIT Press.

Ayache, N. and O. D. Faugeras. (1988) “Building, registrating and fusing noisy visual maps,” *International Journal of Robotics Research*, 7(6), 45-65.

Baldi, Pierre and Y. Chauvin. (1994) “Smooth on-line learning algorithms for hidden Markov models,” *Neural Computation*, 6(2), 307-18.

Balkenius, Christian and N. Hulth. (1999) “Attention as Selection-for-Action: A Scheme for Active Perception,” In *Proceedings for Eurobot ’99*.

Balkenius, C. (2000). “Attention, habituation and conditioning: toward a computational model,”*Cognitive Science Quarterly*, 1, 2, 171-214.

Bluethmann, William, Ambrose, R., Diftler, M., Askew, S., Huber, E., Goza, M., Rehnmark, F., Lovchik, C., and Magruder, D. (2003) “Robonaut: a robot designed to work with humans in space,” *Autonomous Robots.* Mar-May 2003;14(2-3):179-97.

Borthwick, S. and H. F. Durrant-Whyte. (1994) “Dynamic localization of autonomous guided vehicles,” In *Proceedings of 1994 IEEE International Conference on Multi-Sensor Fusion*, 92-97.

Breazeal, Cynthia L. (2002) *Designing Social Robots*, Cambridge: MIT Press.

Breazeal, Cynthia. (1999) "A context-dependent attention system for a social robot," In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1146—1151.

Brill, F.Z., W.N. Martin, and T.J. Olson. (1995) “Markers Elucidated and Applied in Local 3-Space," In *Proceedings of the 1995 IEEE International Symposium on Computer Vision*, November 19-21, 1995.

Bushara, Khalafalla O., T. Hanakawa, I. Immisch, K. Toma, K. Kansaku, and M. Hallett. (2003) “Neural correlates of cross-modal binding,” *Nature: Neuroscience*, 6(2), 190-95.

Campbell, Christina. (2003) “RoboImitate”, Robonaut software, NASA, JSC.

Cañas, Jose M. and M. C. Garcia-Alegre. (1999) “Real time EM segmentation of occupancy grid for robots navigation,” In *Proceedings of IJCAI-99 Workshop Adaptive Spatial Representations of Dynamic Environments*, 75-79, Stockholm.

Cave, Kyle R. (1999). “The FeatureGate Model of Visual Selection,” *Psychological Research, 62*, 182-194.

Chrisman, Lonnie. (1992) “Reinforcement learning with perceptual aliasing: the perceptual distinctions approach,” In *Proceedings of the Tenth National Conference on AI (AAAI)*.

Christopher, J. L. Jr. (1999) “A PneuHand for Human-Like Grasping on a Humanoid Robot,” M.S. Thesis, Vanderbilt University, May 1999.

Coradeschi, S. and A. Saffiotti. “Perceptual anchoring of symbols for action,” In *Proc. Int. Conf. on Artificial Intelligence*, pages 407–412, Seattle, WA, 2001.

Cox, Ingemar J. (1991) “Blanche – an experiment in guidance and navigation of an autonomous robot vehicle,” *IEEE Transactions on Robotics and Automation*, 7, 193-204.

Déniz, Oscar, M. Castrillón, J. Lorenzo, M. Hernández and J. Méndez. “Multimodal Attention System for an Interactive Robot,”, *Lectures Notes in Computer Science*, vol. 2652, *1st Iberian Conference on Pattern Recognition and Image Analysis*, Pto. Andratx, Mallorca, Spain, 4-6 June 2003.

Diftler, M.A., Platt, R., Culbert, C.J., Ambrose, R.O., Buethmann, W.J. “Evolution of the NASA/DARPA Robonaut Control System,” *Proceedings of the 2003 IEEE Conference on Robotics and Automation (ICRA)*, Taipei, Taiwan, May 2003.

Drescher, Gary. (1991) *Made-Up Minds: A Constructionist Approach to Artificial Intelligence*, Cambridge: MIT Press.

Driscoll, Joseph A., R.A. Peters II, and K. R. Cave. (1998) “A Visual Attention Network for a Humanoid Robot,” In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems.*

Dudai, Yadin. (2002) *Memory from A to Z : keywords, concepts, and beyond,* New York : Oxford University Press.

Edmondson, A. C. (1986) *A Fuller Explanation: The Synergetic Geometry of R. Buckminster Fuller*, Birkhäuser, a Pro Scientia Viva title, Design Science collection, CIP.

Floreano, D. and F. Mondada. (1996). "Evolution of plastic neurocontrollers for situated agents". In *From animals to animats* 4, 401-10, Cambridge: MIT Press.

Flynn, A.M. (1988) "Combining sonar and infrared sensors for mobile robot navigation," *International Journal of Robotics Research*, 5-14.

Gonçalves, Luiz. "Towards a Learning Model for Feature Integration in Attention Control," In *Proc. of IEEE Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI 2001)*. August, 20-22, Baden-Baden, Germany.

Hong, L. and G.J. Wang. (1994) "Integrating multisensor noisy and fuzzy data," In *Proceedings of 1994 International Conference on Multi-Sensor Fusion and Integration for Intelligent Systems*, 199-206.

Iftekharuddin, Khan M., R. P. Malhotra. (2002) "Role of Multiresolution Attention in Automated Object Recognition," *Proceedings of the 2002 International Joint Conference on Neural Networks*, 3, 2255-60.

Itti, Laurent, C. Koch, and E. Niebur. (1998) "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 20(11), 1254-59.

Kam, Moshe, X. Zhu, and P. Kalata. (1997) "Sensor Fusion for Mobile Robot Navigation," In *Proceedings of the IEEE*, 85(1), 108-18.

Kawamura, K., T.E. Rogers and K.A. Hambuchen. (2002) "Towards a Human-Robot Symbiotic System," Invited Paper, *International Manufacturing Automation Leaders Forum (IMLF)*, Adelaide, Australia, February 8-10, 2002.

Kawamura, K., A. Alford, K. Hambuchen, and M. Wilkes. (2002) "Towards a Unified Framework for Human-Humanoid Interaction," *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*, September 2000.

Kayama, Kentaro, K. Nagashima, A. Konno, M. Inaba, and H. Inoue. (1998) "Panoramic-Environmental Description as Robot's Visual Short-Term Memory," In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation*, 3253-58.

Klatzky, R. L. (1998). Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In C. Freksa, C. Habel, & K. F. Wender (Eds.), *Spatial cognition – An interdisciplinary approach to representation and processing of spatial knowledge* (*Lecture Notes in Artificial Intelligence* 1404) (pp. 1-17). Berlin: Springer-Verlag.

Klute, G. K., J. M. Czerniecki, and B. Hannaford. (1999) "McKibben Artificial Muscles: Pneumatic Actuators with Biomechanical Intelligence", *Proc. IEEE/ASME 1999 Int'l Conf. on Adv. Intell. Mechatronics (AIM '99)*, Atlanta, GA, September 19-22, 1999

Konolige, K., K. L. Myers, and E. Ruspini. (1997) "The Saphira Architecture: A Design for Autonomy," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 9, *Special issue on Architectures for Physical Agents*, pp. 215-235.

Kraetzschmar, Gerhard, S. Sablatnög, S. Enderle, H. Utz, S. Simon, and G. Palm. (2000) "Integration of Multiple Representation and Navigation Concepts on Autonomous Mobile Robots" *Workshop SOAVE-2000*: Selbstorganisation von adaptivem Verhalten. Kurztitelaufnahme der Deutschen Bibliothek, Boston, Basel, Stuttgart.

Lang, Sebastian, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G.A. Fink and G. Sagerer. (2003) "Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot," In *Proceedings of the International Conference on Multimodal Interfaces*, November 5-7, 2003, Vancouver, British Columbia.

Littman, Michael J. (1993) "An optimization-based categorization of reinforcement learning environments," In *From Animals to Animats: Proceedings of the Second International Conference on Simulation of Adaptive Behavior*, Cambridge: MIT Press.

Liu, Yaning. (2000) "Human-Robot Audio Interaction," Masters Thesis, Vanderbilt University, December 2000.

Maravita, A., C. Spence, S. Kennett, and J. Driver. (2002) "Tool-use changes multimodal spatial interactions between vision and touch in normal humans," *Cognition* 83, B25-B34.

Masumoto, Daiki, H. Yamakawa, T. Kimoto, and S. Nagata. (1994) "Hierarchical Sensory-Motor Fusion Model with Neural Networks," In *Proceedings of the 1994 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 630-38, Las Vegas.

McCallum, R. Andrew. (1996) "Learning to Use Selective Attention and Short-Term Memory in Sequential Tasks," In *From Animals to Animats, Fourth International Conference on Simulation of Adaptive Behavior*, Massachusetts.

Meeden, Lisa A. (1996). "An incremental approach to developing intelligent neural network controllers for robots," *IEEE Transactions on Systems, Man, and Cybernetic*s, 26, 3, 474-85.

Navalpakkam, Vidhya and L. Itti. (2002) "A Goal Oriented Attention Guidance Model," In *Proc. 2nd Workshop on Biologically Motivated Computer Vision , Tuebingen, Germany,* 453-461.

Nolfi, Stefano and D. Floreano. (2000). *Evolutionary Robotics*. MIT Press, Cambridge.

Pack, R. Todd. (1997) "Intelligent Machine Architecture," Ph. D. Thesis, Vanderbilt University, May 1998.

Peters , Richard A. II, K. Hambuchen, and K. Kawamura. (2001) "The Sensory Ego-Sphere as a Short-Term Memory for Humanoids," *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, 451-60.

Peters, Richard A. II, K. Hambuchen, R. Bodenheimer. (2003) "The Sensory Ego-Sphere: A Mediating Interface between Sensors and Cognition," submitted to *IEEE Transactions on Systems, Man and Cybernetics*.

Qiu, Bei-fang. (1997) "Face and Facial Feature Detection in a Complex Scene", Masters Thesis, Vanderbilt University, August 1997.

Sablatnög, Stefan, S. Enderle, M. Dettinger, T. Boss, M. Livani, M. Dietz, J. Giebel, U. Meis, H. Folkerts, A. Neubeck, P. Schäffer, M. Ritter, H. Braxmeier, D. Maschke, G. Kraetzschmar, J. Kaiser, and G. Palm. (1999) "The Ulm Sparrows 99," RoboCup-99 Team Descriptions, Simulation League, Team Ulm-Sparrows, 144-48.

Sekmen, Ali. (2000) "Human-Robot Interaction Methodology," Ph.D. Dissertation, Vanderbilt University, August 2000.

Soyer, Ç., Bozma, H.I., Istefanopulos, Y. (2000) "A New Memory Model for Selective Perception Systems" *Proceedings of the 2000 EIII/RSJ International conference on Intelligent Robots and Systems.*

Srikaew, Atit. (2000) "A Biologically Inspired Active Vision Gaze Controller," Ph.D. Dissertation, Vanderbilt University, May 2000.

Stewart, I. (1991) "Circularly covering clatharin," *Nature*, 351(9), 103, 1991.

Tremblay, Marc R. and M. R. Cutkosky. (1995) "Using Sensor Fusion and Contextual Information to Perform event Detection during a phase-based manipulation task," presented at *1995 International Conference on Intelligent Robots and Systems.*

Urner, K. (1991) "The Invention Behind the Inventions: Synergetics in the 1990s," *Synergetica Journal*,1(1).

Vijayakumar, Sethu, Conradt, J., Shibata, T. and Schaal, S. "Overt Visual Attention for a Humanoid Robot," In *Proc. International Conference on Intelligence in Robotics and Autonomous Systems*, 2001, Hawaii.

Weisstein, Eric W. (1999) "Geodesic Dome." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/GeodesicDome.html.

Wolfe, J.M. (2001). "Guided Search 4.0: A guided search model that does not require memory for rejected distractors," [Abstract] *Journal of Vision*, 1(3), 349a.

Wolfe, Jeremy M. (1994) "Guided Search 2.0: A revised model of visual search" *Psychonomic Bulletin and Review,* 1(2), 202-38.

Ziemke, Tom and M. Thieme.(2002) “Neuromodulation of Reactive Sensorimotor Mappings as a Short-Term Memory Mechanism in Delayed Response Tasks,” *Adaptive Behavior*, 10(3/4), 2002.

Ziemke, Tom. (1999). “Remembering how to behave: Recurrent neural networks for adaptive robot behavior,” In *Recurrent Neural Networks: Design and Applications*, Medsker & Jain (eds.), 355-89, New York: CRC Press.