ELECTRICAL ENGINEERING

IMAGE MAPPING AND VISUAL ATTENTION ON A SENSORY EGO-SPHERE

KATHERINE ACHIM

Thesis under the direction of Professor Richard Alan Peters II

The research in this thesis focuses on two problems related to the Sensory Ego-Sphere (SES), a short-term memory structure for a robot: (1) the mapping to the SES of high-resolution sensory information in the form of imagery, and (2) the concurrent processing of visual attention. Neither problem had been studied previously. The SES coordinates sensory information for further processing and thereby acts as an interface between sensing and cognition. It is an egocentric, spherical mapping of the robot's locale. This research is based on previous work in the areas of multi-modal sensing, sensory-motor coordination, and attention. The paper describes a procedure to composite on the SES an image sequence taken by a camera head, a task that is complicated by significant overlap between successive images in the sequence. Two approaches to the problem of finding and ranking areas of visual interest are compared. One combines visual attention points in the overlapping images on the SES and the other computes attentional points directly on the composited visual scene. Computational structures for mapping imagery to the SES and managing it are described. The problem of attention in bio-vision is discussed as are some algorithms that mimic visual attention behaviors in humans.

IMAGE MAPPING AND VISUAL ATTENTION ON A SENSORY EGO-SPHERE

By

Katherine Achim

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Electrical Engineering

August, 2005

Nashville, Tennessee

Approved

Professor Richard Alan Peters II

Professor Robert E. Bodenheimer

À ma famille pour tout votre amour, support, et confiance en moi.

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor, Dr. Peters, for his insight, support, and guidance. This experience would have been much more difficult without his encouragement and faith in my abilities. I would also like to thank Dr. Bodenheimer for taking time to review my thesis and provide insight as to how to improve it. Thank you to Dr. Wilkes as well for helping me with concepts while my advisor was away. Thanks also to Flo Fottrell for all of her help. Without her, this process would have been much more stressful and disorganized! Thank you for all of the work you do for the CIS department.

I would like to thank the Electrical Engineering Department for their financial support during the completion of this thesis. I would also like to thank all the members of the Center for Intelligent Systems for their help, support, and camaraderie. This experience would not have been as enjoyable without you all.

Finally, I would like to thank my family and Paul for all their love and support. Without them, I would not be where I am today; I am grateful for all you have done for me. Thank you for everything.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

The Sensory Ego-Sphere (SES) is a biologically inspired, short term memory structure for robots that acts as an interface between sensing and cognition [30]. It can be envisioned as a virtual spherical shell surrounding the robot. Information about a point in space is stored on the shell in the direction of the point from the center of the sphere. Thus the SES is an egocentric, spherical mapping of the locale. To date, it has been used to recall the locations of discrete objects in the vicinity of a robot. As such the SES is a sparsely populated map. It is, in theory, capable of providing a dense map of the environment, wherein every point (within the granularity of the sensors) contains data from any number of sensors. Within such a dense mapping, discrete objects or areas of interest can be tracked if they are marked as such. This thesis reports on a test of those two conjectures. It reports on a mapping of high-resolution sensory information (in the form of visual imagery) onto an SES. It also addresses the problems of finding and ranking areas of interest in the images that form a complete visual scene on the SES.

The primary practical use for robotics today is in industrial automation, where it has been successful. In that setting, a robot's environment is well defined and unchanging. The environment is designed to make it easy for the robot to work and it is controlled to keep it that way; it is a machine-centered world. Robots have yet to enter the human-centered world (where the environment is dynamic and unpredictable) as useful tools, despite many

potential applications. Human-centered robotics has the potential to become a key technology of the 21st century.

Problem statement

The grounding problem, defined here as the coupling of action and perception in the real world or the matching of a robot's objectives and resources, is a central problem of cognitive science. It is not completely understood how animals ground themselves in the world, how they learn to interact effectively with their environment and how they understand the effects of their actions on the world (neuro-ethological problem) [22]. How does an abstract representation come to be associated with a physical object or phenomenon instead of with another abstract representation? This problem is related to the philosophical problem of meaning [14]: how can the meaning of arbitrary symbols be grounded in non-symbolic representations instead of other meaningless symbols? One solution states that symbols or abstract representations must be grounded in non-symbolic representations obtained from sensory information [14].

One of the major unsolved problems in robotics is precisely how to combine sensory information of different modalities so that signals are correctly attributed to objects in the environment. Sensory-Motor Coordination (SMC) is clearly necessary for animals and robots. It may also be fundamental for categorization. Pfeifer has shown that SMC data—recorded during simultaneous action and sensing by a robot that is executing a fixed set of tasks in a simple but changing environment—can self-organize into descriptors that categorize the robot-environment interaction [31]. Learning SMC could solve the grounding problem for robots.

There exist certain requirements for learning Sensory-Motor Coordination. As a robot operates, multimodal sensory information must be associated with motor activity. This requires sensor binding despite different spatio-temporal resolutions and differing temporal latencies in throughput. The SES, with independent, parallel sensory processing modules (SPM), does this by virtue of its structure [13]. Since resources (sensory, computational, motor) can only be directed toward a small subset of environmental features available at any one time, learning SMC also requires attention. The SES can combine attentional events detected by different sensors with task- and environment-specific context to produce a ranked set of critical areas in the environment.

The SES is an egocentric, spherical mapping of the environment. To date, it has been used to keep track of the position of known objects in the vicinity of a robot. It is able to combine attentional signals to direct the focus of attention. It is also capable of sensitization and habituation with respect to attention [13]. These three capabilities imply that the SES is currently limited to a sparse mapping of the environment. It has not been able to map high-resolution data such as visual imagery, which requires a dense mapping and, therefore, has not been involved in high-resolution attention.

The contributions of this thesis include methods for and analysis of the high-resolution mapping of visual imagery to the SES, including temporal updating. Methods for and analysis of visual attention on the SES are also explored.

Image Mapping and Visual Attention

Attention is a vital process in vision; it facilitates the identification of important areas in a visual scene. It has been described as a spotlight, illuminating a particular region while neglecting the rest. Corbetta has characterized this selection of a region as necessary because of "computational limitations in the brain's capacity to process information and to ensure that behavior is controlled by relevant information [5]." Moreover, research done on attention mechanisms in the brain has been useful in identifying areas of the visual system as well as their behavior and function. Such information has aided the development of visual attention models that are particularly useful in computer vision and robotics [2].

The Sensory Ego-Sphere replicates some of the functionalities of the mammalian hippocampus which is critically involved in short-term memory, integration of multiple sensory inputs, sensory-motor integration, egocentric mapping of the environment, and position (of self) with respect to objects in the environment [30]. The SES is an egocentric mapping of sensory data. It is structured as a geodesic dome that is indexed by azimuth and elevation angles, and is generally centered on a robot's base frame [30][1]. This thesis focuses on mapping high resolution data, in the form of imagery, to the SES. Imagery is projected onto the sphere at azimuth and elevation angles that correspond to the current pan and tilt angles of the camera-head, and their description, as well as other relevant information, is stored into a database. The vertex on the sphere closest to an image's optical axis angle becomes the registration node, or the location where information is posted (*i.e.*, stored). One of the problems associated with the registration of imagery on the SES is how to composite an image sequence taken by a camera-head rotating with respect to a robot's base frame over time. This problem is not trivial because successive images in the sequence overlap

[1] For this work, the SES was centered on the pan/tilt origin of the camera-head.

significantly. It was solved by an algorithm to extract foveal windows[2] from each image to be posted to the SES. The size of the windows depends on a pixels-per-degree measure (determined by examining overlapping adjacent images) as well as the focal length of the camera and angle between adjacent images. This will be described in detail in chapter IV.

A second problem is the detection and mapping of visual attention points. A visual scene contains much information, much more than can be processed at a high level by an active agent with limited computational abilities. Visual salience is a measure of conspicuousness or relevance; a location with a high salience value should be attended while a location with a low salience value can be disregarded. Salience values of locations in a scene must, therefore, be computed to determine which locations to attend. Thus, a visual attention system is employed to map salience levels to locations on the Sensory Ego-Sphere. Like image registration, this task is complicated by the overlap of images on the SES. Attentional points found in an individual image often will not fall within the foveal window of the SES node and, therefore fall on other nodes, to which they must be mapped. The visual attention system described herein finds and ranks attentional points, both in individual images from a time-sequence and on the composite scene registered on the SES. Two point selection approaches were tested: the first was to find the attentional points and their salience values in individual images, map them to the appropriate nodes, and then sum the salience values of each node to determine the foci of attention on the SES. The second approach was to discard the portions of individual images that were not in the foveal window. An image was then constructed from these foveae and processed by the attention system. This will be described in detail in chapter IV.

[2] The fovea is the high resolution central portion of the retina of the eye. The fovea is centered on the optical axis of the eye. A small image, from the central window that surrounds the optical axis of a camera has come to be known as the camera's fovea, even though that region is of no higher resolution than the rest of the image.

Thesis Organization

The remainder of this thesis is structured as follows: Chapter II contains background information on visual attention mechanisms in humans, models of visual attention, and previous work on the Sensory Ego-Sphere. Chapter III describes the hardware and software used for this research. Chapter IV contains the algorithms for image mapping and visual attention processing using a Sensory Ego-Sphere, and Chapter V contains the results of experiments performed. Chapter VI offers final thoughts and possible future work extending the application of research completed for this thesis.

CHAPTER II

BACKGROUND MATERIAL AND PREVIOUS WORK

Visual Attention in Humans

Attention is a vital process in vision; visual attention is the identification of salient (or relevant) areas in a visual scene. A salient feature is one that is made noticeable by differing from its surrounding neighborhood. The relevance of an area is determined by the task to be performed or by how closely its features match those of a known target. The mammalian visual system, if thought of as a computational process, is an intricate, non-localized network of modules with much interaction through feed-forward and feedback paths [37]. Much of the separation of the visual system from other parts of the brain is conceptual rather than anatomical, since other sensory modalities are processed at some of the same location as visual information. Its actual structure is not completely known. Its architecture appears hierarchical at times and parallel at others. Perhaps the best assumption is to qualify it as a distributed system where many modules interact with each other and affect one-another in both serial and parallel ways [6]. Mechanisms and processes of the visual system such as eye movement, motion detection, attention, or object recognition are neither self-contained nor attributed to only one module, but instead depend both on a variety of brain structures. Therefore, this discussion of attentional representations in the visual system does not attempt to identify structures that are entirely responsible for attention (*i.e.*, an attention module), but instead identifies several areas where evidence of attentional mechanisms has been observed. Attention affects (or exists in) areas which span from the primary visual cortex to higher

areas of visual processing [15]. Studying attentional representations in the mammalian brain gives insight on a very efficient visual system and this knowledge can help the development of an efficient robotic visual system.

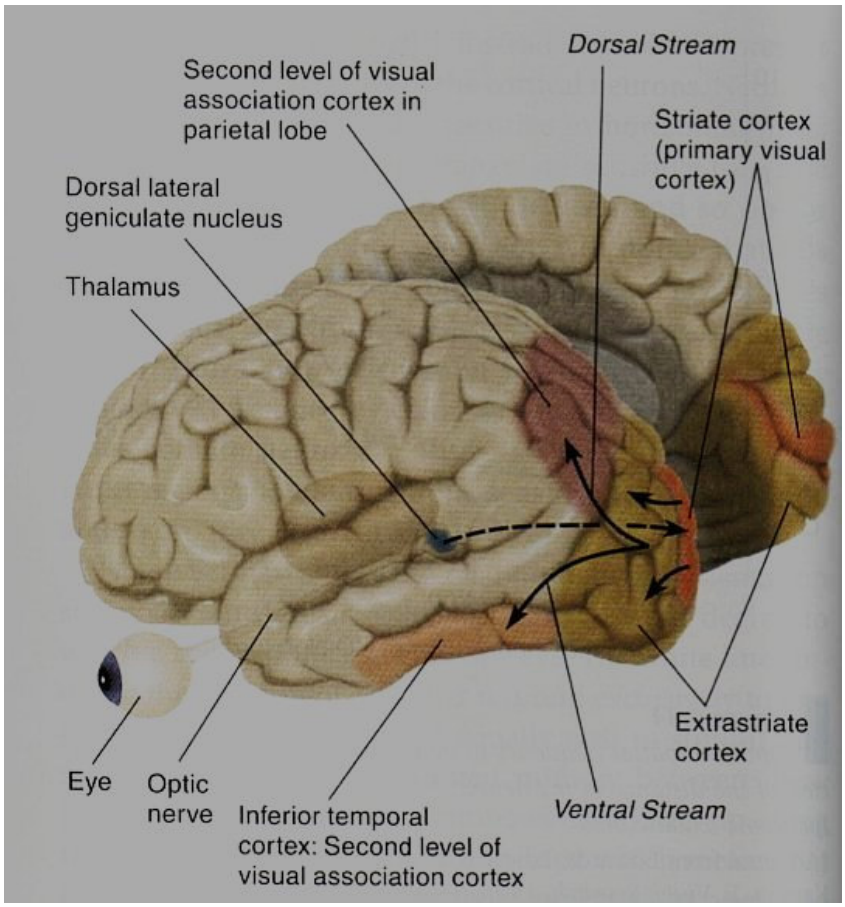


Figure 1. Vision pathways [44]

The visual system is often thought of as being divided into two separate streams: a ventral pathway for object recognition, the "what" pathway, and a dorsal pathway for motion detection, the "where" pathway. Both of these streams, shown in figure 1, have an effect on, and are affected by attention. Attention emerges from multiple processes in the visual system. Attentional activity can be detected by neuroimaging techniques as modulation of

cell responses in different visual areas [1]. These include the frontal cortex, occipital cortex, parietal cortex, medial thalamus, and superior colliculus, Attentional modulation can also be observed in the frontal eye field (FEF), cingulate, premotor, lateral prefrontal, orbitofrontal, opercular, posterior parietal, lateral and inferior temporal, parahippocampal, and insular and subcortial regions [3]. Some of these areas are discussed below since they have influenced some of the work described in this thesis.

Primary Visual Cortex (V1 or Striate Cortex)

The primary visual cortex (or V1) is the first area of the visual system modulated by attention (excluding eye movements). The optical mapping of the visual scene onto the retina maps in turn onto each eye's ganglion cells, which transfer information to the lateral geniculate nuclei (LGN). The LGN regulates the transfer of information to the primary visual cortex where the first transformations are made on the information coming from the retina. V1 separates and packages information from the retina and LGN and then sends it on to more specialized visual processing areas [37]. V1 contains several representations of the visual scene through spatial selective cells, orientation selective cells, direction selective cells, and disparity selective cells, among others. It sends information mainly to V2 to be distributed to both the ventral and dorsal streams (the "what" and "where" pathways) [37].

Attentional modulation has been observed in V1 with the help of neuroimaging techniques such as positron-emission tomography (PET) and especially functional magnetic resonance imaging (fMRI). PET studies showed that specific cell locations in V1 were modulated as a function of size, and that the degree of modulation increased with the number of items present in the visual field [32]. This finding is in agreement with a model of

attention in which objects in a crowded scene compete for representational resources, since all of the information available to the system cannot be processed simultaneously. The competition can be described as dependent on both bottom-up and top-down factors. Bottom-up factors are attributes of a stimulus, such as color or shape. If an attribute of a particular stimulus inherently differs from the attributes of the surrounding stimuli, this particular stimulus will have a better chance in the competition. Top-down factors are task-dependent and depend on a desired behavior. For example, if a subject knows that the target is a red circle, stimuli that look most like a red circles (red in color, round in shape) will have the best chances in the competition. Top-down factors can override bottom-up factors when the former are not relevant to the task at hand [37]. The object or feature in the visual scene that wins the competition is attended to [7]. Attentional processing on the SES is based on these ideas.

A 1999 fMRI study by Somers et al. that involved complex visual tasks found evidence of response modulation in V1 when attention was directed at either the fovea or the periphery of the eye [32]. Two other functional MRI studies by Gandhi et al. and by Martinez et al. used simpler tasks, but the attention necessary to complete them was still significant. V1 activation was definitely observed [32]. In summary, attentional modulation of the response of cells in V1 was found to depend on (1) the complexity of the task, (2) the competition from other items in the visual field, and (3) context integration (such as contours and curve tracing) [32]. The attentional processing algorithm used in this thesis replicates some of these findings. It has both bottom-up and top-down processing components wherein locations in a scene (or image) compete to become the focus of attention.

Extrastriate Areas

Extrastriate areas are areas outside of the striate cortex, which is another name for primary visual cortex, V1. Modulations have also been observed in color and motion sensitive extrastriate areas, V4 and V5 respectively, when a subject's attention is directed to such features. Chawla et al. [3] conducted a study and found that task demands can modulate the response of cells in these areas. They established that the response of a cell in extrastriate area V5 is enhanced when attention is directed at a moving stimulus in its receptive field[3]. This enhancement is not observed when a moving stimulus is viewed passively (motion is not attended to), but it is observed when a subject attends to motion even if no motion is present in the visual field. Similarly, cells in V4 exhibited enhanced responses when the color of objects was attended. Furthermore, cells become increasingly selective with increases in the complexity of the task, as more attention is required. This demonstrated that responses to separate attributes of objects can be attentionally modulated. This finding is used in many models of visual attention, including the one implemented here, where the visual scene is decomposed by attributes into feature maps.

Another study was performed by Brefczynski et al. [1] to identify the neural mechanism that enables covert shifts in attention, or the ability to direct attention to an area without making a saccade[4] to it. They used functional MRI techniques to observe the gross patterns of activation throughout the entire visual system while a task involving the shifting of attention (but not of gaze) was executed. The results of this study showed a spatially-mapped attentional modulation, homeomorphic to the mapping of the visual world onto the retina. Modulation was observed in primary visual cortex as well as in extrastriate areas such

[3] The receptive field of a cortical cell is the area on the retina which when stimulated causes a response in the cell.
[4] A saccade is a rapid shift in gaze, caused by a ballistic eye movement.

as V2, V3, VP, V4v, medial occipital cortex, and ventral occipitotemporal cortex. This study also showed that in addition to being retinotopic, attentional modulation can be object-based.

The parietal cortex, which is part of the dorsal pathway, is important for visual space analysis, movement, and attention. More specifically, a study by Gottlieb et al. [12] has shown that the lateral intraparietal area (LIP), which is part of the parietal cortex, contains a partial representation of the visual scene that only strongly represents salient or behaviorally significant stimulus. The response of a cell was significant when a stimulus was flashed in its receptive field (RF). In contrast, the response was much less when a saccade was made to the center of an array of stable (non-flashing) stimuli, which put the same stimulus element in the cell's RF. (The act of making a saccade to place a stimulus in the fovea is known as an overt shift in attention.) It can then be concluded that neurons in the LIP respond to recent-onset stimulus and not older stimuli at different locations in the field. The recent onset makes the stimulus salient. This finding can be applied to the combination of visual attention points on the SES; attentional points from new images can be given higher activation values.

Studies by Corbetta [5] identified a network of signals in the frontal and parietal cortex that directs covert attention to relevant locations in the visual field. It also modulates the responses in the ventral pathway, which deals with object recognition. Covert attention involves attending to a stimulus without directly fixating it[5]. This finding can be related to the presence and significance of multiple attentional points on the Sensory Ego-Sphere.

Attention is necessary for deterministic visual searches. A study by Klein et al. [21] of overt, orienting visual search observed a phenomenon which they called inhibition of return (IOR) in the superior colliculus. In one of their experiments, a complex scene was

[5] Fixation is the act of maintaining a gaze on a point in space, typically after a saccade to the point.

searched overtly for a hidden target. After a period of time, a probe[6] appeared somewhere in the visual scene and had to be detected by making a saccade to it, thereby placing the particular probe in the fovea. When the probe was placed at a location where there had previously been a saccade, there was inhibition of return; that is, the amount of time needed for its detection was lengthened. It is thought that inhibition of return facilitates complex visual searches by preventing the visual system from searching the same places repeatedly to acquire new information about a visual scene. Inhibition of return was not observed when the visual scene being searched was removed before the probe was presented, giving evidence that IOR is directly connected with objects in a scene. The concept of inhibition of return will be discussed further in the next section, since it can be applied to attentional processing on the SES.

Research performed by Desimone [7] explored a link between memory and attention. Although it is usually thought that attention plays an important role in memory and the learning of stimuli or the environment, it is also true that memory plays a reciprocal role in attention. The inferior temporal cortex (IT) of monkeys was studied. The IT, located at the end of the ventral stream, is thought to be partly responsible for complex visual processing and object recognition. The response of a class of cells in IT was found to be enhanced only when the stimulus matched a previously viewed sample stimulus (held in memory). It was also found that a representation of a behaviorally important stimulus at a particular time is created by maintaining activity during stimulus delay periods in IT. Cells in prefrontal cortex were found to have a similar activity pattern. It is thought that the feedback path from this cortex to IT modulates the activity in favor of the stimulus matching the sample held in memory. This agrees with the biased competition model of attention in favor of behaviorally

[6] In this context a probe is a unique, visually discernible marker.

relevant objects described earlier. Biasing of attention in this manner can be performed with the SES. Additionally, the study [7] found evidence for memory mechanisms influencing attention in experiments involving visual searches: response to stimulus was enhanced when the stimulus was a match to the sample, and decreased when it did not match.

Representations of Visual Attention Mechanisms in Artificial Vision

One way to artificially implement a naturally occurring process is to identify and understand the various stages of the process and then model them mathematically. A model of visual attention in the mammalian brain, or an attentional processor for a robot could be constructed from the psychophysical and physiological knowledge that has been acquired through the many studies mentioned above. Effective models of visual attention in artificial vision systems could be applied to such tasks as navigational aids, humanoid robotics, surveillance, and automatic target detection [17]. For a robot and human to be able to work together they must be able to direct each other's attention to relevant objects. This implies that the robot's visual attention system should be similar, at least in results, to that of the human. Several computational models of visual attention have been implemented by a number of researchers. They are discussed below.

Itti Algorithm

Itti et al. [16, 18] developed a saliency-based model of visual attention. Salience refers to how conspicuous a location or feature is in an image; salient locations are locations which "locally stand out from their surround" [18]. Their model is strictly bottom-up. It ranks discrete points in the image by salience and directs attention to the location of the

maximum. An image of a scene inputs into the model which computes several feature maps in which objects or locations in the scene then compete for saliency. Only the most prominent or noticeable[7] locations in a region are kept within each map. This is done on all of the maps in parallel and is, therefore, very fast. Once several salient locations have been identified, they can be sent on to other processes for further analysis. There are forty-two feature maps implemented in the Itti model: six maps for intensity contrasts (six different center/surround size combinations), 12 for color discrimination, and 24 for orientation discrimination. These features are extracted using center-surround operations which are computationally similar to the receptive fields of some cells (usually edge detectors) in the visual system. The architecture has proven to be effective in identifying salient locations in a scene [18]. The color feature maps are implemented using center-surround organizations similar to the red/green and blue/yellow double-opponency receptive fields in the visual system [37] There are 4 different color combinations (red/green, green/red, blue/yellow, yellow/blue) and six different center/surround size combinations. Maps account for red/green and green/red double opponency simultaneously (similarly blue/yellow, yellow/blue double opponency is computed simultaneously) for a total of 12 feature maps. A receptive field will respond to a given color strongly, but to its opponent color weakly, in its center and vice-versa in the surround. The orientation maps in the Itti algorithm have receptive fields represented by Gabor pyramids, a model first proposed by Marcelja in 1980 [24]. They model the sensitivity profiles of the orientation-selective cells[8] in the primary visual cortex [18]. There are four orientations selected, 0°, 45°, 90°, and 135°, and six different center/surround size combinations, for a total of 24 maps. All these feature maps are the

[7] by Itti's definition.
[8] those cells that respond to oriented edges.

inputs to a final saliency map, which ranks the importance of locations in the scene. Because all feature maps are combined, it is possible that features deemed salient in only a few maps may appear less salient than non-important objects or noise present in more feature maps. To mitigate this problem, the feature maps are normalized to a common fixed range before they are combined in the saliency map. Each map's global maximum is found and the average of the local maxima is computed. The map is multiplied by the square of the difference between the two. These computations enhance the difference between a maximum and the overall map average so that if that difference is big, the location will stand out more than if that difference is small. The feature maps are then combined into three categorical salience maps, one each for color, intensity, and orientation. These three are normalized then summed to form the final saliency map. The location of the maximum in this map indicates the location where the focus of attention should be directed. If this location is found by another process to be incorrect, then the location is inhibited for a period of time to give the next-most salient location the focus of attention. This inhibition delay is much larger than the time required for an attention shift from one location to another so that several locations will be attended before returning to the originally attended location. This mimics the inhibition of return behavior observed in the visual system. There is also a positive excitation applied to locations closest to the currently-attended location to bias the model to attend to these locations next. Figure 2 shows the architecture of the Itti model.

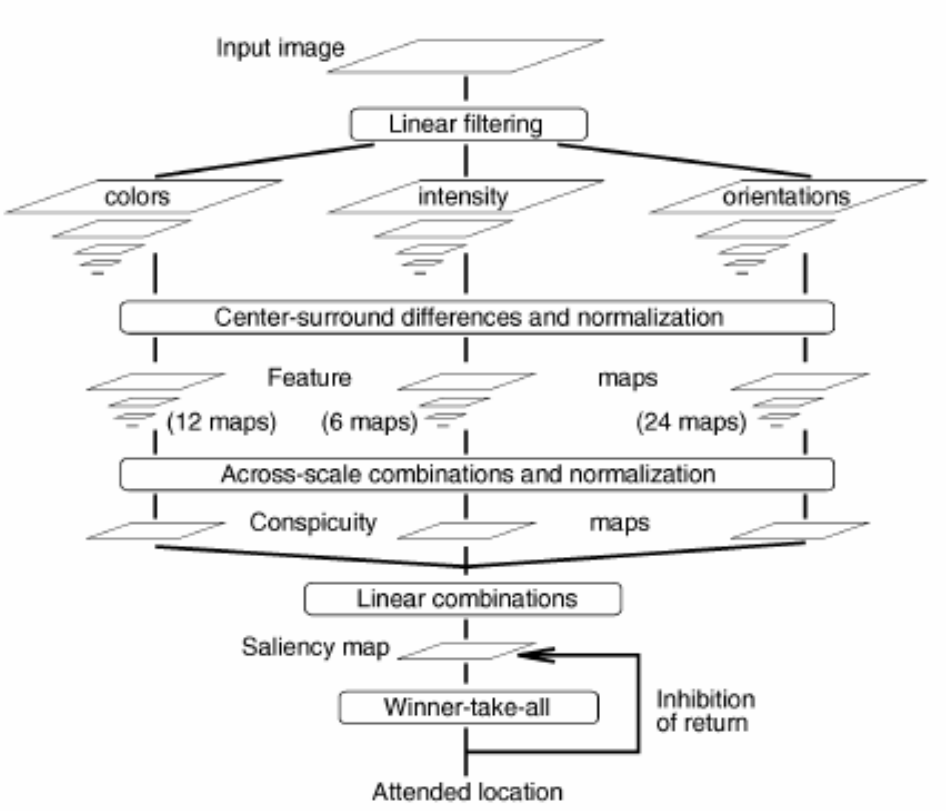


Figure 2. Architecture of Itti's saliency-based model of visual attention [18]

The Itti model was observed to find a target on the first try if this target differed from distractor locations[9] by one feature category (color, intensity, or orientation). However, if the target differed in more than one feature category (called a "conjunctive search"), the search took a longer time period which was proportional to the number of distractors present. These are the same results observed in humans.

Recently, the Itti model was modified to include a top-down component that incorporates task relevance in visual attention [27]. The model first determines what to look for by parsing a task specification using a knowledge base of entities and their relationships. The most relevant task-related entity is then searched for in the visual scene. Task relevance

[9] locations with one or more of the features of the target

is taken into consideration by biasing the attention system for those low-level features (intensity, color, or orientation) that are present in the target (task-relevant entity). Figure 3 illustrates the top-down biasing model for object detection in a visual scene. This is similar to the attentional biasing done on the SES, although learning internal target representations is not specifically implemented. Targets are specified to the SES either manually or by another computational object.

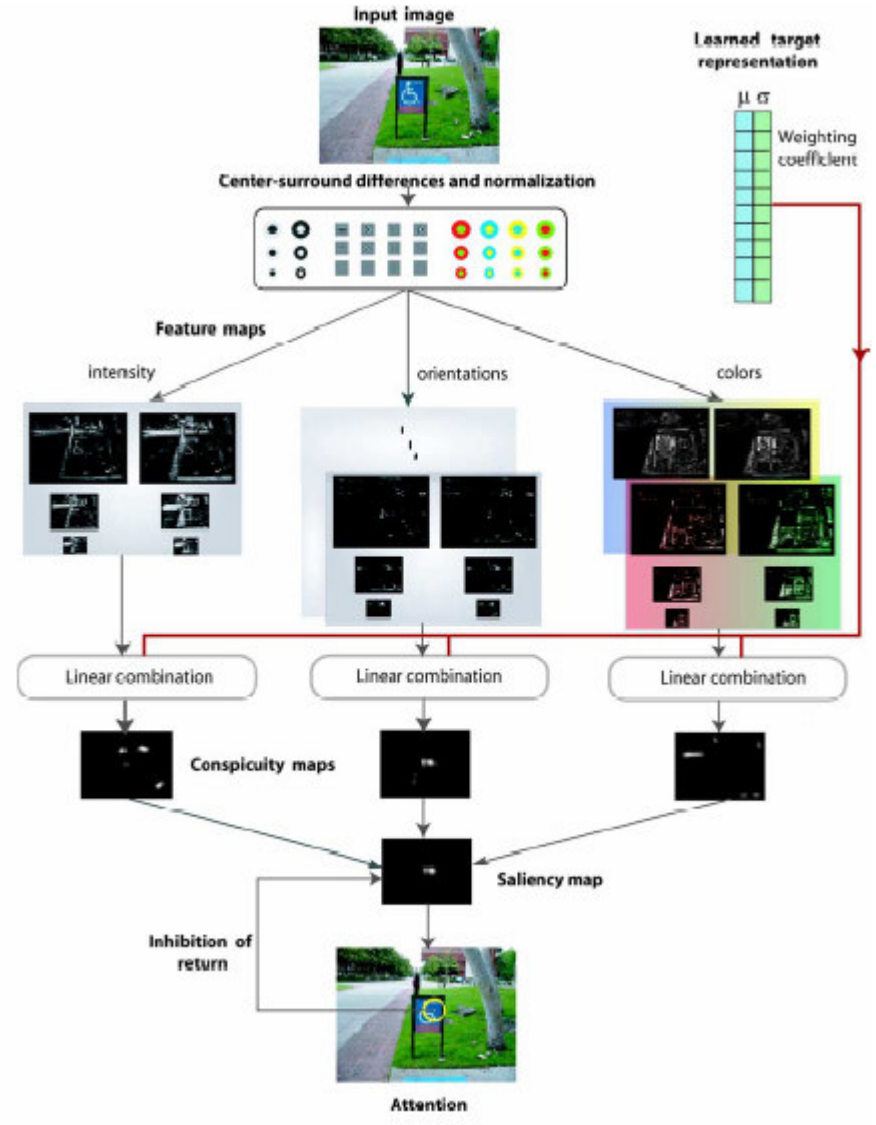


Figure 3. Top-down biasing model for object detection. To detect a specific target in any scene, the learned target representation is used to bias the linear combination of different feature maps to form the salience map. In the salience map thus formed, all scene locations whose features are similar to the target become more salient and are more likely to draw attention[27].

FeatureGate

Cave [2] observed that inhibition during attentional tasks was directed at objects in the visual scene but not at blank locations. He also observed that, during target searching with distractors present, the distractors closest to the target were inhibited more than the ones

farther away. Another study performed by Cave concluded that distractors with more target features were less inhibited than those without any target features. From these results, Cave concluded that attention is not quite a spotlight that inhibits everything equally outside its beam. Instead, attention seems to inhibit objects in the scene based on both their locations and their similarity to the target. Using this information, Cave designed the FeatureGate model of visual attention.

Each location in the visual scene has a vector of basic features such as orientation or color (as above). Each location also has an attentional gate that regulates the flow of information from there to the output. The gate depends on that location's features and the features of surrounding locations. The entire vector of feature values present at a location is passed to the output if the location is deemed most important for the task. FeatureGate is hierarchical; the visual scene is partitioned into neighborhoods. The "winning" location in each neighborhood is passed to the next level but the others are not. This proceeds iteratively until there is only one location remaining. The remaining location is the output of the model.

FeatureGate can be envisioned as in Figure 4. The network contains a number of feature modules and an activation module. There is one feature module for each element in the feature vector. They are separate pyramidal data structures with the same number of levels, which Cave calls *layers.* Each layer of a feature module is a network of units distributed in a uniform rectangular grid. Units hold feature values. A layer of the activation module is a rectangular array of bins. A bin contains a real number called the activation. All layers on the same level have the same array dimensions; layers on different levels have different dimensions.

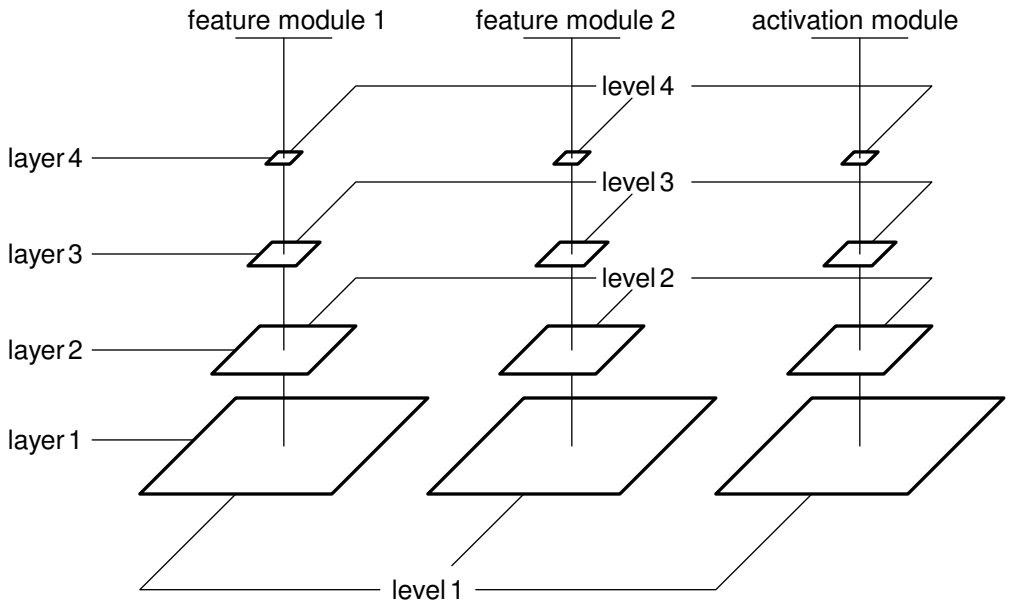


Figure 4. A FeatureGate network with 2 feature maps and 4 levels [45].

FeatureGate contains two subsystems to handle bottom-up and top-down attentional mechanisms. A top-down process is task-related. For example, the task may be to search for a particular person in a scene. A bottom-up process will identify the most salient location in the scene, independently of the task. The top-down subsystem passes locations with target features and inhibits those with dissimilar features. This subsystem is activated only when a target features is specified. The bottom-up subsystem compares features at one location to features at neighboring locations. It passes (allows through the gate) the vector that differs most within the neighborhood and blocks the others. An activation value proportional to that difference is calculated for each passing location. The activation values are summed within each subsystem and combined across the systems to yield the total activation for each location. The locations with the *n* highest activations are noted on the next level.

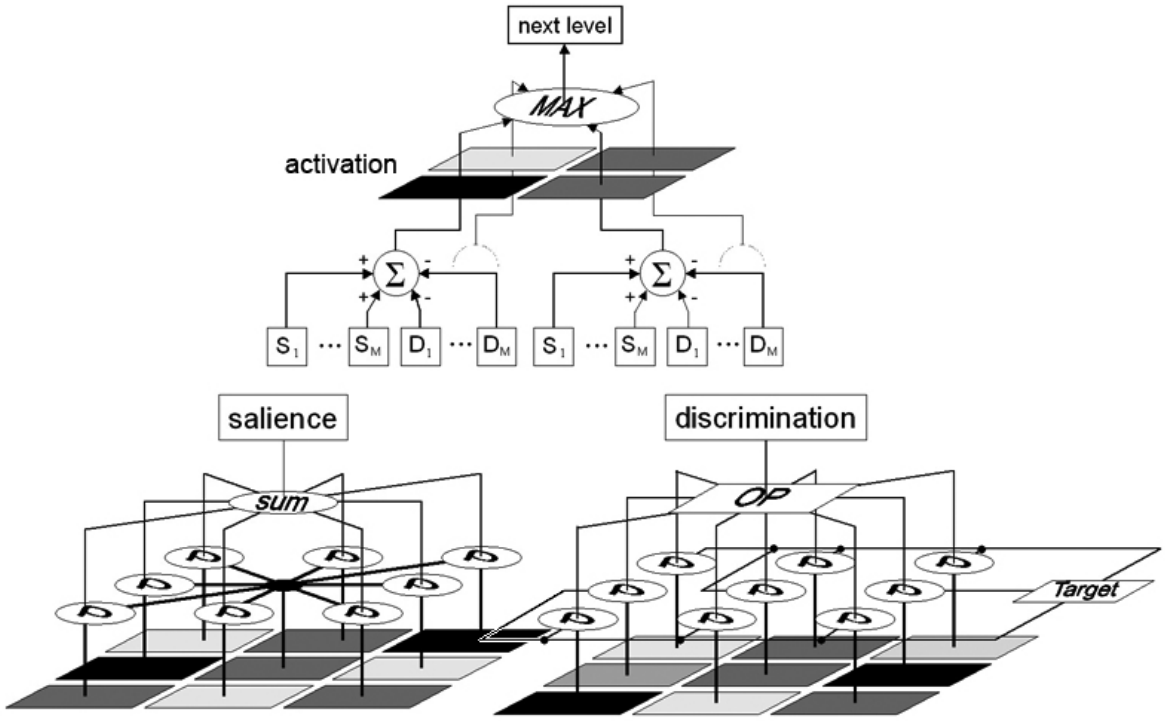


Figure 5. illustration of hierarchy and selection in FeatureGate [45]

The results of FeatureGate were similar to the results for the Itti model; a target is usually found when it differs from the distractors in only one feature. However, if a conjunctive search is undertaken (if the target differs from the distractors in more than one feature), the output location is not usually correct. This is due to the fact that a distractor may differ from the surrounding distractors significantly in a neighborhood, thereby generating a high activation value. To address this problem, inhibition of return is used to inhibit the winning location so that another location can reach the output. This is repeated if necessary until an appropriate target is found.

The FeatureGate model is used in the control system of ISAC, the humanoid robot at Vanderbilt University [9]. FeatureGate was implemented to identify a fixation point in a

stereo pair of images. The results were shown to be consistent with human attention and, therefore, justify the use of FeatureGate as a model for visual attention [9]. FeatureGate is used for attentional biasing on the SES.

Guided Search

Guided Search is a model of attention similar to FeatureGate. The idea behind Guided Search is that high-level processing cannot be performed on the visual scene as a whole. Instead, basic processing is done on a visual scene to identify regions of potential interest. It guides attention to these regions so that higher-level processing can be performed efficiently. Feature maps for color and orientation are created. The activation of each pixel in the image is calculated in two steps: the bottom-up component of the activation is computed as the difference between the value of a particular pixel and its neighbors. The top-down component is computed as a function of the similarity between a pixel and the values of the target, as is done by FeatureGate. Activation values from the feature maps are summed to form an activation map, which directs attention to the area with the highest activation [40]. Figure 6 illustrates the structure of the Guided Search Model.

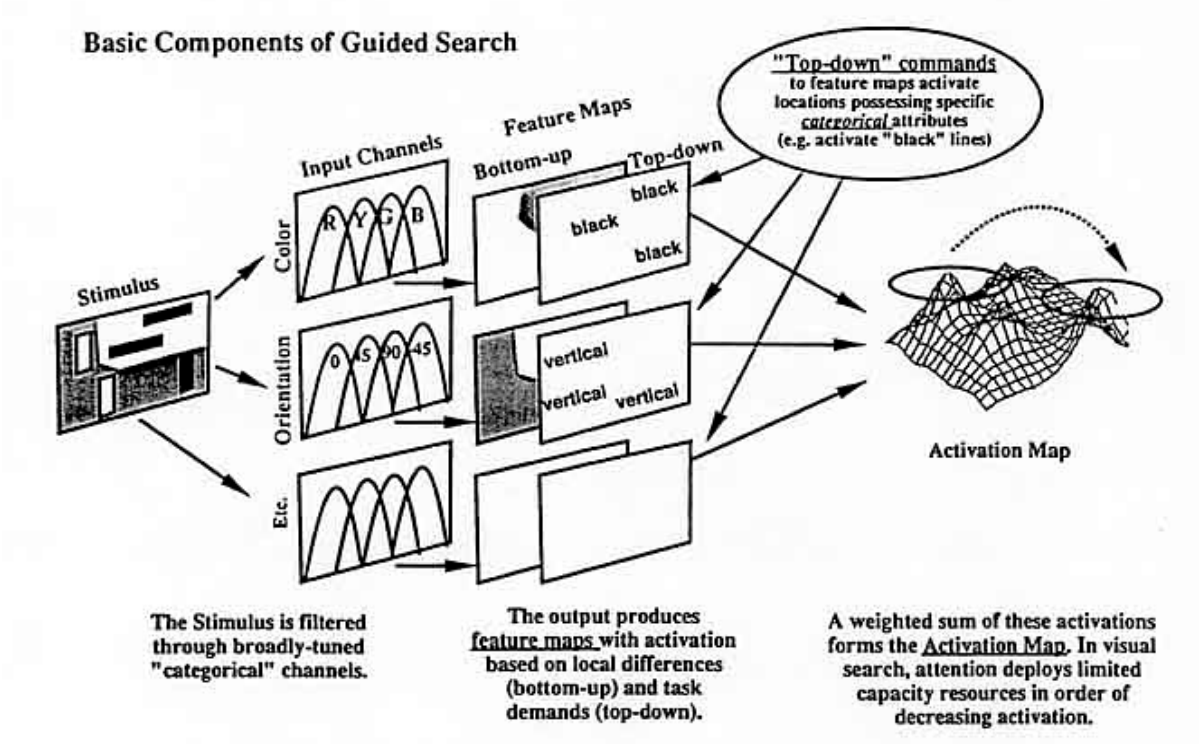


Figure 6. Structure of Guided Search [40]

To model the human vision system more accurately, a later revision incorporated eye movements and made use of the spatially varying acuity of the retina [41].

Figure 7. Architecture of Guided Search 3.0 [41]

A further revision incorporated a size feature map and did not have memory of rejected locations of interest (as did the first model) to be more accurate in the modeling the human visual attention system [42].

All models described above use feature maps. In a recent Nature Neuroscience review, Wolfe and Horowitz [43] listed the attributes that can guide attention to salient locations in a visual scene. As can be seen from table 1, the models rely only on a few features, mainly orientation and luminance. The table is reproduced from data presented in the article and identifies features that are likely to direct attention and others that probably do not.

Table 1. Attributes and likelihood of guiding attention [43]

Undoubted Attributes	Probable Attributes	Possible Attributes	Doubtful Cases	Probable Non-Attributes
-Color -Motion -Orientation -Size (including length and spatial frequency)	-Luminance onset (flicker -Luminance polarity -Vernier offset -Stereoscopic depth and tilt -Pictorial depth cues -Shape -Line termination -Closure -Topological status -Curvature	-Lighting direction (shading) -Glossiness (luster) -Expansion -Number -Aspect Ratio	-Novelty -Letter identity (over learned sets in general) -Alphanumeric category	-Intersection -Optic flow -Color change -Three-dimensional volumes -Faces (familiar, upright, angry, and so on) -Your name -Semantic category (ex: ‘animal’, ‘scary’)

Previous Work

The Sensory Ego-Sphere

The Sensory Ego-Sphere (SES) is a mediating interface between sensors and cognition that structures and coordinates sensory information for further processing. It mimics some of the functionalities of the mammalian hippocampus, which is critically involved in short-term memory, integration of multiple sensory inputs, sensory-motor integration, egocentric mapping of the environment, and position (of self) with respect to objects in the environment [30].

The SES is structured as a geodesic dome, which is a quasi-uniform triangular tessellation of a sphere into a polyhedron [10]. The geodesic dome structure was chosen to represent the SES because it is "the optimal solution to the problem of how to cover a sphere with the least number of partially overlapping circles of the same radius" [35]. A geodesic dome is composed of twelve pentagons and a variable number of hexagons that depend on the frequency (or tessellation) of the dome. The frequency is determined by the number of vertices that connect the center of one pentagon to the center of another pentagon, all pentagons being distributed on the dome evenly. The number of vertices (V) can be determined from the frequency (N) using equation 2.1.

$$V = 10N^2 + 2 \tag{2.1}$$

To create a geodesic dome, an icosahedron (which is made up of 12 vertices) is first created. This is a dome with a frequency of 1. To increase the frequency of the dome to two,

a new vertex is added at the midpoint of each edge. Each new vertex, already connected to the two vertices on the original edge, is connected to the 4 four new vertices nearest it, This results in a total of six neighbors for each new vertex and, at the same time, adds hexagons to the overall structure. The original pentagon centers are connected to five neighbors while the new hexagon centers are connected to six neighbors. Subdivision continues until the polyhedron has the desired frequency. Once all vertices have been added to the polyhedron structure, all vertices are moved so as to be an equal distance away from the center, creating the geodesic dome. Figure 8 illustrates this process.

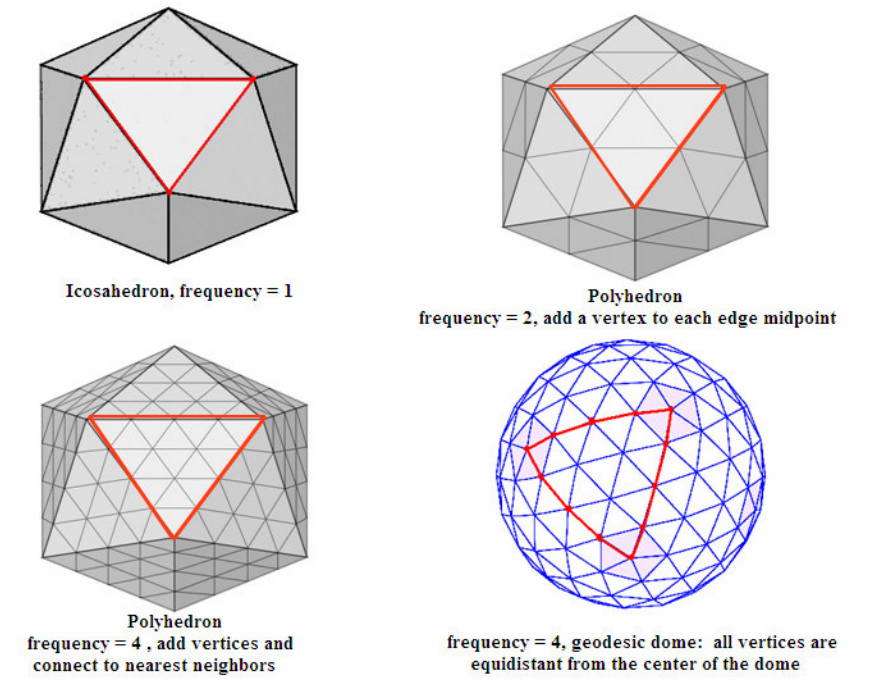


Figure 8. Tessellation of an Icosahedron into a Geodesic Dome [45]

The SES is centered on a robot's coordinate origin to provide an egocentric representation of the environment (Figure 9). The SES is indexed by azimuth and elevation angles.

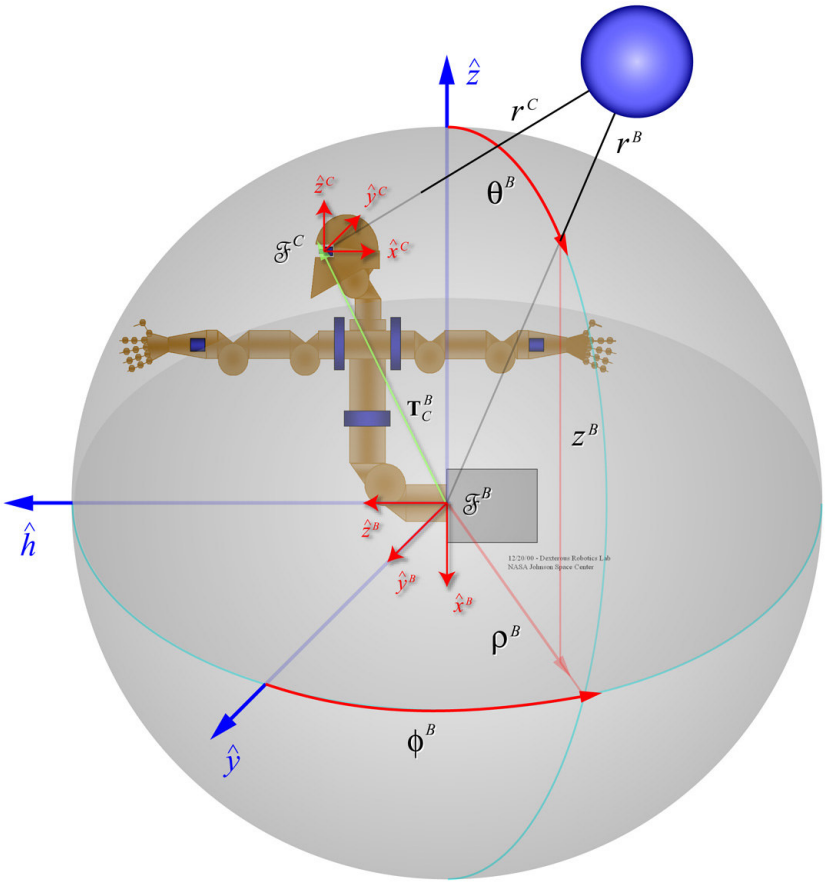


Figure 9. A robot within its SES [30]

Sensory data is stored onto the sphere at the node closest to the origin of the data (in space). For example, an object that has been visually located in the environment is projected onto the sphere at azimuth and elevation angles that correspond to the pan and tilt angles of the camera-head when the object was seen. A label that identifies the object and other relevant information is stored into a database. The vertex on the sphere closest to an object's projection becomes the registration node, or the location where the information is stored in the database, as illustrated in Figure 10 [30].

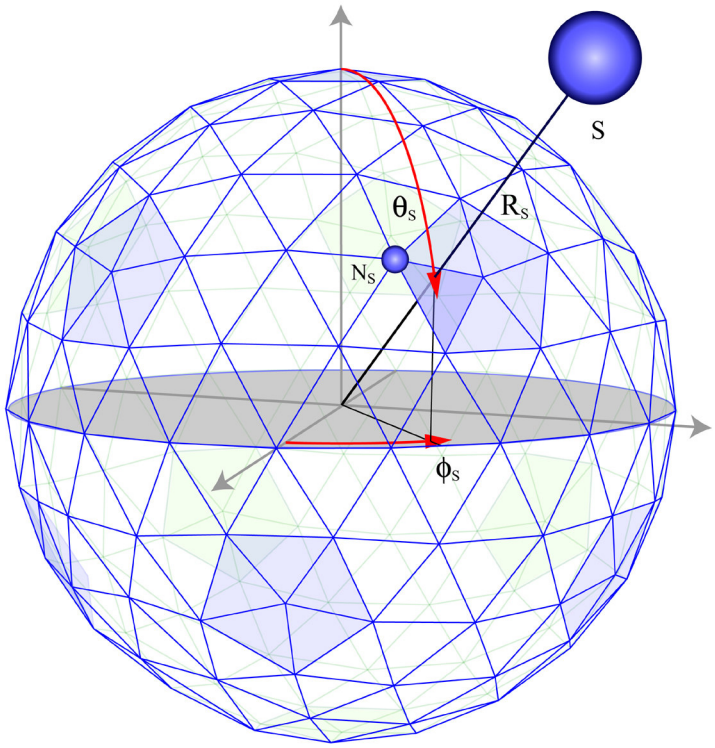


Figure 10. Projection of an object onto the SES [30]

Recently it was shown that sparse and/or simple sensory modalities can be combined accurately using an SES and that attention can be cued by such combinations [13]. More specifically, sensory information having the same source (for example: the sound of the voice and the image of the face of a person sensed simultaneously from the same direction in space) can be recognized as such through spatio-temporal coincidence of stimuli[10], and can increase the salience of a particular location on the SES [13].

The SES also serves as a short-term memory. This is necessary for temporal binding, since sensory information must be stored in memory to be combined with sensory information detected later. Different sensory events at the same location in space but different locations in time can also be accumulated to form a focus of attention on the sphere.

[10] Occurring at approximately the same time and/or in the same location.

The work to date on the SES has involved mapping and combining sparse sensory information such as localized sounds and discrete objects. It has not addressed complex sensory modes such as vision, proprioception, or touch. This thesis describes the first work done on the SES to map and to combine dense sensory information in the form of images.

CHAPTER III

RESEARCH SYSTEM

Robotic Platform: ISAC

Experiments for this thesis were performed using ISAC (Intelligent Soft Arm Control), a humanoid robot developed at Vanderbilt University [20, 28]. Originally, ISAC was developed as a robotic aid system for the physically disabled [19] but has evolved into a research platform for the study of human-robot interactions. ISAC (Figure 11) is comprised of two 6 degrees-of-freedom (DOF) arms that are controlled by McKibben artificial muscles [19]. ISAC also has an active vision system implemented using two pan-tilt units and Sony XC-999 cigar cameras [36]. The PTU-46-17.5 pan-tilt units are manufactured by Directed Perception Inc. The range of the pan-tilt units is approximately ±159° in pan (for a total of 318°), and + 31° to -47° in tilt, with option of 80° down, for a total of 111° in tilt [8]. Figure 12 illustrates ISAC's pan-tilt axes. The pan/tilt units have a resolution of 0.051428 degrees [8]. The cameras are ½" CCD color video cameras with 768 (H) x 494 (V) effective picture elements [34].

The images processed were 320x240 color images. The full images were processed but only a portion of each image, the region around the optical center of the image corresponding to a node of the SES, was displayed.

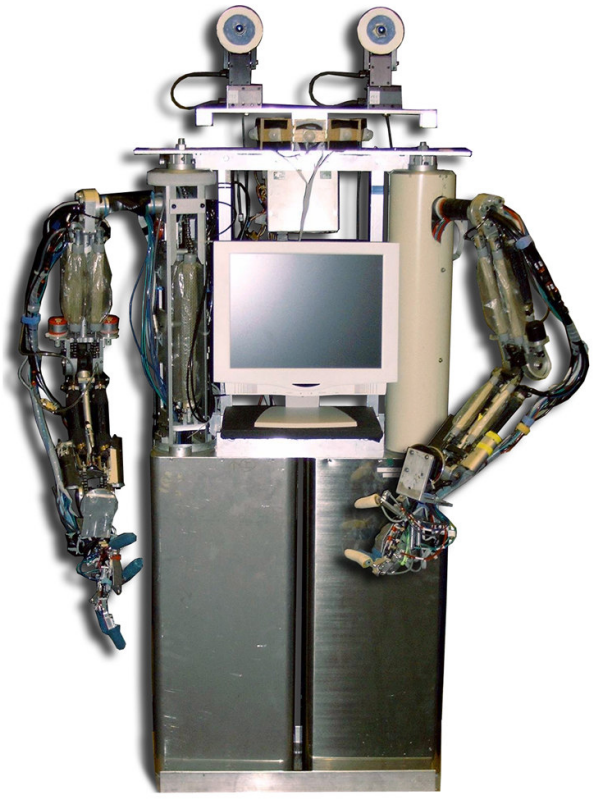


Figure 11. ISAC Humanoid Robot [45]

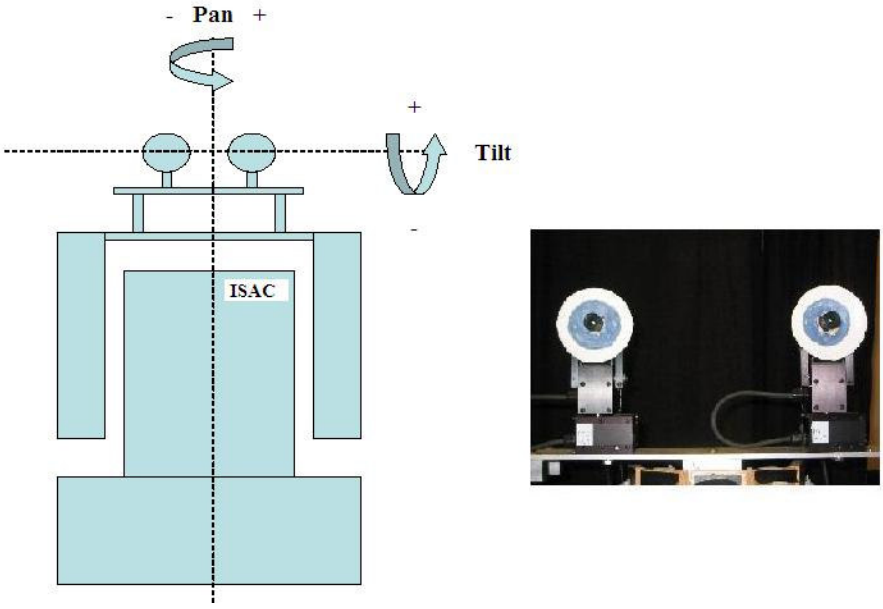


Figure 12. ISAC pan/tilt axes and camera-head

Sensory Ego-Sphere

The Sensory Ego-Sphere used in this work has a tessellation of 14; it is comprised of 1962 vertices or nodes where information can be stored. The distance between nodes varies between 4 and approximately 6 degrees [13, 30]. This will be explained in greater detail in Chapter IV (cf. p. 48). The SES is centered at Vanderbilt's humanoid robot ISAC's head origin (cf. p. 43).

Database

The informational structure of the SES is a connected graph of pointers to data structures that are indexed by a unique ID. Each vertex (or node) has 6 or 7 pointers, depending on the number of neighbors a node possesses. A node at the center of a pentagonal neighborhood will have five nearest neighbors while a node at the center of a hexagonal neighborhood will have six. The extra pointer in both cases points to a list of variable length that contains pointers to data structures. These structures are database records that contain the unique ID of the node at which information was posted, its location, and a timestamp [13]. Other relevant information about the posting can be found in several database tables and is accessed through the unique ID. Figure 13 demonstrates the database connectivity.

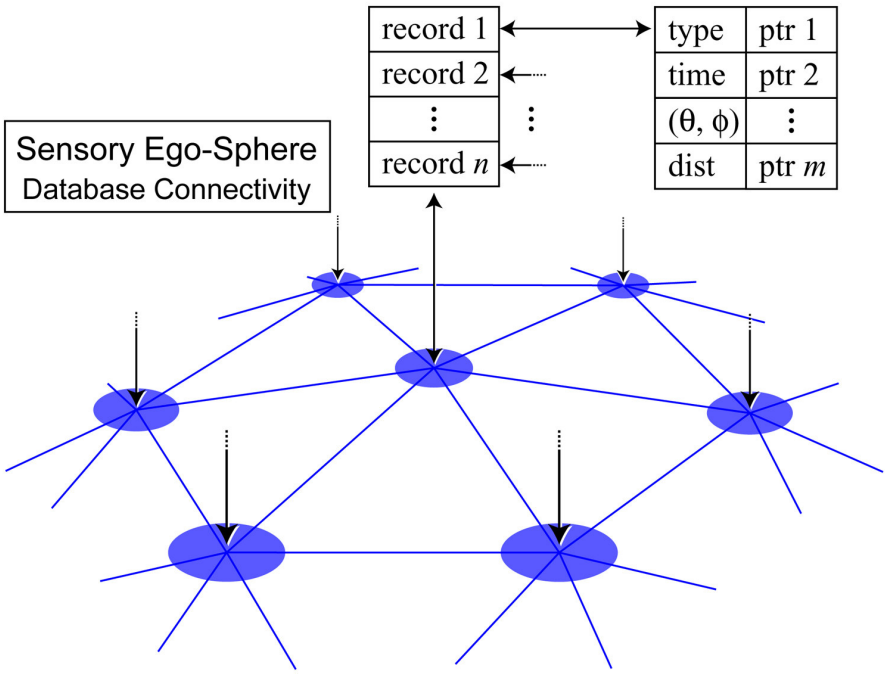


Figure 13. Database Connectivity. Each node points to its neighbors and to a database record that holds relevant information about the node posting and other related records [30].

For this work, a MySQL server was used and a SES database was created for ISAC. This database contained four tables, (1) a table of postings on the SES, (2) a table of all nodes on the SES, their locations, and a list of their nearest neighbors, and (3) & (4) two tables used for attentional processing. A description of these can be found in Table 2. TblNodes contains each node's azimuth/elevation angle pair (phi/theta) as well as i and j integer indices which correspond to the elevation and azimuth angles respectively. The indices facilitate finding a node's nearest neighbors [13]. The contents of tblActivation and tblAttentionalLocations will be described in greater detail in a subsequent section.

Table 2. List of tables in the SES_ISAC database

Table Name	Table Contents
tblNodes	Each node's phi/theta location, i/j indices, and list of nearest neighbors
tblSES	Information about each posting made to the SES, including pan/tilt location, node ID, name, type, image name, and a timestamp
tblActivation	A list of top 12 attentional points found in the camera image taken at the pan/tilt location of a particular nodeID at their activation values
tblAttentionalLocations	A list of all attentional points that map to a particular node ID from neighboring camera images

Sensory Ego-Sphere Interface

The SES interface is a Visual Basic 6.0 application that allows posting to, and retrieving from the SES. There are six different retrieval methods. They are displayed in table 3 along with their description. Note that the methods involving retrieval by location will return all postings at the node closest to the location and its immediate neighbors (5 or 6 depending on the connectivity around the given node). It is also possible to enter a pan/tilt angle pair to retrieve the ID of the node closest to that pair. The contents of the SES database can be cleared with a single command. A new SES can be created from this application, and different Sensory Ego-Spheres can be accessed from the same application. Image centers could be posted using this application; however, due to the amount of images to post for this work, the process was automated with a Matlab program. The SES interface is shown in Figure 14.

Table 3. SES retrieval methods

Retrieval Method	Description
Retrieve by name	Returns all postings in tblSES with matching name
Retrieve by name and type	Returns all postings in tblSES with matching name and type
Retrieve by type	Returns all postings in tblSES with matching type
Retrieve by name and location	Returns all postings in tblSES with matching name and pan/tilt angle location (or immediate neighbor)
Retrieve by type and location	Returns all postings in tblSES with matching type and pan/tilt angle location (or immediate neighbor)
Retrieve by location	Returns all postings in tblSES with matching pan/tilt angle location (or immediate neighbor)

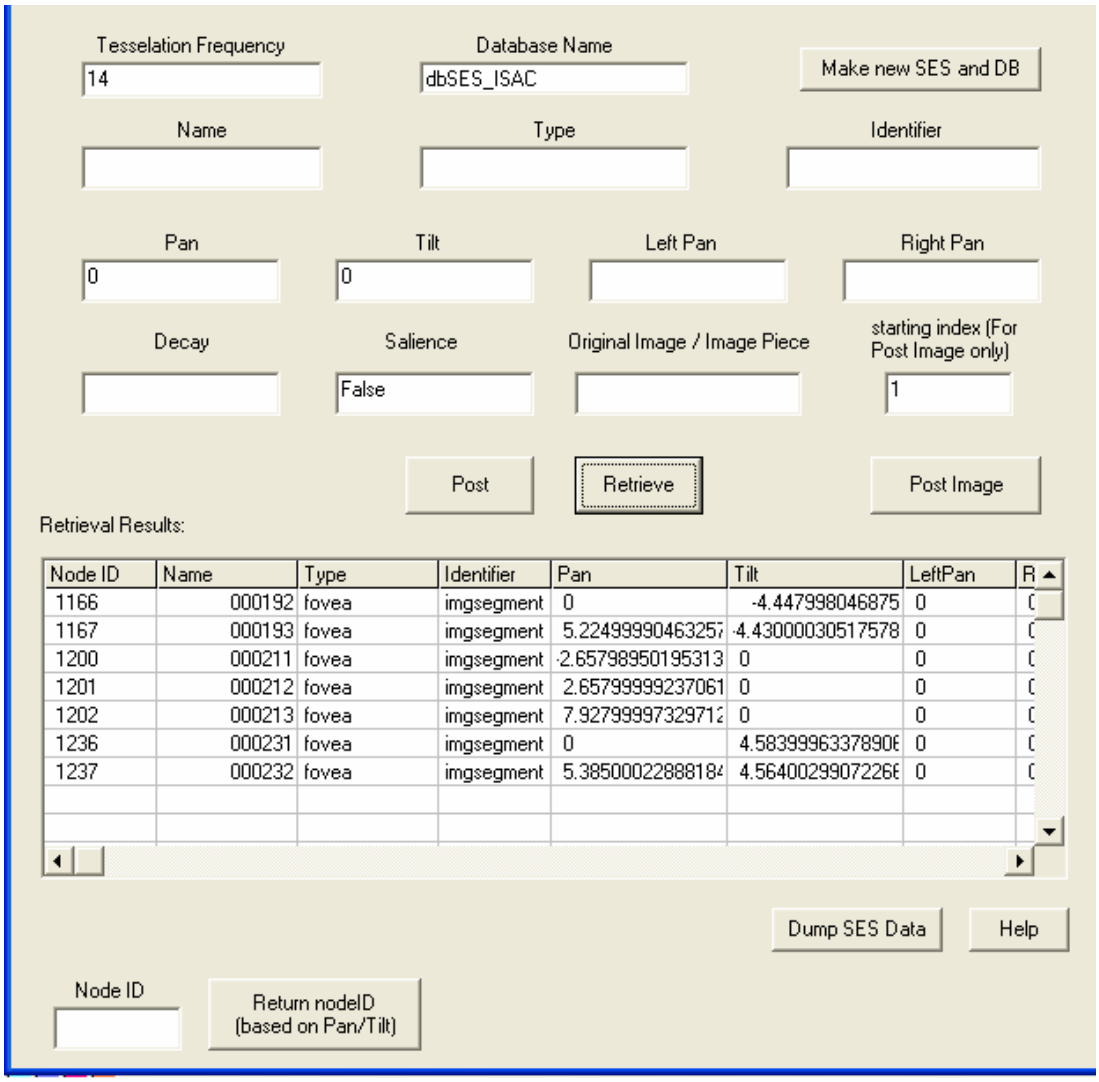


Figure 14. SES Interface

Figure 14 gives an example of the retrieval by location method. The tessellation and name of the database were specified, along with the desired pan and tilt angles (in this case, pan = 0 and tilt=0). After pressing the "Retrieve" button, all records in tblSES posted at the node corresponding to the pan/tilt location or any of its immediate neighbors are returned in the "Retrieval results" window.

Visual Attention Implementation

The FeatureGate model of visual attention was implemented for this research. The implementation uses three separate feature maps: one each for color, luminance, and orientation, (like the second stage of the Itti algorithm, cf. p. 16). The orientation processing is implemented by a Frei-Chen basis [33]. There are nine Frei-Chen basis images, into which the original image can be decomposed uniquely. They are as follows: (1) average pixel value, (2) horizontal edge, (3) vertical edge, (4) diagonal edge (down to the right, (5) diagonal edge (up to the right), (6) ecks corner detector (diagonal corner), (7) plus corner detector (horizontal / vertical), (8) plus intersection detector, and (9) ecks intersection detector. Examples of features 2 - 9 are illustrated in figure 15.

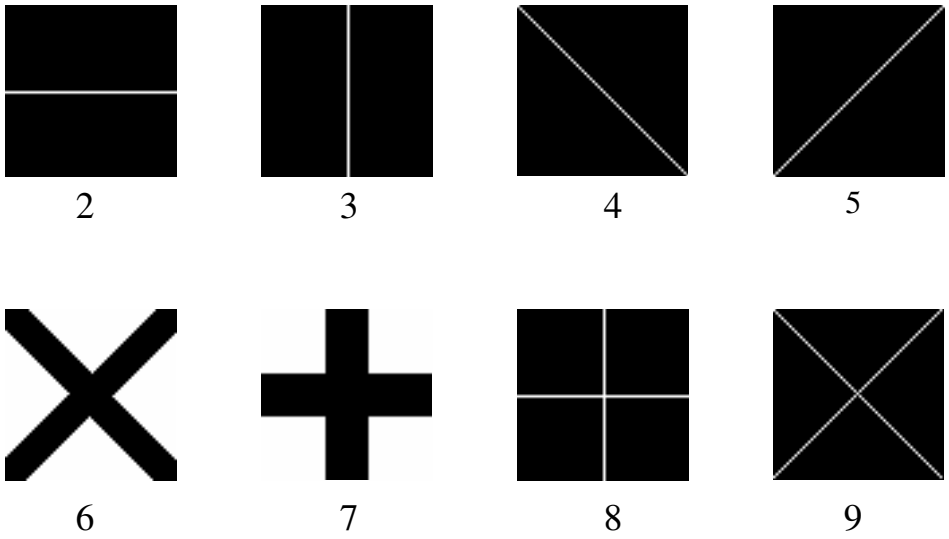


Figure 15. Examples of 8 features included in the respective Frei-Chen Components of an image

In the bottom up process, each pixel location's features are compared to the features of its 8 nearest neighbors by Euclidean distance and the results are added and saved in the

activation map. In the top-down process, each pixel location's features are compared to the known target features. In usual FeatureGate implementations, an arbitrary number of locations with high activation values move on to the next level (example: downsample each row and column sizes by 2 at each level), where the process is repeated until only one location is left, the focus of attention. In this implementation, the locations with the highest activations from the first level are selected as foci of attention and recorded to reduce the processing time. For individual images in the image sequence, this implementation returns the 12 most salient locations (row and column location) and their activation values in a saliency array structure. This array also includes the pan and tilt angles of the image being processed. Only one salient location per neighborhood of 15 by 15 is returned, because locations very close together most likely refer to the same physical feature and a good covering of the image is wanted. A variable number of activation locations can be chosen on the reconstructed scene from the SES. For better results, the incoming images were first blurred using a constant filter. The filter was 3x3 and filled with ones. This was found to help maintain the locations of salient points in sequential images. A series of 10 images taken from the camera at the same pan/tilt angle was processed in FeatureGate three times: once without a filter, once with a constant filter, and once with a median filter. The attentional points selected in the series were most constant from one image to the other with the constant filter. By blurring the image, FeatureGate processing is less susceptible to minuscule, insignificant changes that occur from one image to the next.

CHAPTER IV

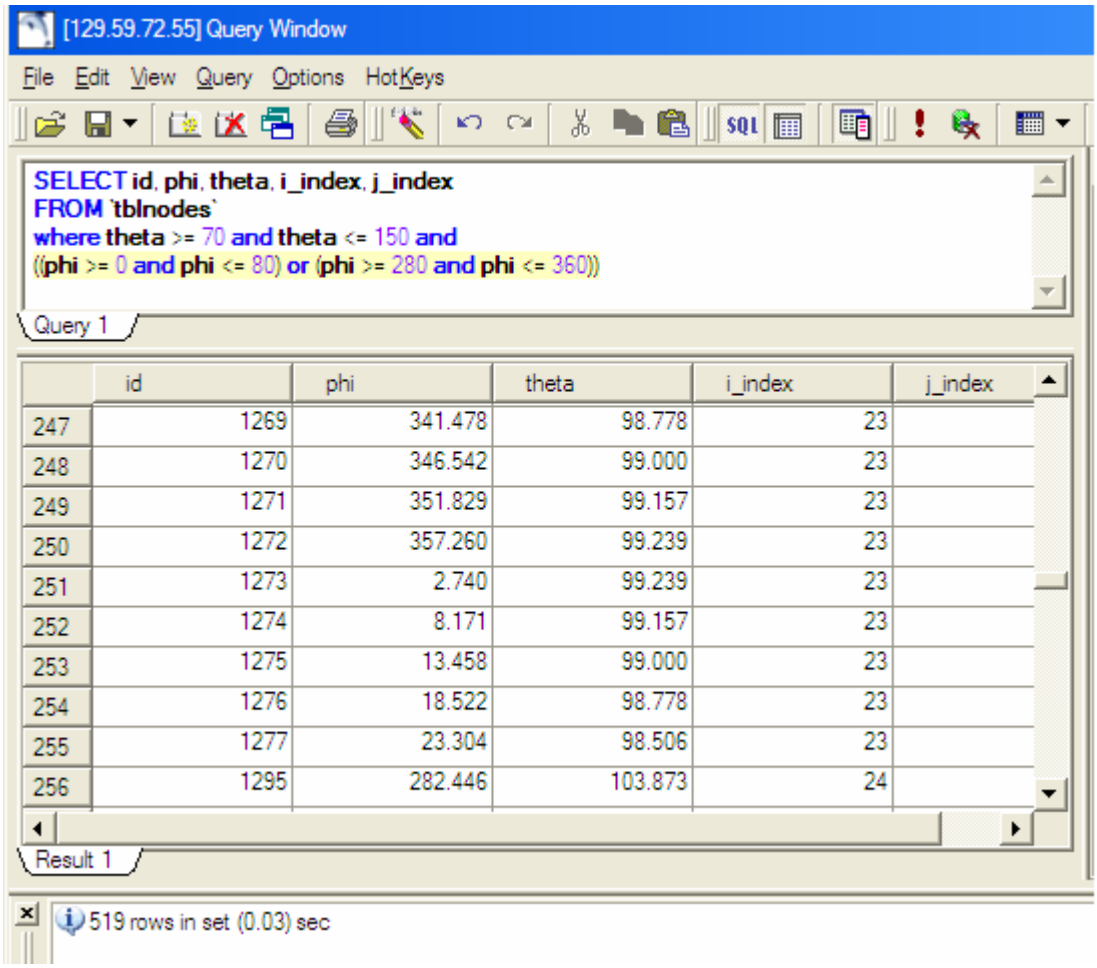
METHODS

In previous works, images were preprocessed through color segmentation and edge extraction to identify and extract relevant objects from them. These objects were then posted on the Sensory Ego-Sphere at their particular location in space; this location has less-than-pixel resolution because of the uncertainty in object localization [13]. A collection of objects is a sparse set of data on the SES. In contrast, this work involved the mapping of dense sensory data to the SES. These data were images taken by ISAC's camera-head. The images were not preprocessed and no particular objects were identified. The camera-head was caused to traverse its workspace while grabbing images. The result was a complete mapping of the visual scene onto the SES. Note that for this work the visual scene was mapped with only one of the two cameras (the left one). The same procedure could be followed for each of the two cameras, and would have to be if stereo image pairs are to be mapped onto the SES.

Image Sequence Generation

ISAC has a forward-looking camera-head platform. Although it has pan-tilt-and-verge capabilities, the cameras cannot rotate through 360 degrees (see below) and cannot, therefore, map the entire SES. A connected subset of the SES within the area of +20 to –60 degrees in tilt and +80 to –80 degrees in pan was chosen because both cameras can cover it, and the ±80º pan range is consistent with the human field of view [39]. The task of mapping

a complete visual scene onto the Sensory Ego-Sphere was accomplished by first compiling a list of all the SES nodes within the field of view. Since the nodes on the Ego-Sphere are indexed by elevation and azimuth angles (phi and theta), these measurements correspond to theta values between 70 and 150 degrees and phi values from 0 to 80 degrees and from 280 to 360 degrees. The list was obtained by querying the database table tblNodes (cf. Table 2) to return all nodes on the SES in the chosen region. Figure 16 shows an excerpt of the query results.



	id	phi	theta	i_index	j_index
247	1269	341.478	98.778	23	
248	1270	346.542	99.000	23	
249	1271	351.829	99.157	23	
250	1272	357.260	99.239	23	
251	1273	2.740	99.239	23	
252	1274	8.171	99.157	23	
253	1275	13.458	99.000	23	
254	1276	18.522	98.778	23	
255	1277	23.304	98.506	23	
256	1295	282.446	103.873	24	

Figure 16. tblNodes query for image sequence generation

Once this list of node ID, phi, and theta values was obtained, the corresponding pan/tilt angle pairs were calculated for the camera-head to position the pan/tilt units so that the optical center of the camera was at each node's location. The conversion between phi (φ)/theta (θ) and pan/tilt is necessary since node locations are expressed in φ, θ and the pan/tilt units in the camera-head use (obviously) pan and tilt angles for positioning. Figure 17 demonstrates the relationship between phi/theta and pan/tilt representations on the SES with respect to ISAC.

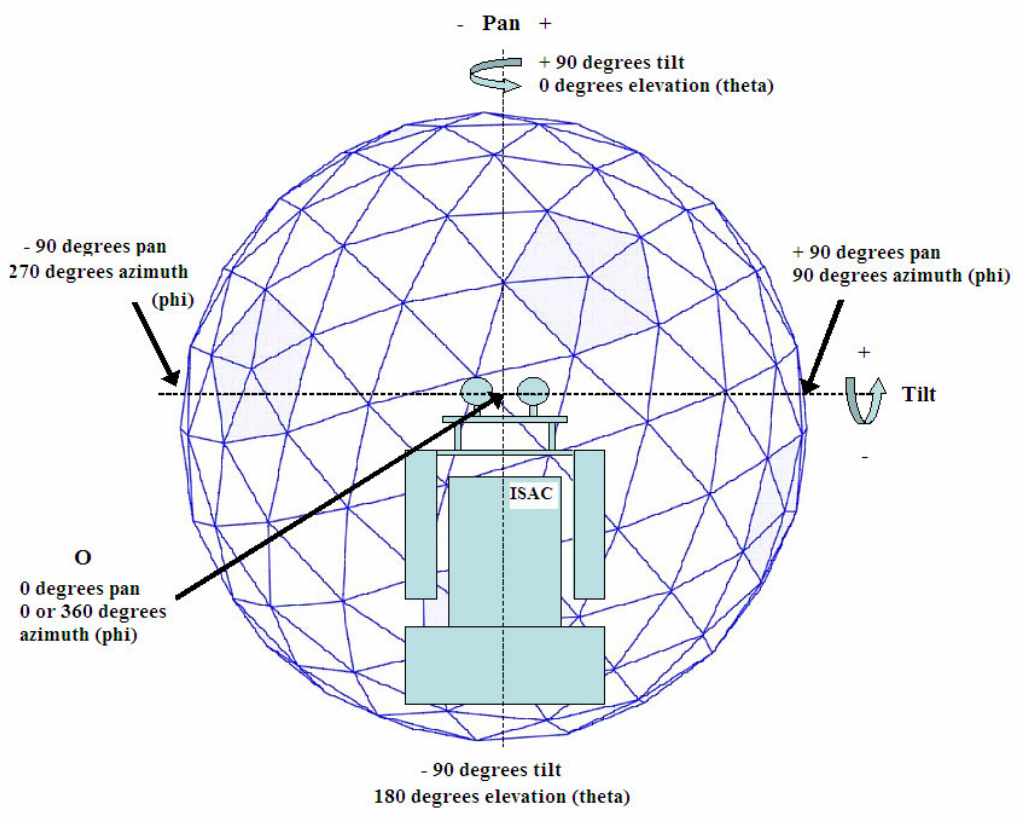


Figure 17. Angles with respect to ISAC's SES

As can be observed from this figure, theta (θ) ranges from 0º at the North pole to 180º at the south pole. Phi (φ) ranges from 0º at the front center of the SES (point O in the figure)

to 360º at the same point, moving in a counter clockwise direction around the vertical axis. For ISAC's camera-head, positive tilt is defined as being upward while negative tilt is downward, for a total range of ±90º. Similarly, positive pan is defined as being to the left (ccw), negative tilt is to the right (cw, with respect to the vertical axis), meeting in the back center at ± 180º. The conversion equations between pan/tilt and phi/theta are.

$$\text{tilt} = 90 - \theta \tag{4.1}$$

$$\text{pan} = \varphi \text{ , } \varphi \leq 180 \quad \text{pan} = \varphi - 360, \varphi > 180 \tag{4.2}$$

Since the pan range of a Directed Perception pan/tilt unit is limited to ±159º and tilt to +31º and -80º [8], the whole SES cannot be populated with images. But the range chosen (±80º pan, +20º to -60º tilt) is adequate as explained above.

A sequence of 519 images was then generated by taking a picture at each of the pan/tilt locations in the list. For reference within the database (and the computer's file system) all the images were given the same base name and each a unique number that corresponded to the index of the node in the list (Figure 18).

Index	PAN	TILT
330	-24.67	-25.333
331	-19.659	-26.132
332	-14.315	-26.784
333	-8.705	-27.247
334	-2.922	-27.487
335	2.922	-27.487
336	8.705	-27.247
337	14.315	-26.784
338	19.659	-26.132
339	24.67	-25.333
340	29.309	-24.431

Image000333.bmp **Pan/tilt angle pairs array**

Figure 18. Example of an image and its corresponding pan/tilt angles

Camera Calibration

The mapping of an image to a node of the SES requires metric information about the images (*e.g.*, angular resolutions of the pixels) and about the cameras. For the latter, we need to know the true optical centers and the focal lengths. That requires camera calibration. The left camera was calibrated using Caltech's Matlab Camera Calibration Toolbox [25]. The procedure is outlined on the Calibration Toolbox's website.[11] The focal length array was calculated to be: [307.97804 307.13756] ± [1.91436 2.03466][12] and the optical center was calculated at: [160.59226 125.41502] ± [2.68781 2.10154]. These measurements are in pixels. The camera's field of view is ½", or 12.7mm [34], and the image dimensions are 320 by 240. Using these values, we find that a pixel is .0397mm in pan and .0529mm in tilt. The focal length is then calculated to be [12.2267, 16.2476] in millimeters. Since each pixel

[11] http://www.vision.caltech.edu/bouguetj/calib_doc/index.html#parameters
[12] Due to imperfections in the lens system, the effective pin-hole focal length differs for projections in the *x*-direction and projections in the *y*-direction, hence the dyadic focal length measurement.

is square, whereas the image has an aspect ratio of 4-to-3, the smaller pixel size and focal length (12.2267mm) were used.[13]

The optical center location was used to define a rectangular foveal window in each image to display on the SES: instead of using the center pixel, (160,120), of the image as the center of the fovea, the calculated optical center rounded to the nearest pixel, (161,125), was used.

Populating the Sensory Ego-Sphere

A program was written to populate the Sensory Ego-Sphere with a sequence of images. Each image was taken at a pan/tilt angle pair that corresponded to a particular node on the SES; more precisely, the image center corresponded to that angle pair. A foveal window at the center of each image in the sequence was extracted and posted on the SES at the correct node location. Figure 19 illustrates this procedure.

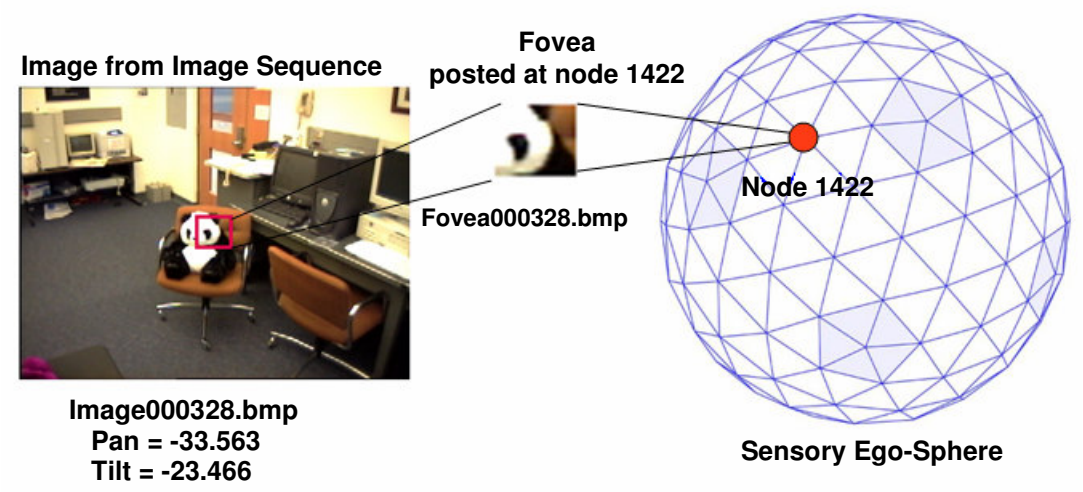
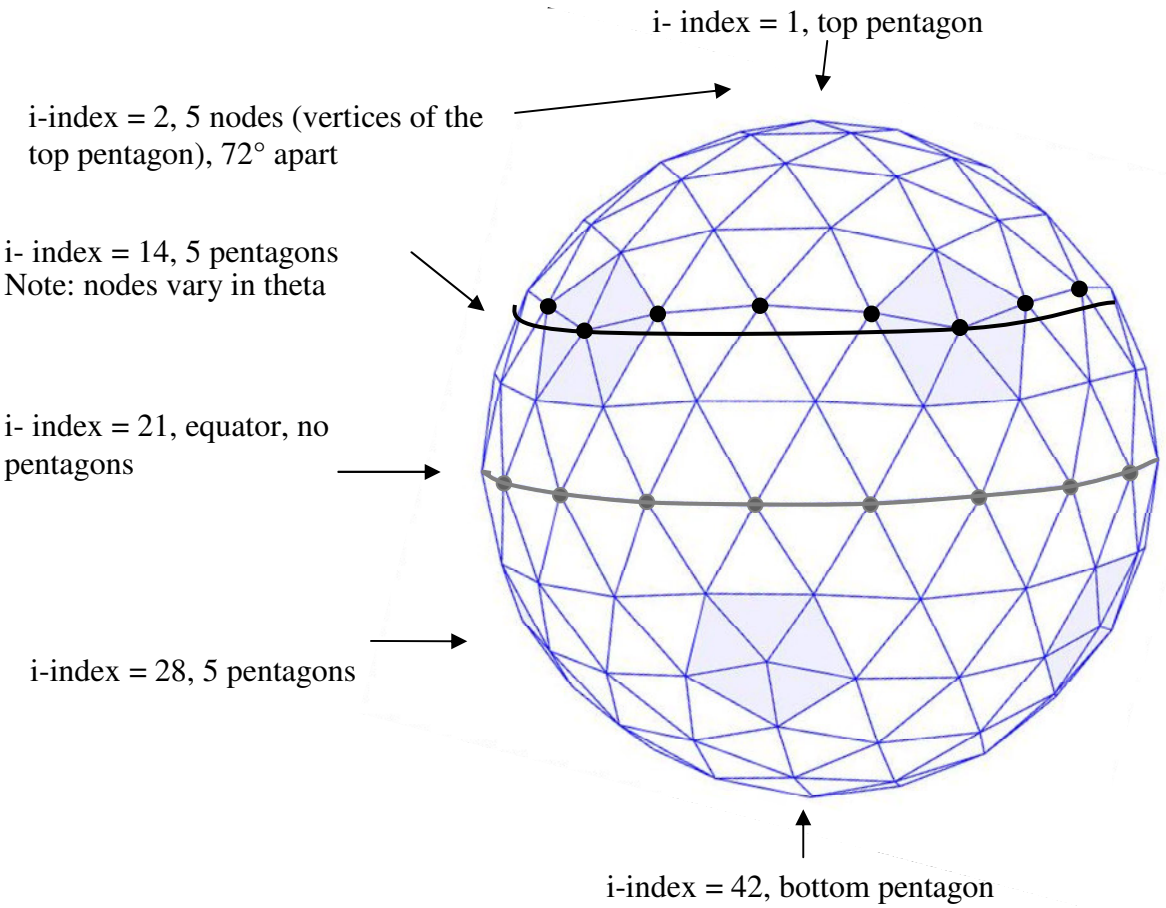


Figure 19. Posting a fovea onto the SES

[13] Although the sensor array is square, standard NTSC cameras have a 4-to-3 rectangular aspect ratio of, typically 640×480 rectangular pixels. That implies that ¼ of the vertical extent of the array, presumably 1/8 on the top and 1/8 on the bottom are unused.

The size of the foveal window taken from the center varied but generally corresponded to approximately 5º in pan and 5º in tilt since this is the distance that separates most nodes on a geodesic dome with a frequency of 14 [30]. However, because both pentagons and hexagons make up the dome, edges between nodes on a geodesic dome do not all have the same length, even if they have the same i-index. For example, nodes with an i-index equal to 10 have theta values ranging from 40.501º to 46.5529º. As was mentioned in chapter III, i and j integer indices correspond to the elevation and azimuth angles respectively. These indices facilitate the task of finding a node's nearest neighbors. All nodes that lie within a particular theta (tilt or elevation) range are assigned the same i-index. A j-index starting at 0 is assigned to the node whose phi (pan or azimuth) is closest to 0. Indexing proceeds in the counter-clockwise direction around the vertical axis. Nodes increasingly closer to either pole of the SES are farther apart in pan than those close to the equator; for example, there are only 5 nodes at an i-index of 2 (top of the sphere) and they are 72º apart. Therefore, the dimensions of the foveal window must be selected as a function of the number of degrees between the node and its neighbor nodes. The neighbor node of reference has the same i-index (or relative elevation angle), but at a j-index incremented by 1 (greater azimuth angle). All nodes on the equator (i-index = 21) have theta values of 90º (gray line on figure 18). The pentagons are distributed on the dome as follows: There is one at each pole and 5 each at index i = 14 (black line on figure 18) and index i = 28, which correspond to thetas of approximately 60º and 120º respectively. Figure 20 illustrates this while table 4 gives a listing of node distances at several i-indices.



Note: this SES has a tessellation of 4, but i-indices for tessellation of 14 used in this thesis are displayed. Tessellation refers to the number of edges that separate the center nodes of pentagons.

Figure 20. Pentagon distribution on the SES

Table 4. Distance between nodes at different elevations on the SES

Location	i-index	min phi distance	max phi distance	min theta value	max theta value
Top of sphere	2	72º	72º	3.8054º	3.8054º
Line going through centers of pentagons	14	4.0873º	5.9255º	58.2827º	63.4349º
Equator	21	4.934º	5.316º	90º	90º

For precise results, the distances between each node and its 4 closest neighbors (top, down, left, and right) were calculated in degrees from the tblNodes data and converted to pixel measures. An appropriately-sized fovea was then extracted from the center of the image and the row/column location of the center of the fovea as well as its size and the image index and node ID were stored in a text file. This information is necessary to recreate an image of the visual scene from the foveae.

The pixel-per-degree measure was determined experimentally. Two images were taken, differing by a known degree measure in pan only (the same was done for tilt with two different images), and the images were placed one on top of the other so that their features overlapped as closely as possible. The pixel difference between these two images was then calculated and divided by the number of degrees separating the images. The pixels-per-degree measure for pan was found to be 6.072 pix/deg while the tilt measure was 5.536 pix/deg. The 28x30 average fovea size was calculated by multiplying these measures by 5º and rounding to the nearest pixel. Figure 21 shows the overlapped images used to calculate the pan pixels-per-degree measure.

Each fovea record was posted in the table tblSES in the SES_ISAC database and included the node ID where the fovea was posted on the SES, the name of the fovea (which is its index in the list of pan/tilt angle pairs), type, identifier, pan, tilt, and a timestamp. The size of each foveal window is kept in a separate file but could also be kept in this table. Figure 22 shows an excerpt from tblSES.

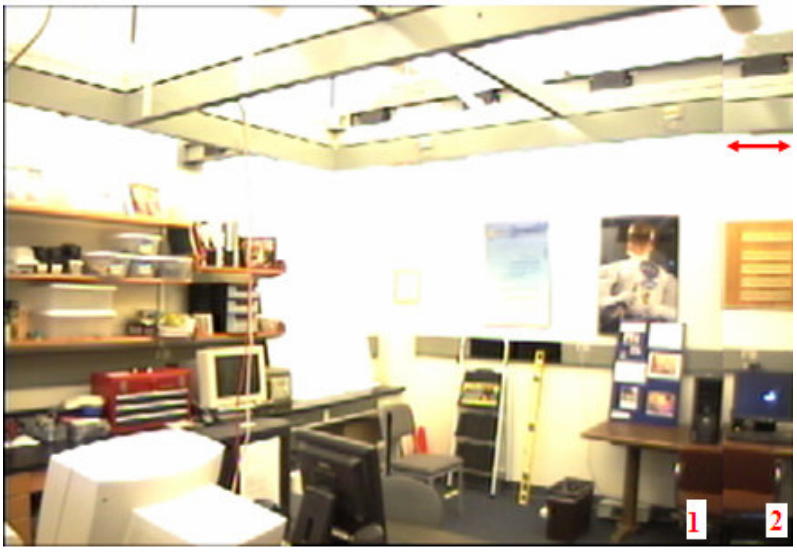


Image 1 pan: 43.928 tilt: 0, Image 2 pan: 38.658 tilt: 0

Figure 21. Image used to determine pan pixels-per-degree measure

	ID	name	type	identifier	actual_pan	actual_tilt	image_name
43	723	000043	fovea	imgsegment	30.6160	4.5640	fovea000043.bmp
44	724	000044	fovea	imgsegment	36.0000	4.5840	fovea000044.bmp
45	725	000045	fovea	imgsegment	41.3850	4.5640	fovea000045.bmp
46	726	000046	fovea	imgsegment	46.6760	4.5050	fovea000046.bmp
47	727	000047	fovea	imgsegment	51.7890	4.4120	fovea000047.bmp
48	728	000048	fovea	imgsegment	56.6580	4.2900	fovea000048.bmp
49	729	000049	fovea	imgsegment	61.6350	4.3760	fovea000049.bmp
50	730	000050	fovea	imgsegment	66.7750	4.4300	fovea000050.bmp
51	731	000051	fovea	imgsegment	72.0000	4.4480	fovea000051.bmp
52	732	000052	fovea	imgsegment	77.2250	4.4300	fovea000052.bmp
53	757	000053	fovea	imgsegment	22.9340	0.0000	fovea000053.bmp
54	758	000054	fovea	imgsegment	28.0730	0.0000	fovea000054.bmp

Figure 22. tblSES postings

A filename for the foveal window posted at the specific node ID was also recorded in tblSES. This filename was used to display all image pieces onto a graphical representation of

the Sensory Ego-Sphere. Figure 23 shows ISAC's empty Sensory Ego-Sphere while Figure 24 shows a visual representation of all the foveal images on the Sensory Ego-Sphere with respect to ISAC. Clifton's SESDisplay program was used to display the content on the sphere [4].

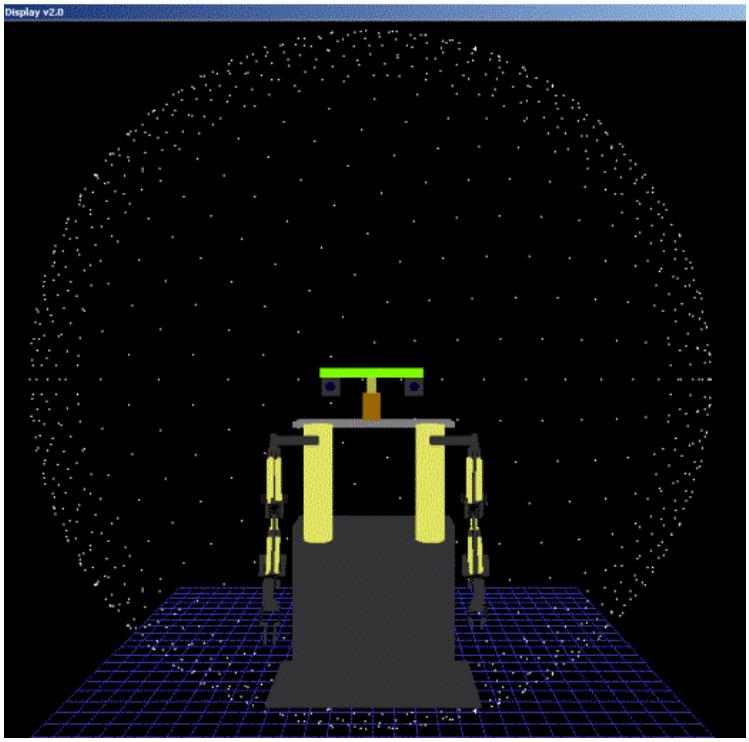


Figure 23. ISAC in its empty Sensory Ego-Sphere

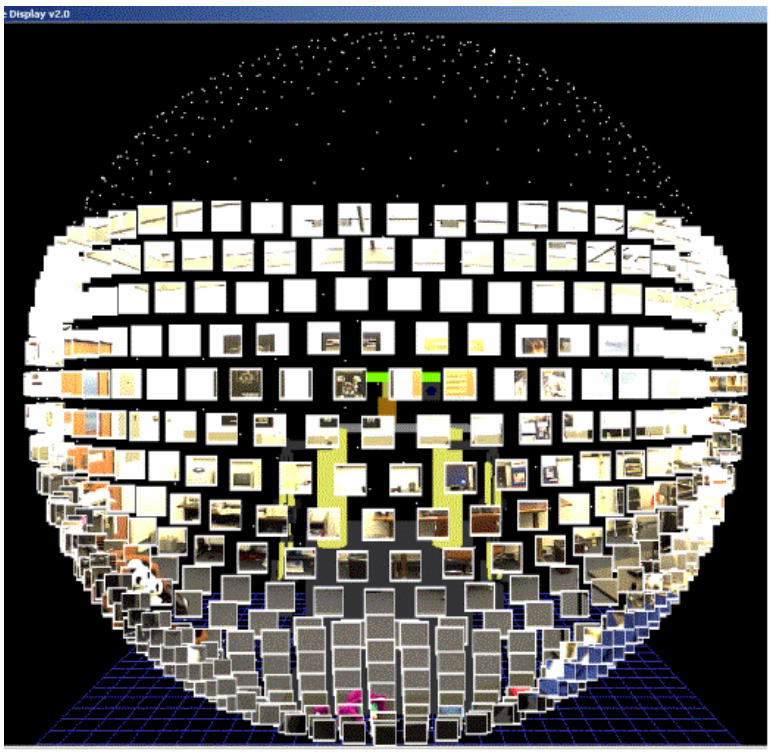


Figure 24. Visual scene posted on ISAC's Sensory Ego-Sphere

Image Reconstruction

A program was written that uses tblSES to reconstruct from all the foveal images, an approximately continuous image of the visual scene. The program generates a node map of the reconstructed scene image. The map associates each pixel in the image with a node on the SES. A reconstructed image is illustrated in Figure 25.

Figure 25. Reconstructed scene from SES fovea images.

The distance between each node on the SES varies as a function of tilt angle and 12 of the nodes are connected in a pentagonal arrangement whereas the rest are connected hexagonally. Therefore the pixel-wise mapping of any one foveal image to the reconstructed scene image varies as a function of (φ_k, θ_k). In particular the regions of the reconstructed image that correspond to pentagons on the SES require more pixels from the associated foveal images than do the hexagonally-connected regions. Figure 26 shows the reconstructed scene image if this adjustment is not made. Note the unpopulated regions where the SES has pentagonal connectivity.

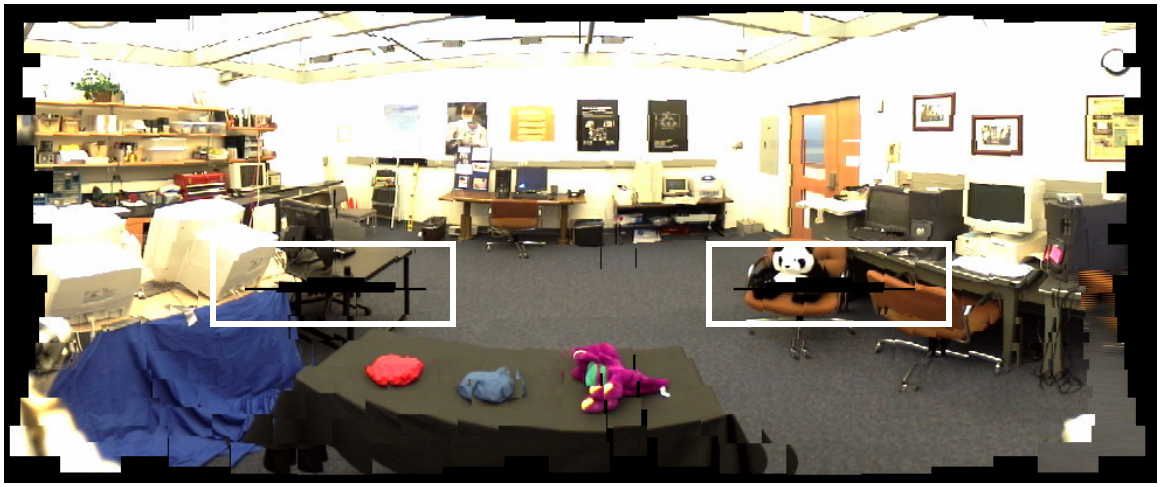


Figure 26. Scene reconstructed from SES fovea images without compensation for pentagonal regions.

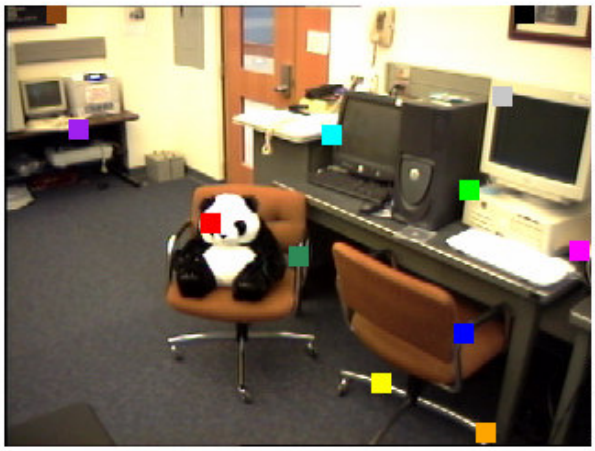
Attentional Processing on Individual Images

The problem of attention arises once the SES is populated with dense information. Because of limited computational resources, only regions of interest — determined by safety, opportunity, and by the task — can be attended to, if the robot is to interact with a human-centered environment in real time. The problem is how to perform attentional processing given a populated SES and an image input stream. How should visual attention be selected on the entire SES? There are at least two possibilities. One is to perform visual attention processing on the entire SES. The other is to detect points of interest within the individual images and combine them with the imagery that is already present. This section describes the latter approach.

Attentional processing was performed on each image in the image sequence and the results were recorded at the node corresponding to the optical center. FeatureGate, described in Chapter III, was used for this task. No top-down control was used in this part of the experiment since the addition of targets would limit the generality of the procedure. Attentional points were chosen solely on their salience. Because of the lack of top-down

processing, the model is very similar to Itti's algorithm [18]. The implementation does not process "empty" pixels (value: [0, 0, 0]) since these correspond to a lack of information, and such pixels are not treated as neighbors to non-empty pixels. Because of the computational processes in FeatureGate zero values can create spurious attentional points. For example, the color difference between an empty and a non-empty neighbor pixel would be significant. That would add to the salience of any pixel lying on the boundary of the image. For this reason, if a non-empty pixel has an empty neighbor pixel, that neighbor is not counted in the equation. FeatureGate was implemented to return the 12 most salient locations (row and column location) and their activation values in a saliency array structure. The number of locations returned by the program was set to 12 arbitrarily because it was found that this number usually results in a relatively uniform distribution of attentional points throughout the image. Moreover, there are rarely more than 12 important locations in any single image taken by ISAC's cameras. The saliency array also includes the pan and tilt angles of the image being processed.

The results of the attentional processing were then recorded in table tblActivation. Each image has 12 entries of the following form in the table: the node ID corresponding to the optical center at the given pan/tilt angles, the activation of the focus of attention (FOA) in question, and the row and column pixel location of the FOA in the image. Figure 27 illustrates this. The ranking of attentional points is in the following order: red, orange, yellow, lime green, blue, indigo, violet, magenta, black, gray, brown, and hunter green.



ID	activation	row_location	col_location
1421	5587.04969506144	119	113
1421	4344.75109759038	233	263
1421	4180.43026413515	206	206
1421	3955.76308262245	101	254
1421	3929.49435445951	71	179
1421	3927.4712481641	179	251
1421	3890.00855876116	68	41
1421	3856.84105896786	134	314
1421	3819.20628725172	137	161
1421	3648.34297607508	5	284
1421	3628.38015052536	50	272
1421	3564.51942163999	5	29

Image of the 12 most salient locations in Image000327, posted at node ID 1421

Database entries of the 12 most salient locations in image posted at node ID 1421

Figure 27. Top 12 attentional points displayed on image and recorded in database

Although only a subsection (the central foveal region) is displayed on the graphical SES representation, a full-size image is taken and processed at each node location. Because of this, there is considerable overlap between nodally-adjacent images from the sequence. The overlap means that attentional points from different images will often refer to the same location in space. In the ISAC vision system a single image spans approximately 55° in pan and 45° in tilt. Therefore, if two images are less than 55° in pan and 45° in tilt apart, they will overlap. Since we associate only a foveal window with each node, images that lie within approximately 30° in pan and 25° in tilt will overlap in the fovea. As stated before, the distance separating most nodes is approximately 5°. This yields approximately 30 images that overlap any central foveal window. The amount of overlap is a function of the location of the node in the scene. If the node is in the top left corner, for example, it has fewer images overlapping in its fovea because not many images were taken above and to the left of the node. Nodes in the center of the visual scene have maximal overlap. Overlap is also

influenced by the size of the fovea. A node with a larger fovea will have more images map to it than one with a smaller fovea. The size of the fovea depends on the length of the edges connecting it to its neighbors. Because attentional points will often (but not always) cluster on the same image feature, FeatureGate was implemented to select no more than one salient location per 15×15 neighborhood of pixels. It was desired that there be one overall attentional salience value associated with each node of the SES. Since the average fovea is 28 x 30 pixels, more than one attentional point from one original image could map to a particular fovea, depending on where the neighborhood lies with respect to the fovea. Moreover, attentional points from overlapping images will often, if not always, project onto the fovea of a particular node. To compute a single salience value for a node, the salience of all attentional points that map to the node should be combined. It was presumed that an attentional location that is identified in many images is more salient (and should, therefore, have a higher value) than an attentional location found in one image only. The process followed to combine attentional points and to identify scene locations of high salience is described below.

After attentional data is obtained from an image, each of its 12 salient points is mapped to the SES node that corresponds to its location. The correspondence is determined as follows: The distance in pixels of the image center from the attentional point is first calculated then converted into a displacement in degrees using the values (4.3) and (4.4). These values were determined experimentally: a span of 5 degrees in tilt was approximately 28 pixels and a span of 5 degrees in pan was approximately 30 pixels (see pages 48-50).

$$1 \text{ tilt degree} \approx 5.536 \text{ pixels} \qquad (4.3)$$

$$1 \text{ pan degree} \approx 6.072 \text{ pixels} \qquad (4.4)$$

Once that information is known, it is used in conjunction with the pan/tilt angle of the optical center to find each attentional point's actual pan and tilt angle which is used in turn to map the attentional point to the appropriate node. The results for the 12 attentional points associated with each of the 519 images, including the activation levels of each attentional point and its pan/tilt location, were stored in the table tblattentionalLocations. An example of this is illustrated in Figure 28. The excerpt from shows all original images (imageCenterID column) with an attentional point that maps to node 1421 (ID column) on the SES as well as each attentional point's calculated pan and tilt angles.

imageCenterID	activation	row_location	col_location	ID	new_pan	new_tilt
1302	3528.45635151368	197	146	1421	-38.769	-26.631
1626	4406.0892136347	47	212	1421	-37.660	-26.918
1624	3865.28720791141	41	140	1421	-39.610	-25.834
1421	3819.20628725172	137	161	1421	-38.602	-26.537
1682	4790.87021411756	26	236	1421	-37.308	-27.323
1340	3567.10112978288	173	134	1421	-39.200	-26.030
1424	4096.69460730313	131	233	1421	-36.692	-27.320
1679	4030.1043628122	17	116	1421	-39.962	-25.698
1501	4254.13706313962	98	236	1421	-36.789	-27.206
1303	4170.34775382598	197	173	1421	-38.141	-26.680
1733	4671.13303692383	5	266	1421	-37.252	-27.576
1681	4482.2756277841	17	197	1421	-38.315	-26.187
1376	4538.00367299199	134	11	1421	-36.106	-25.248
1263	3904.98696794347	218	107	1421	-39.967	-26.208
1379	3482.11379688946	149	125	1421	-39.578	-26.128
1564	3961.55470308158	68	158	1421	-38.870	-26.482
1498	3336.89809198234	89	143	1421	-39.126	-25.980
1341	4141.19056990857	173	161	1421	-38.672	-26.148
1500	4229.77678573337	92	206	1421	-37.650	-26.522
1627	4673.82780267508	53	248	1421	-36.879	-27.403
1680	3419.09533910259	17	155	1421	-38.956	-26.187
1304	4063.41376530996	197	200	1421	-37.513	-26.631
1560	4264.77466462499	80	5	1421	-37.175	-25.682

Figure 28. All attentional points that map to node 1421.

To determine the saliencies of the nodes, the activation (*i.e.*, the numerical saliency value) of each attention point posted at a node was summed. Figure 29 shows the top 12 overall most salient locations in the scene. Colored rectangles show perimeter of the fovea associated with each node, with order of salience according to the following colors scheme: red, orange, yellow, lime green, blue, indigo, violet, magenta, black, gray, brown, and hunter green (for the rest). For visibility, red ovals have been placed around the salient foveae.

Features selected include panda, chair, doorway, computer, black strip on far wall, printer, posters, and Barney's foot.

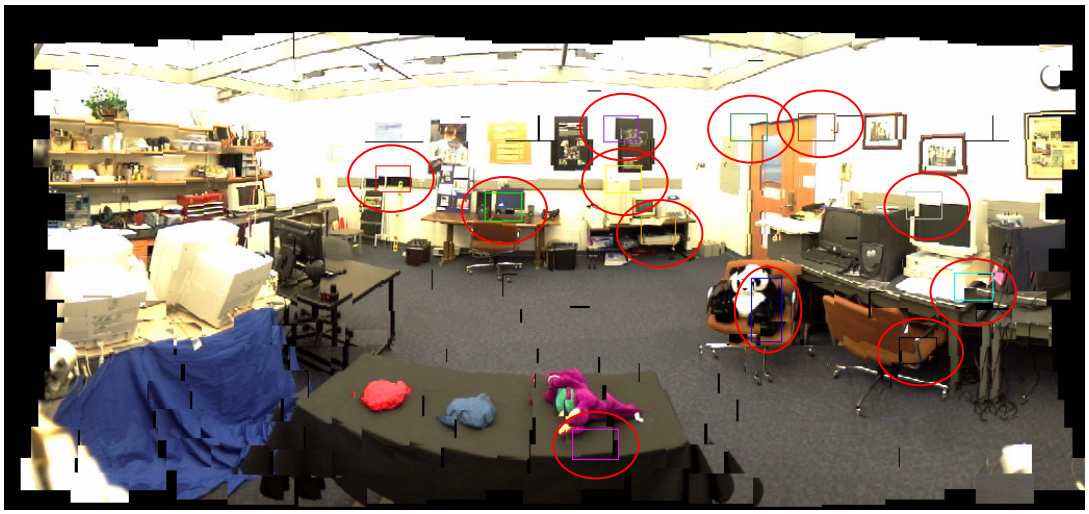


Figure 29. Top 12 most salient locations in scene by activation summation

The number of attentional points at each node was also calculated. Errors in location can cause attentional points from the same feature to be mapped to adjacent nodes. Therefore, an attentional point clustering algorithm was used to find all attentional locations that correspond to a specific environment feature. That feature, along with its collective attention value, was then assigned to a specific node. The procedure was to select each node ID with at least 15 attentional points and calculate the median pan/tilt values of the points. All attentional points in all images that fell within a radius of 2 degrees from the median pan/tilt values were then found. All these points were mapped to the same node — the node with the most attentional points that fall within the radius. A radius of 2 degrees was chosen because it represents approximately ¼ of the average fovea and is compact enough to isolate point clusters. A larger radius could consolidate different groups of points into one attentional location, thereby fusing two features into one.

Attentional Processing on Complete Visual Scene Reconstructed Images

Another way to determine attentional locations on the entire SES would be to process the image of the visual scene (reconstructed from the foveal images as described above) through FeatureGate (for example, the image in Figure 24). The FeatureGate algorithm was modified to include the node map of the reconstructed image (cf. p. 53). This makes it possible to record the node ID the attentional point is associated with for comparison with the other attentional processing technique. The results can be found in Figure 30. The salience of border pixels was decreased to avoid false attentional locations at the edges of the image.

Figure 30. Top 12 most salient locations by attentional processing on reconstructed scene image

Features deemed salient by the attentional processing include the panda, chair, black strip on the far wall, printer, as well as various corners or regions of high contrast. No objects on the front table were selected in this image. Note that no top-down processing was performed on this image; therefore, locations were selected based on neighborhood difference and not based on how closely they matched features of a known target.

Updating the SES

Although not the focus of this thesis, how to continually update images and visual attention on the SES is a problem that needs to be addressed to make full use of the work presented here. For real-time applications it will be necessary to update the information on the SES by replacing older images with new ones and by re-computing the salience values to identify new foci of attention. A related problem is how to combine attentional points from time-separate images of the same nodal area. The age of the image on the node and motion within the region should play a part in the combination. Previous information will need to be weighted by some measure of the change that has occurred at a particular node: if a previous image and a new image are completely different, the previous processing no longer applies and should be discarded.

While practical updating schemes will be implemented in future works, two experiments were performed to explore SES updating methods. In the first experiment, 11 images were taken from ISAC's left camera, starting from the upper right and moving toward the lower left. Nothing was purposefully changed in the visual scene. In the second experiment, the objects on the black table were replaced and images whose foveal windows displayed a piece of the table (33 images in all) were taken. In both experiments, the images taken at each location replaced the previous images of the same locations in the image sequence. Attentional processing and image recreation were then performed as described in the previous sections: processing of individual images followed by combination of attentional locations at each node and the reconstruction of the scene from the foveal images followed by FeatureGate. Since the first method involves processing on full images, old images could potentially have attentional points that map to nodes with new images, and

these points will be combined with new attentional points. This may be a problem in the future. The results of these experiments can be found in chapter V. Figures 31 and 32 illustrate the reconstructed scene for both experiments

Figure 31. Reconstructed scene image for experiment 1, 11 images replaced from upper right to lower left.

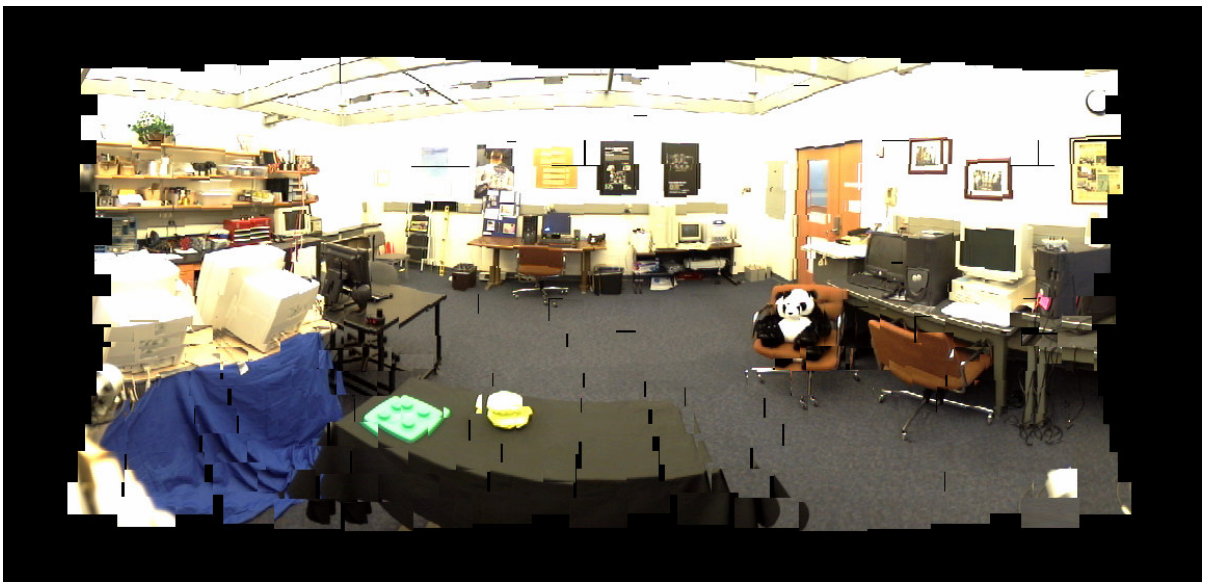


Figure 32. Reconstructed scene image for experiment 2, 33 images replaced making up the black table.

CHAPTER V

RESULTS

SES Population

As can be observed from figure 25 (reproduced here in figure 33), the reconstructed scene was almost entirely covered with imagery. There is, however, some overlap between certain adjacent foveae (especially at the bottom of the image) as well as areas where imagery is missing between foveae. Image fovea size was selected to cover only the area associated with the respective node, and the pixels-per-degree measure (cf. p. 49) was used to estimate that size.

Figure 33. Reconstructed scene from SES fovea images.

Larger foveae could have been used in image reconstruction. Figure 34 illustrates the reconstructed scene with a fovea size that covers a node and extends to its neighboring nodes. This method was not selected because it caused more distortion and overlap than the first

image, especially in the bottom of the image. Moreover, the overlap makes it difficult to create an accurate node map of the reconstructed image. As fovea images are placed on the reconstructed scene, each pixel occupied is marked with the node ID in the node map. Overlap causes nodes entered first to be partially erased by the nodes entered subsequently, which yields inaccurate fovea sizes when displaying attentional points.

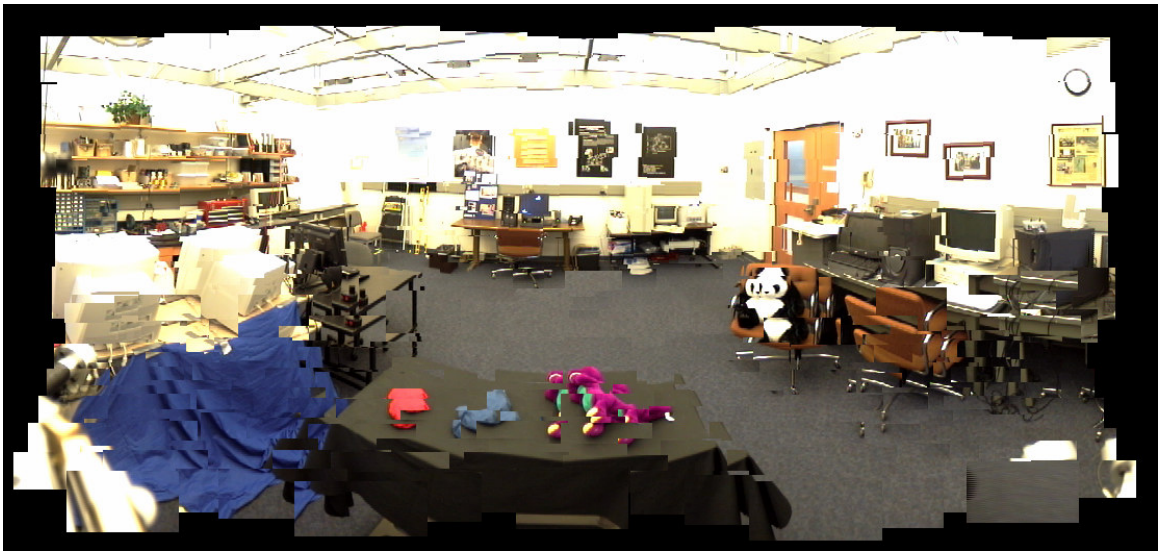


Figure 34. Reconstructed scene from larger fovea images.

Clustering

The significance of a single attentional point in an image with respect to the entire scene on the SES had to be determined to examine the relationship between a single point and the scene. Is the point only found in one image or in many images overlapping on a particular node? Each point was chosen because of its high salience value, which in this context is a measure of contrast or neighborhood difference. Does the attentional point correspond to an important feature in the scene or was it only significant in its neighborhood in the image from which it came? To accomplish this, clusters of attentional points first had

to be found (cf. p. 59). Each cluster was then mapped to a single node ID in order to prevent attentional points referring to the same feature to be registered at different nodes. The threshold of 15 was determined to be a good value from the graphs in figures 35 and 36. It also corresponds to a location being chosen in approximately half of the images that overlap on a fovea (provided only 1 attentional point from each image is mapped to the fovea) since there are generally 30 images that overlap on one node. If more than one point is chosen from an individual image and the number of attentional point at a node exceeds 15, then several points being chosen from a couple of images would still make this location significant. This is based on the assumption that if attentional points mapping to a node show up in many images, it is more than likely a real feature in the scene.

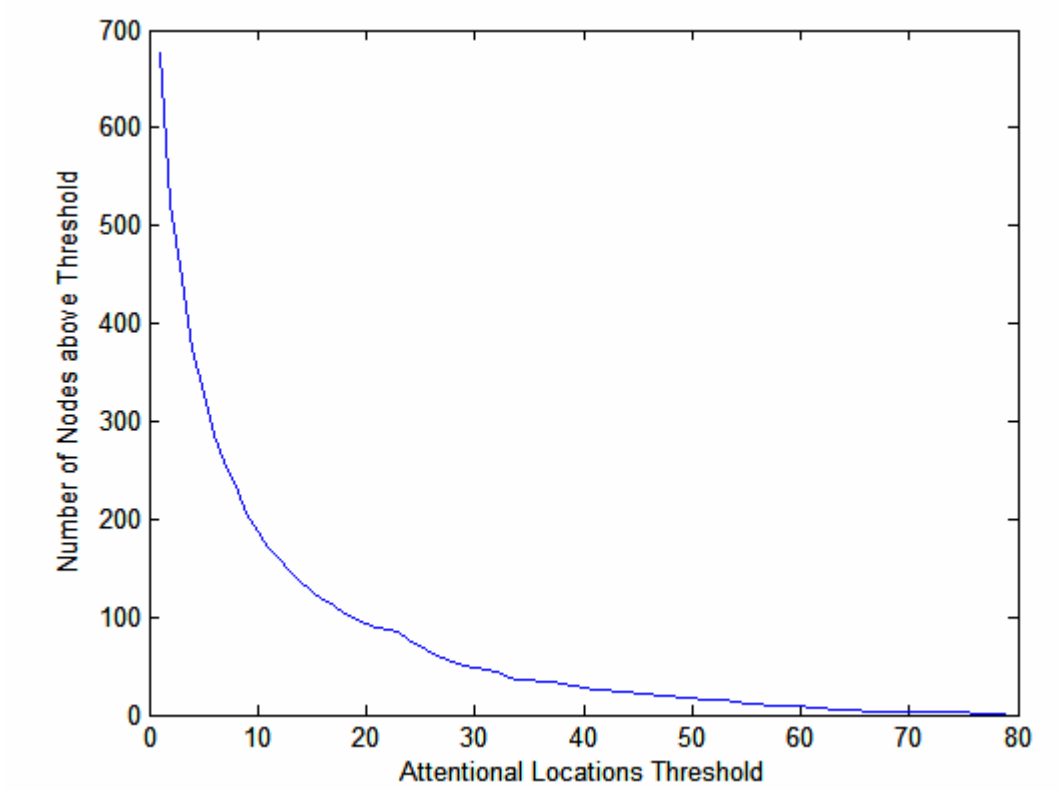


Figure 35. Graph of the number of nodes with more attentional locations than a specific threshold

There were 118 nodes (out of 672) with 15 or more points. Figure 36 illustrates the number of attentional points at each node.

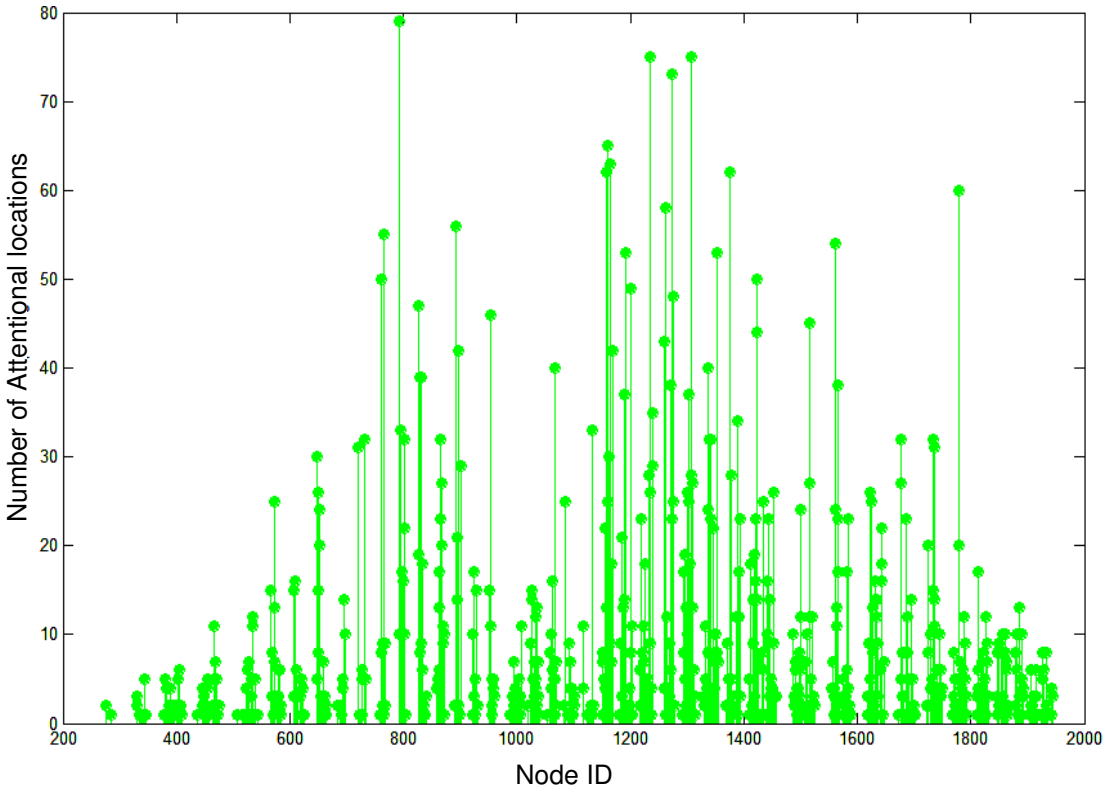


Figure 36. Number of attentional points per node

Once the clusters were processed and attentional points representing the same location in space were assigned to the same node, the data was used to create several graphs. Figure 37 is a graph of the new graph representing the number of attentional points per node.

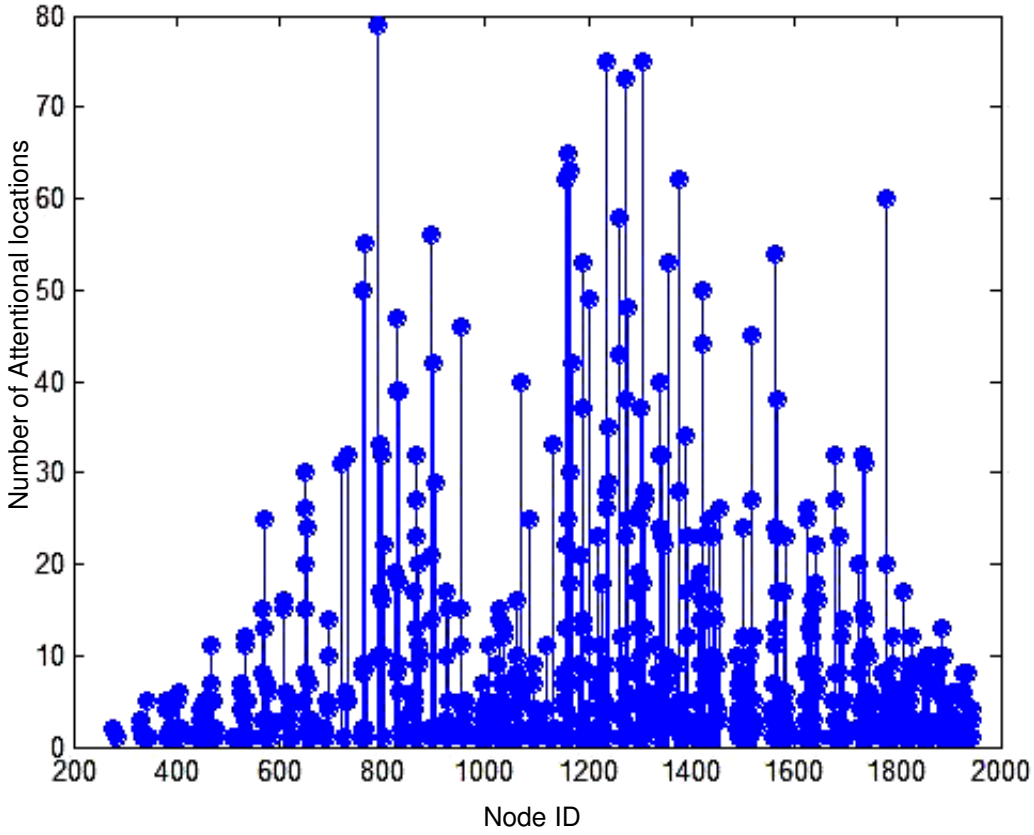


Figure 37. Number of attentional points per node after cluster processing

Since the difference between figures 36 and 37 is not obvious, a graph of the difference between these two can be found in figure 38. Because of clustering, points were removed from some nodes and assigned to other nodes; this caused the first nodes to have a negative difference and accounts for the negative values on the graph.

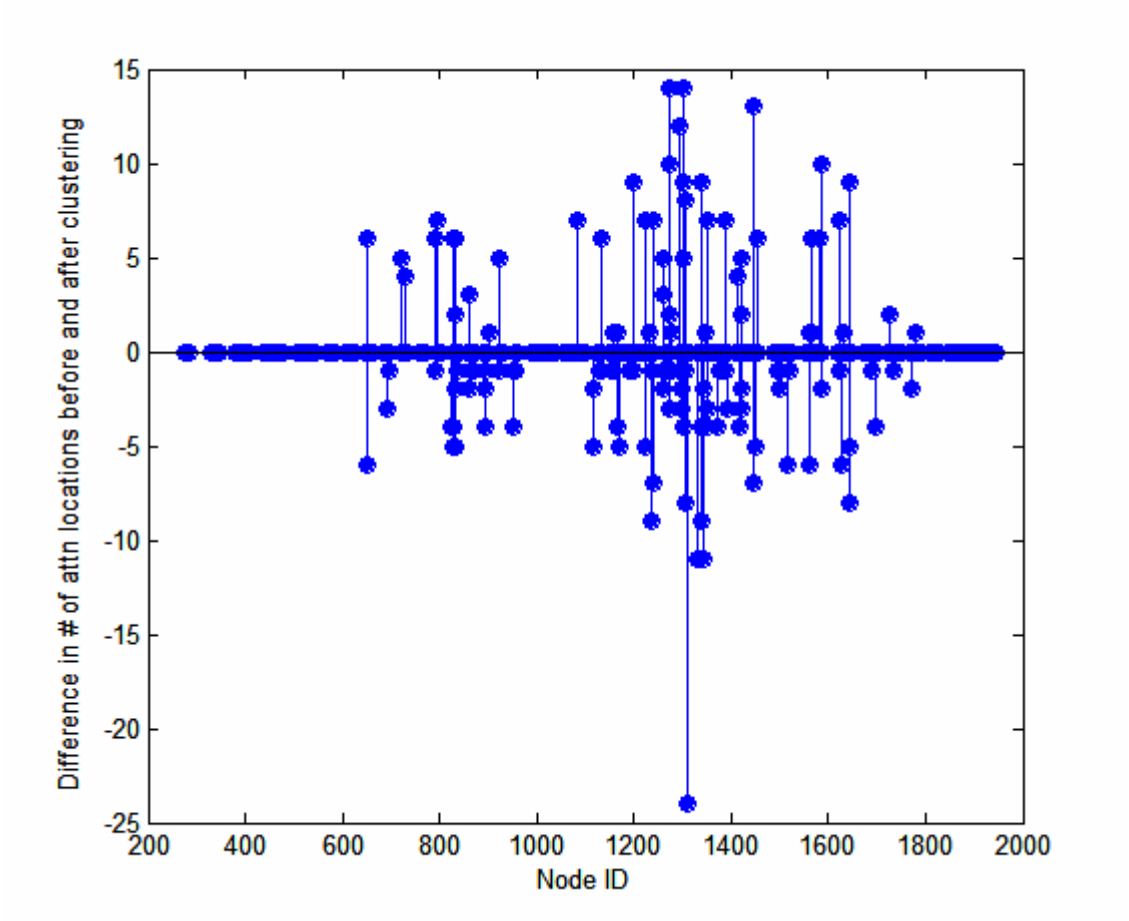


Figure 38. Difference in the number of attentional locations per node before and after clustering

Activation summation and averaging

Graphs for activation threshold versus number of nodes and activation per node were also generated and can be found in figures 39 and 40. The FeatureGate activation output for a single salient location ranged from 1027.9671 to 6354.7972 in this experiment. The minimum and maximum summed activation values per node were 1141.3 and 3.6706 x 10^5 respectively.

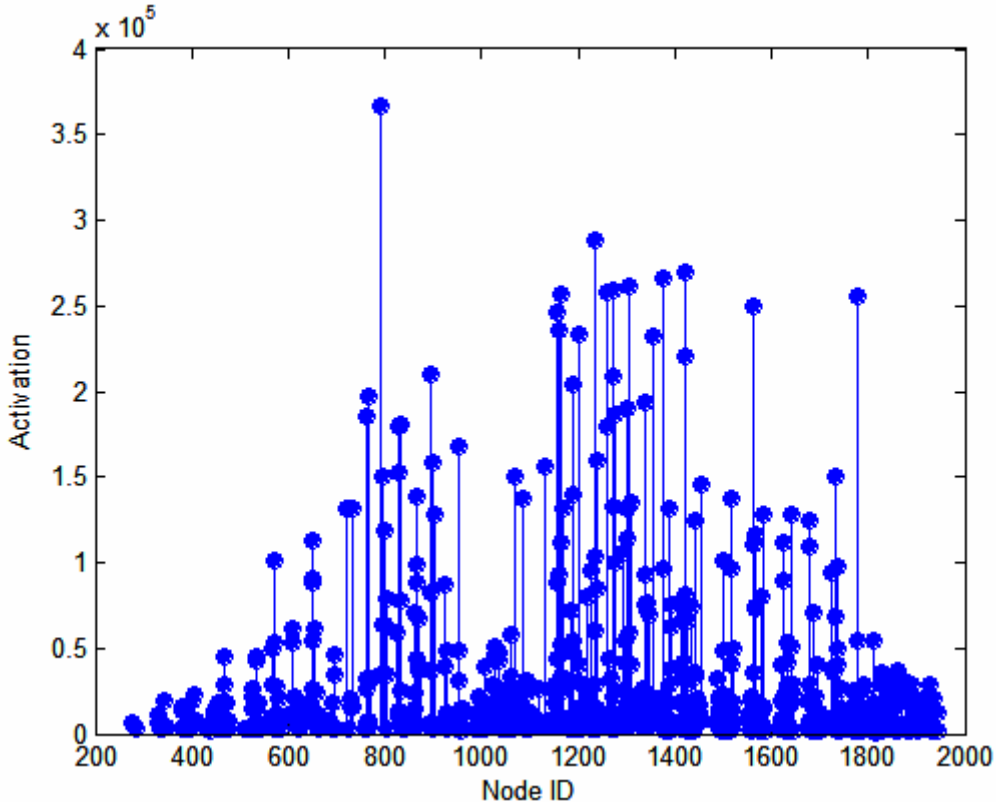


Figure 39. Activation per node ID

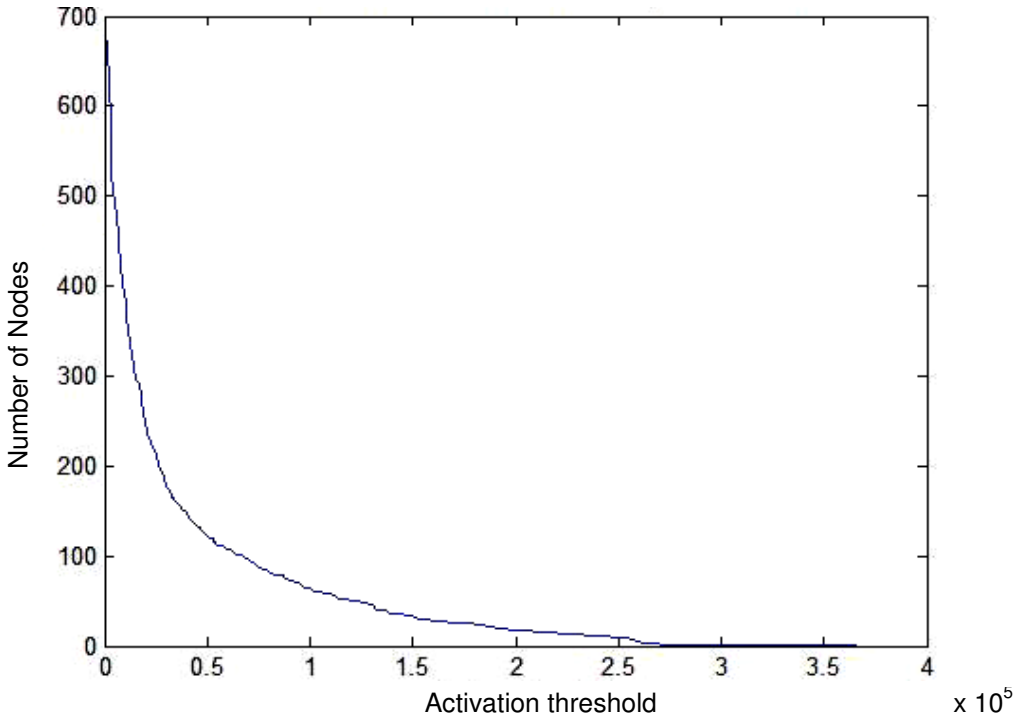


Figure 40. Graph of the number of nodes above a specific threshold

Several thresholds were chosen and the percentage of nodes with activation above threshold level was computed. Table 5 lists these results. These results give a measure of the activation level necessary for a node to be a significant attentional location on the entire SES. For example, to be in the top 10% of attentional locations on the SES, a node would have to have a summed activation value of al least 100000.

Table 5. Activation thresholds and percentage of nodes above thresholds

Activation Threshold	Number of Nodes above Threshold	Percentage of Nodes above Threshold
27000	201	30%
40000	148	22%
45000	134	20%
50000	123	18.5%
100000	64	9.5%

Another way of determining how important a single attentional location is to the overall salience of the SES is to calculate the percentage of individual attentional locations that map to a node with above-threshold activation. There are 6228 total attentional locations (12 points per image x 519 images) made to the SES. These calculations were performed for several thresholds. For example, if the nodes with activation values in the top 10% are chosen (threshold of 100000), the percentage of individual attentional locations that map to one of these nodes is 41%. In other words, 41% of individual attentional locations map to the top 10% node locations on the SES. The percentage calculations for different thresholds can be found in Table 6.

Table 6. Percentage of individual attentional locations above threshold

Percentage of Nodes Above Threshold Chosen	Threshold	Percentage of Individual attention locations at Nodes Above Threshold
Top 10%	100000	41%
Top 20%	45000	65.3%
Top 30%	27000	77%

Another measure of the importance of individual attentional locations on the overall SES salience can be examined by finding the percentage of attentional locations in the top N locations (nodes). This is very similar to the percentage comparison above except that a fixed number of nodes are chosen as opposed to a percentage of nodes. Choosing a fixed number of salient locations can be useful for comparisons. Moreover, no matter how many attentional locations are found in a scene, only a fixed number can and should be attended. For example, 19% of individual attentional locations were found to map to the top 20 node locations on the SES. In order words, the 20 most salient locations on the sphere represent 19% of all individual attentional locations. Table 7 shows the number of attentional locations for several values of N.

Table 7. Percentage of individual attentional locations in top N most salient node locations

N	Percentage of attentional locations in top N node locations
20	19%
30	25.8%
50	36.2%

Because of their location in the visual scene, certain nodes have a higher possibility of attentional locations since more images overlap on their fovea. These locations could have higher activations than nodes lying closer to the edges of the visual scene and therefore have an unfair advantage in the attention competition. To test the effects of this phenomenon on attention, the average activation value at each node was computed and used to determine the most salient locations in the scene. A graph (similar to figure 40) representing the

number of nodes with average activation higher than a specified threshold can be found in figure 41. Results illustrated in tables 5, 6, and 7 were also generated using the average activation of each node (as opposed to the sum of all activations at each node) and can be found in tables 8, 9, and 10. Average activation values per node ranged from 1141 to 5298. As can be seen from the graph in figure 41, the most significant change in average activation values occurs between values of 3000 and 4000.

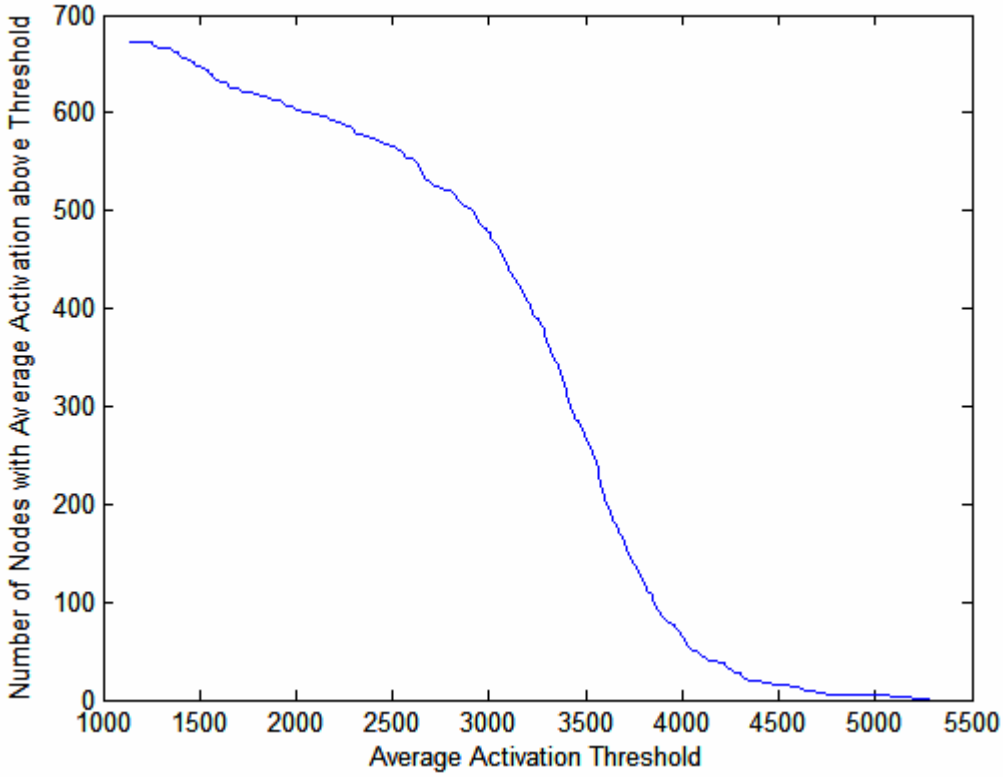


Figure 41. Graph of the number of nodes with average activation above a specific threshold

Table 8. Average activation thresholds and percentage of nodes above thresholds

Average Activation Threshold	Number of Nodes above Threshold	Percentage of Nodes above Threshold
2000	603	90%
3000	477	71%
3615	199	30%
3760	137	20%
4000	65	9.7%
5000	5	0.7%

Table 9. Percentage of individual attentional locations above activation average threshold

Percentage of Nodes above Threshold	Threshold	Percentage of Individual attention locations at Nodes above Threshold
Top 10%	4000	21.2%
Top 20%	3760	42.3%
Top 30%	3615	56.6%

Table 10. Percentage of individual attentional locations in top N most salient node locations

N	Percentage of attentional locations in top N node locations
20	5.3%
30	9.8%
50	14.4%

By comparing tables 6 and 7 to tables 9 and 10, it can be observed that taking the sum of all activations at a node instead of the average to determine the most salient locations in

the scene accounts for more individual attentional locations. For example, the top 30 most salient nodes on the SES identified with the summation method account for 25.8% of individual attentional locations while the top 30 nodes identified with the averaging method only account for 9.8% of individual attentional locations.

Individual Images versus Reconstructed Scene Image

Attentional points found in individual images were compared to attentional points found in the reconstructed scene image. This was done by finding the N nodes with the highest summed (or averaged) activation of all of the attentional locations mapping to a node. A reconstructed scene image (figure 25) was also run through the FeatureGate program as a single image to find the N nodes with highest activation. When attentional processing is performed on full-size individual images, some attentional locations get mapped to nodes that do not correspond to an image piece posted in table tblSES. This occurs in images taken at nodes lying near the edges of the visual scene. For example, when an image taken at a node in the upper left corner of the visual scene is processed in FeatureGate, attentional locations could be identified above and to the left of the foveal window. These locations, when mapped to the nearest node, will be posted at nodes outside of the visual scene reconstructed from the foveae. These locations are not represented in the reconstructed visual scene image and it wouldn't be accurate to compare them to nodes in the reconstructed image; for this reason, the top N locations that correspond to a node in the reconstructed scene image are found. There are 519 such nodes, compared to 672 nodes with an attentional point mapped to it from a full-size image.

The attentional locations found through summation and averaging of the activation values were then compared to the locations found by processing the reconstructed scene image directly. The results can be found in table 11. As can be observed, the attentional locations found when summing the activation values at each node matched the attentional locations found when processing the reconstructed image more closely than did the locations found when taking the average activation value at each node. The combination of these results and the results in tables 8-10 suggests that activation values should be summed rather than averaged.

Table 11. Matching attentional nodes between individual image summing and averaging and reconstructed scene image

N	Number / Percentage of matching nodes by summing activation at each node	Number / Percentage of matching nodes by averaging activation at each node
12	5 / 42%	2 / 17%
20	8 / 40%	5 / 25%
30	13 / 43%	6 / 20%
50	21 / 42%	17 / 34%
100	59 / 59%	55 / 55%

Figures 42, 43, and 44 show the top 20 attentional node locations in the summed activation values image, averaged activation values image, and reconstructed visual scene processed image respectively.

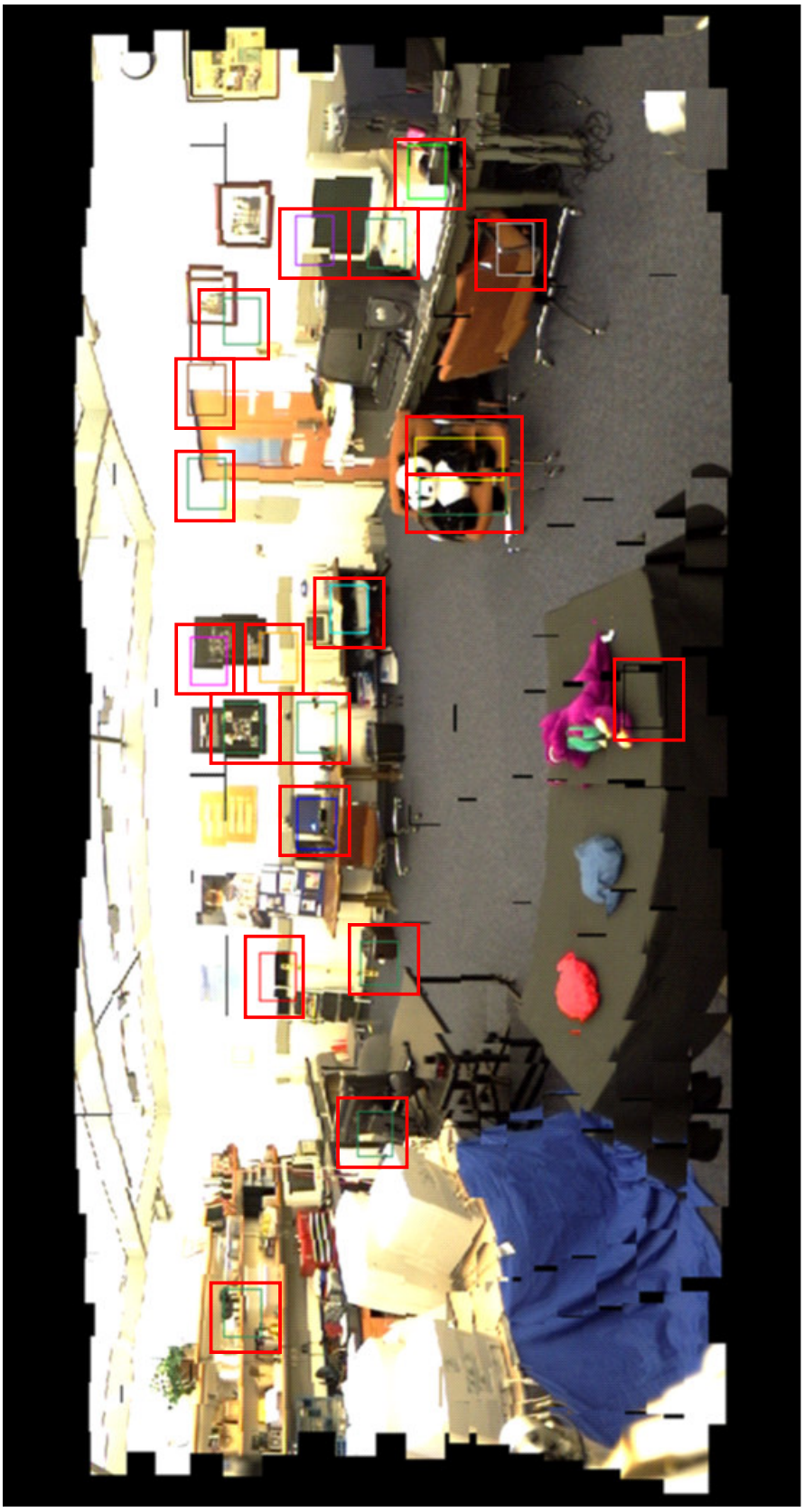


Figure 42. Top 20 attentional locations in summed activations image

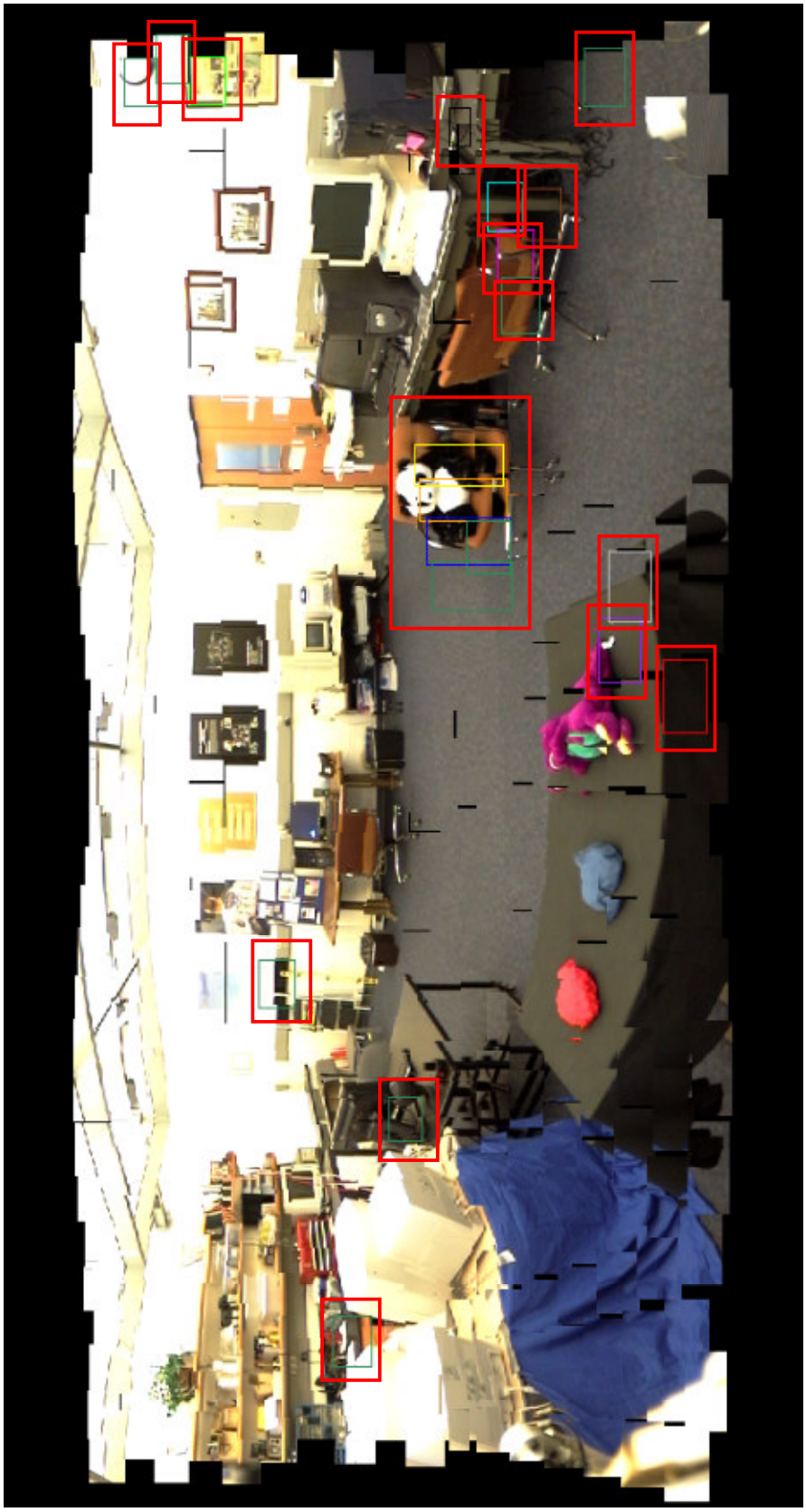


Figure 43. Top 20 attentional locations in averaged activations image

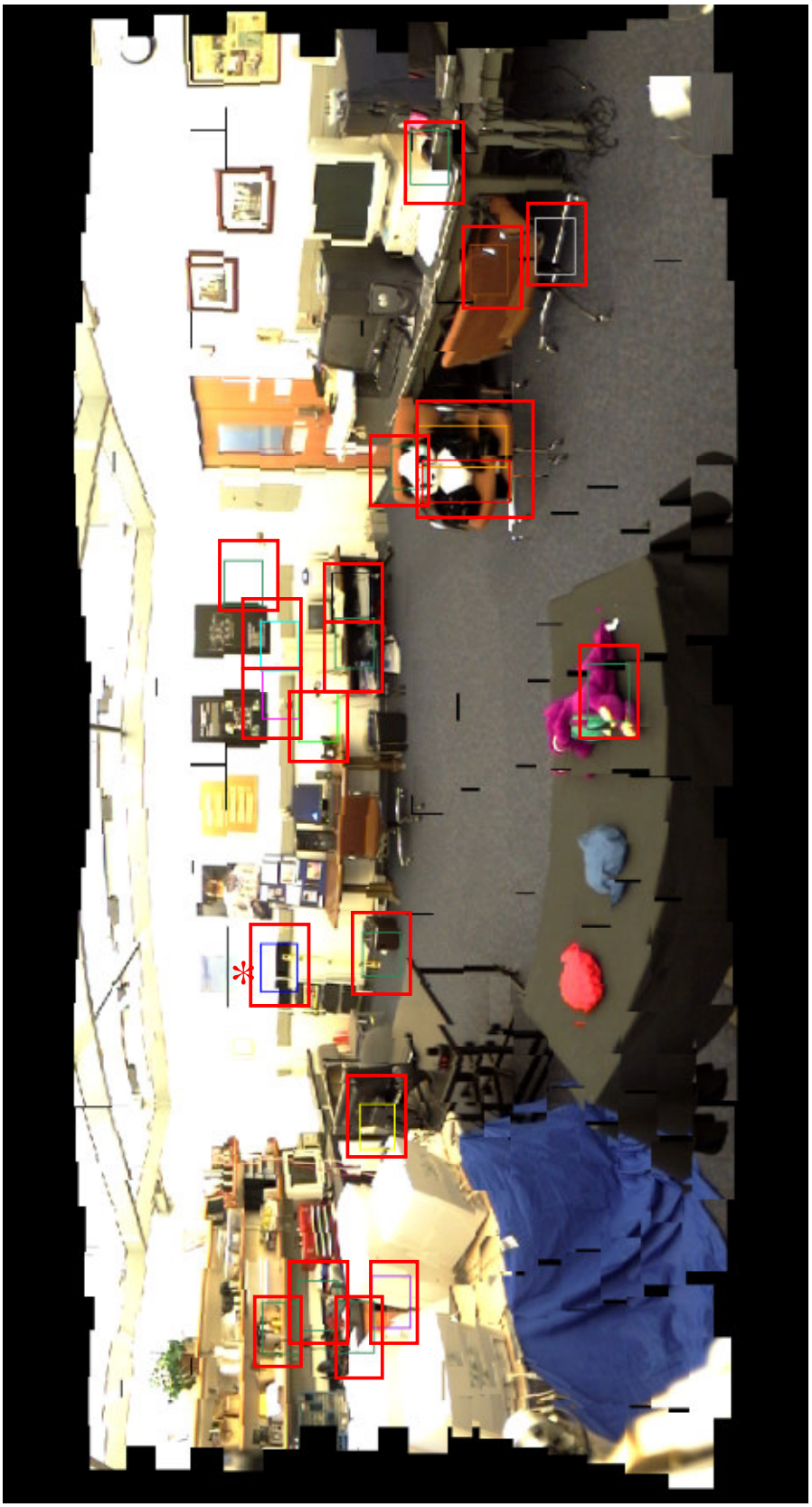


Figure 44. Top 20 attentional locations in reconstructed visual scene image

Salient features, such as the panda, Barney doll, trash can, left side shelves, and chair were detected in both the summed activation image and the reconstructed scene image. They were also detected (except for the trash can and shelves) in the averaged activation image. Features with definite edges and corners, such as the black frames on the front wall and the black wall-strip (marked with an asterisk in figure 41) were also detected in all images. The clock and frame, however, on the right wall were only identified in the averaged image, while the upper corners of the door were only identified in the summed activation image.

Updating Imagery on the SES

As described in chapter IV, two updating experiments were conducted. Experiment 1 involved a replacing swipe of images (11 in total) from the upper right corner to the lower left corner without purposefully changing anything in the room. Experiment 2 involved changing the objects on the black table and replacing the 33 images with foveae of a section of the table. The reconstructed scene images were processed with FeatureGate and each individual image was processed as well. Table 12 compares the results of attentional processing on the original scene image and both experiment updates. Table 13 compares the results between the summed activation of individual attentional locations at each node in the original locations to both experiments and table 14 does the same for the activation averaging method. As can be seen from the data in these tables, updating imagery on the SES did not greatly affect the most salient locations. In the future, imagery will need to be weighted by age and by how much different it is from the previous imagery of the same location (motion of the area).

Table 12. Comparison between original reconstructed scene image and updated reconstructed scene images

N	Matching Nodes in Original and Update Experiment 1	Matching Nodes in Original and Update Experiment 2
12	11	12
20	18	19
30	25	28
50	45	47

Table 13. Comparison between original summed activation image and updated summed activation images

N	Matching Nodes in Original and Update Experiment 1	Matching Nodes in Original and Update Experiment 2
12	12	11
20	18	17
30	26	26
50	43	42

Table 14. Comparison between original averaged activation image and updated averaged activation images.

N	Matching Nodes in Original and Update Experiment 1	Matching Nodes in Original and Update Experiment 2
12	12	11
20	18	17
30	27	24
50	47	43

Figures 45 and 46 show the top 20 most salient locations when the reconstructed image is processed while figures 47 and 48 show the top 20 most salient locations in the

summed activation images and figures 49 and 50 show the top 20 most salient locations in the averaged activation images. As can be seen, these images are very similar to figures 42, 43 and 44 respectively. Most of the same salient locations were chosen, although not always in the same order as in the original. This shows that the updating has changed the salience of the SES. To be efficient, however, the future updating scheme should identify changes such as changing the objects on the black table (experiment 2). The yellow bean bag was identified as the 7th most salient location in the averaged activation image but not in either the summed activation or reconstructed scene images. The averaged activation processing should be re-examined when designing the SES update implementation. The difference between new and previous imagery should also be examined.

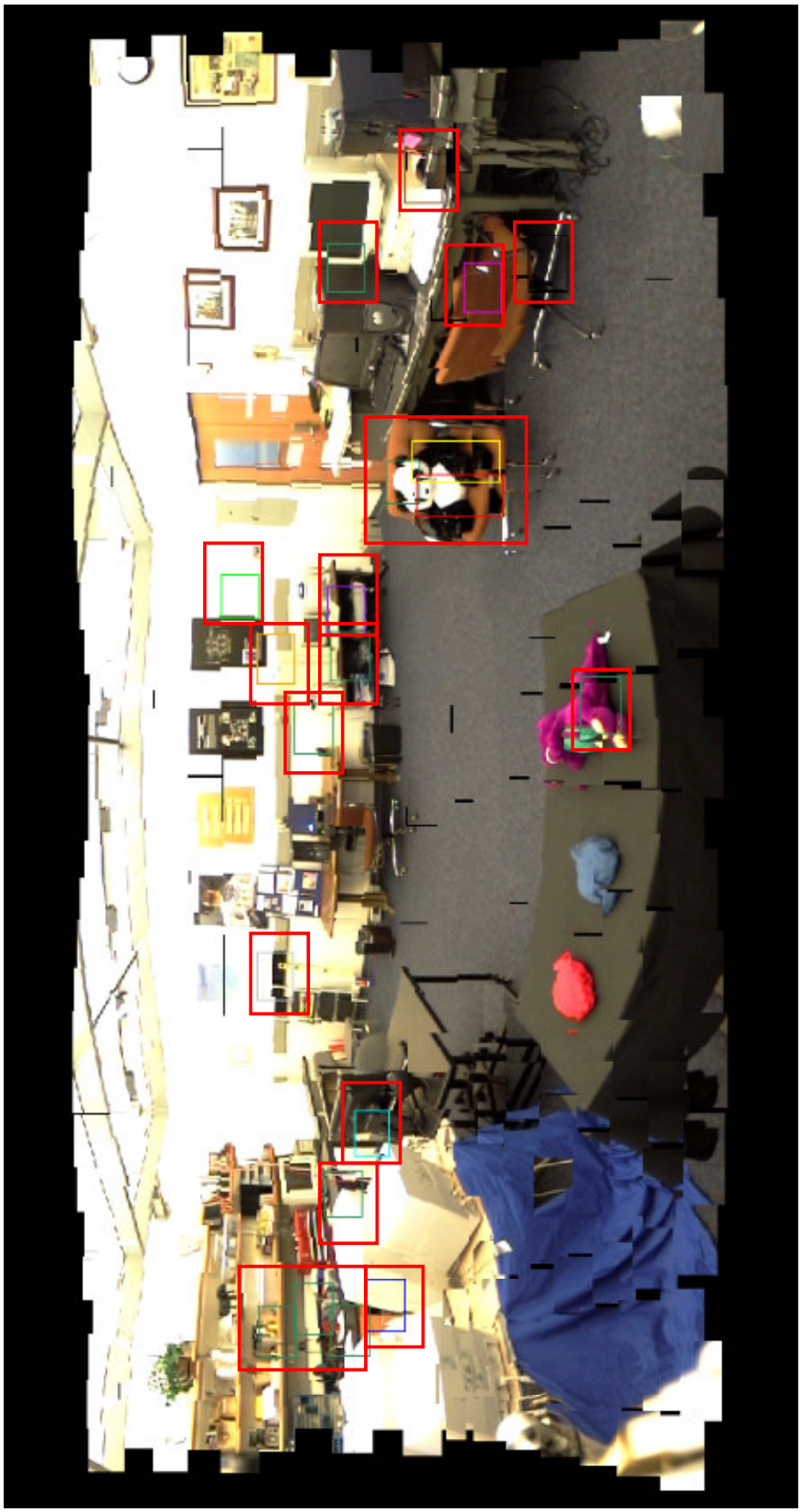


Figure 45. Top 20 locations in reconstructed visual scene for update experiment 1

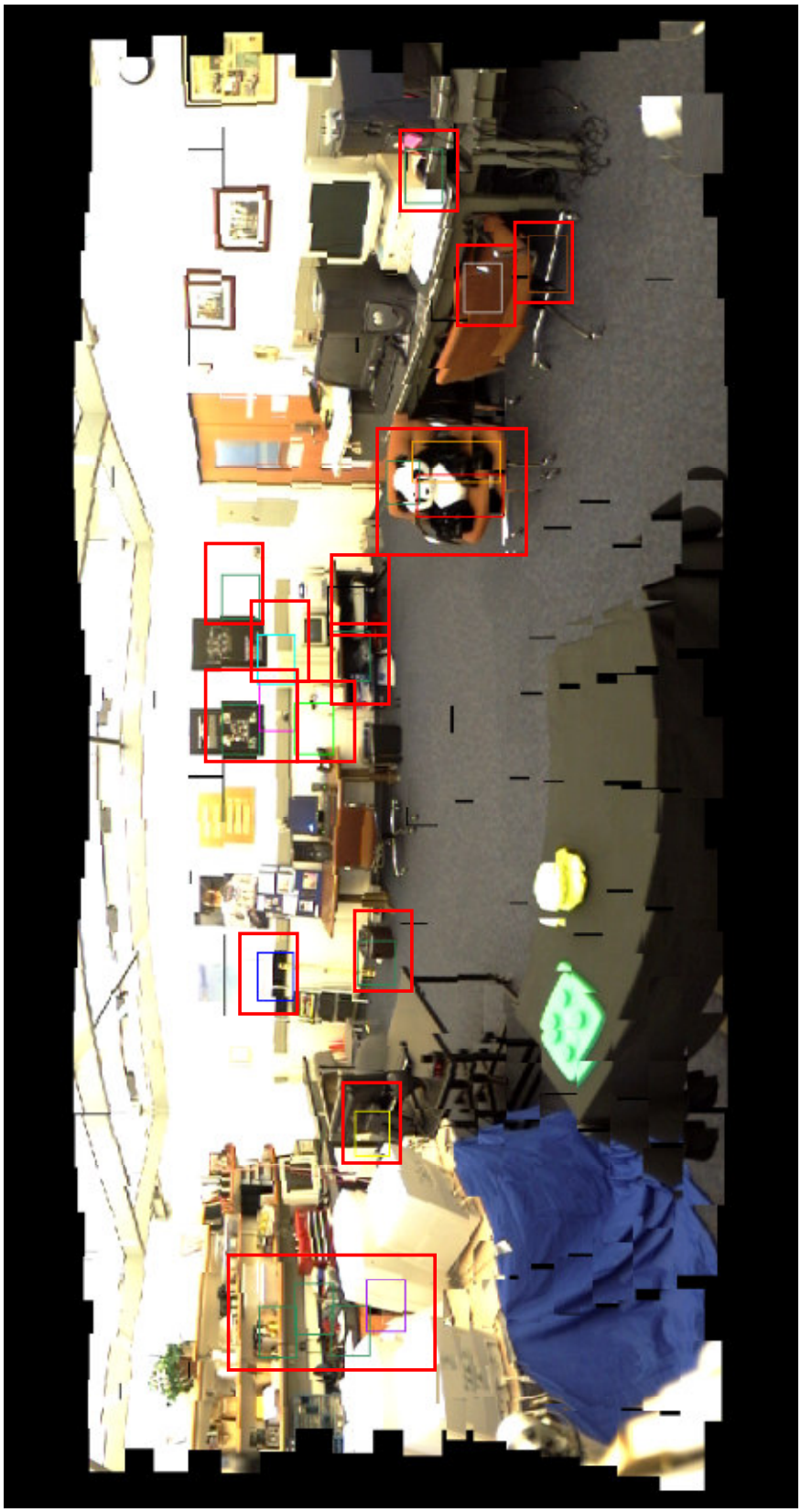


Figure 46. Top 20 locations in reconstructed visual scene for update experiment 2

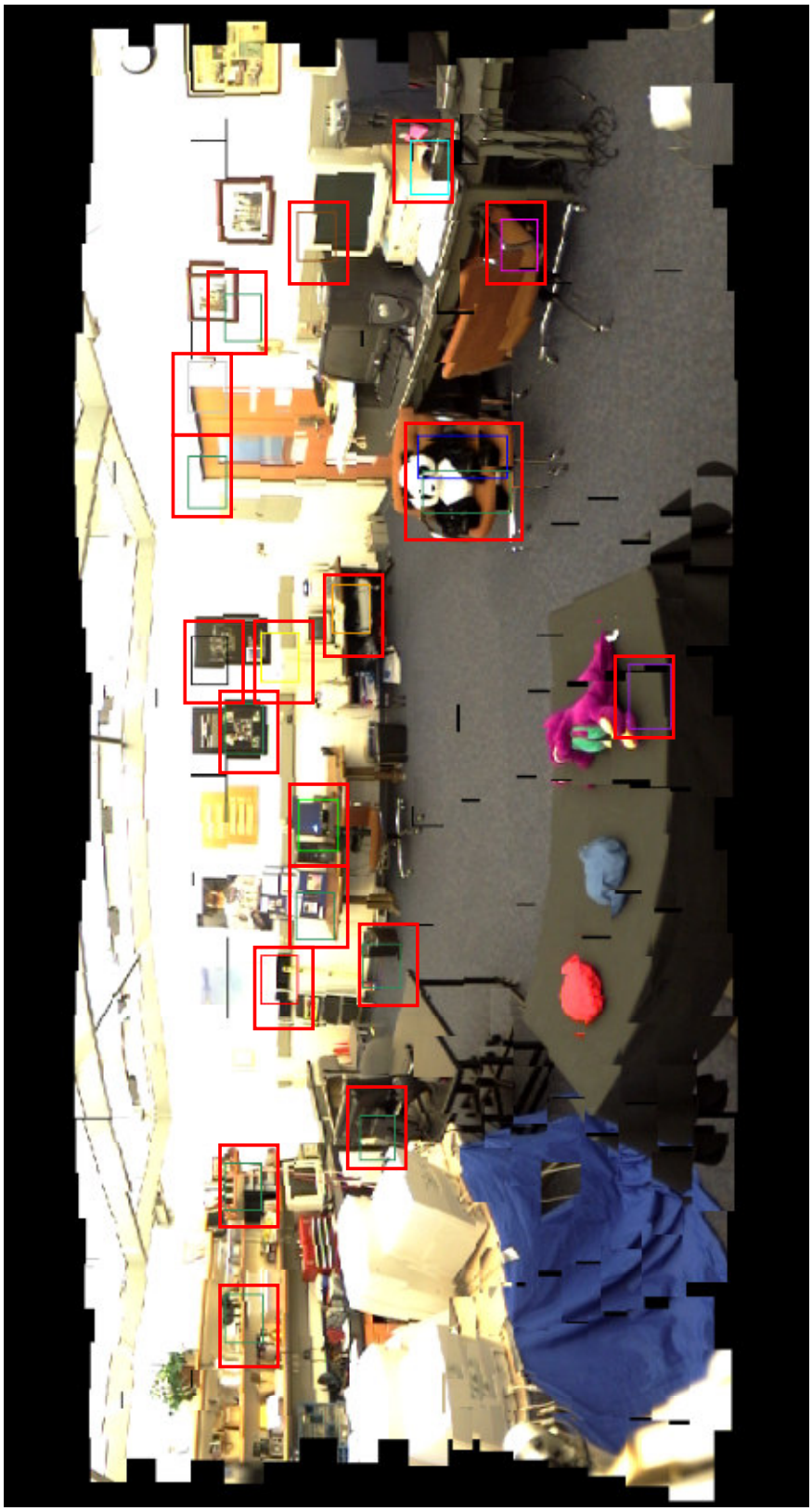


Figure 47. Top 20 locations in summed activation image for update experiment 1

Figure 48. Top 20 locations in summed activation image for update experiment 2

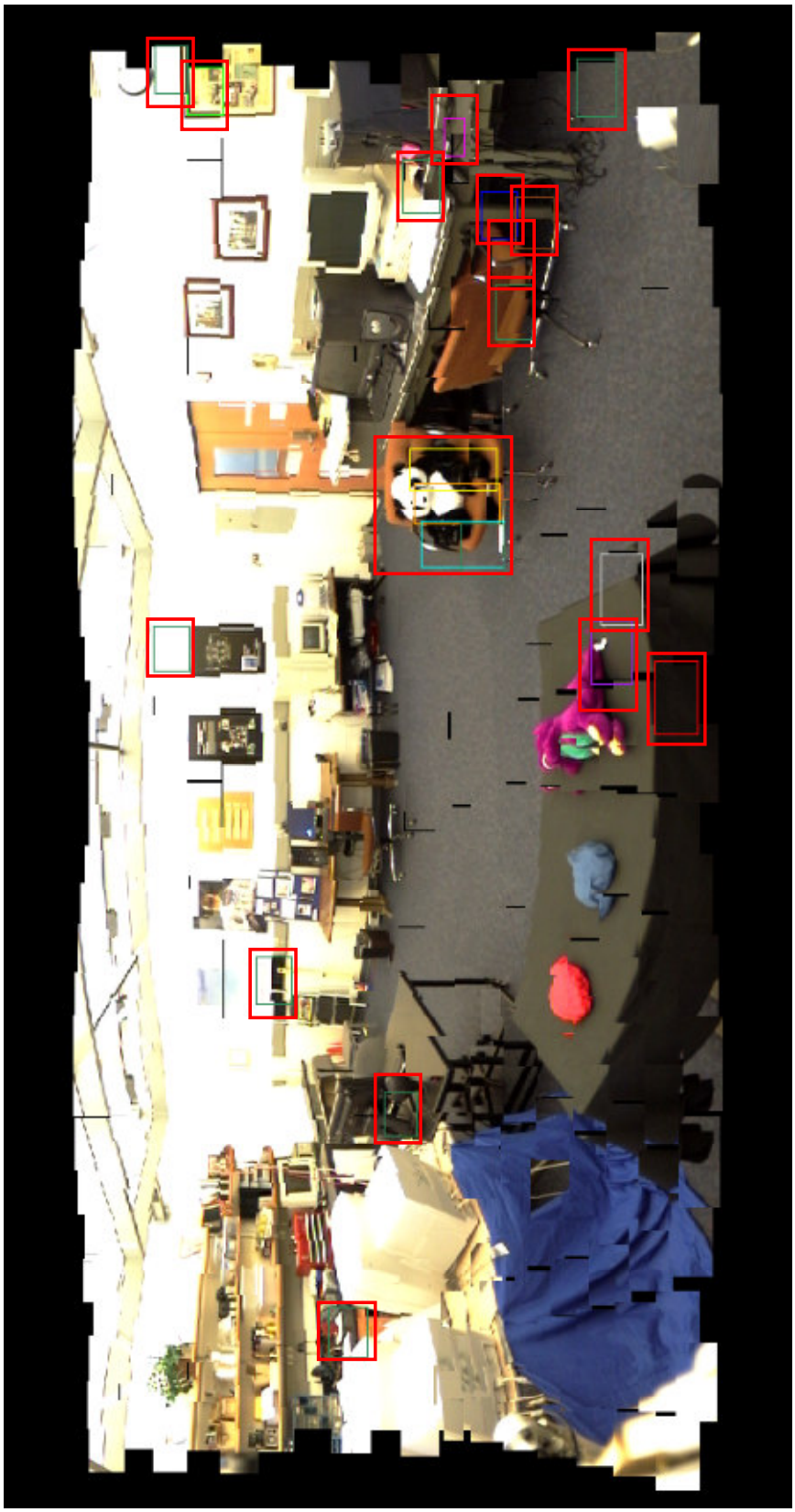


Figure 49. Top 20 locations in averaged activation image in update experiment 1

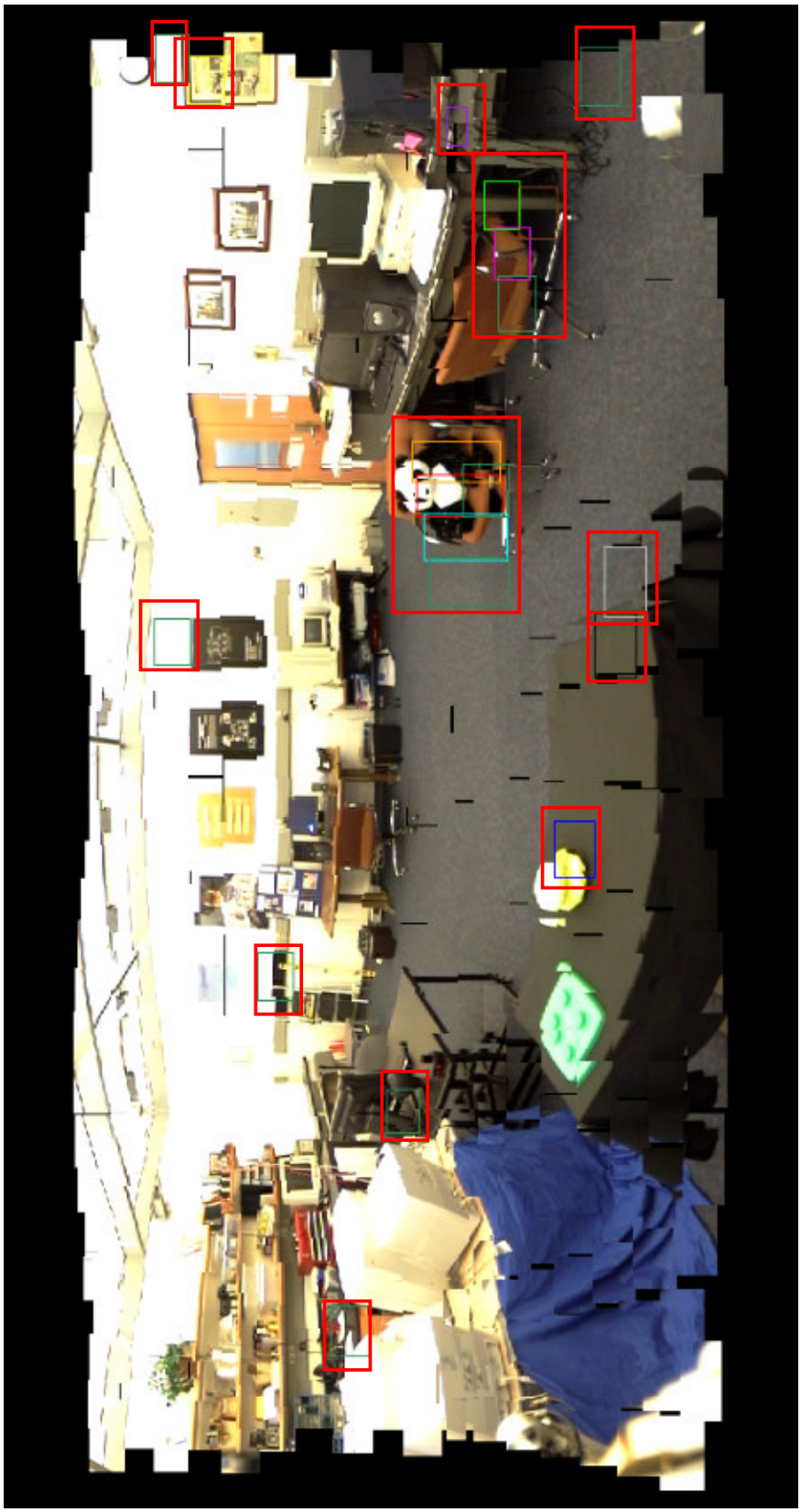


Figure 50. Top 20 locations in averaged activation image in update experiment 2

CHAPTER VI

CONCLUSIONS AND FUTURE WORK

Conclusions

This thesis has presented the procedures to map high-resolution imagery on the SES using an image sequence and to perform visual attention processing on the SES. The image sequence was taken from the humanoid robot ISAC's rotating camera head. An image was taken at each SES node falling inside a pre-determined area. The selected area provided a good representation of the robot's environment. Problems such as overlap between adjacent images of the image sequence and variable distance between nodes were addressed to obtain a continuous mapping of the robot's visual scene. Although a full image was captured at each node location, a foveal window was extracted from the center of the image. The foveal windows were then used to populate the SES and reconstruct the visual scene with minimal overlap. Because of the variable distance between adjacent nodes on the SES, the size of each foveal window was determined based on the distance in pan/tilt degrees of each node to its 4 nearest neighbors. This distance was then converted to a distance in pixels using a pixel-per-degree measure determined experimentally.

A mechanism for attention is necessary if the SES is populated with dense imagery. Because of limited computational resources, only regions of interest can be attended if a robot is to interact with a human-centered environment in real-time. Two possibilities of selecting visual attention on the SES were examined in an attempt to address the problem of how attentional processing should be achieved. Both methods used the FeatureGate model of visual attention [2, 9].

The first method involved performing attentional processing on individual full-size images from the image sequence to identify the most salient locations. Because of overlap present between images in the sequence, attentional points found in different images could refer to the same location in space. For this reason, the salient locations found in each image were then associated with the node closest to their location on the SES (instead of the node corresponding to the optical center of the full-size image). To eliminate possible errors in location causing attentional points from the same feature to be mapped to adjacent nodes, nodes with clusters of 15 or more attentional points were identified. A threshold of 15 was chosen both from a graph and because approximately 30 images overlap on any given fovea and a threshold of 15 signifies that a feature was found salient in half of the images. The median pan/tilt locations of each cluster was calculated and all attentional points falling within a 2 degree radius of the cluster's median were remapped to a single node. Attentional points were then summed at each node to find the most salient node locations in the entire visual scene. Based on the assumption that the more often a location is selected in separate images, the more likely it is that there is an actual relevant feature at that location, an attentional point that has persisted in several adjacent images will have a higher activation value and, therefore, will be deemed more salient than an attentional point found in only one image. Updating the salience of the SES as new images become available could easily be done with this method by processing new images and combining the new attentional points found with the attentional points already present. The activation at each node could be weighed by the age of each attentional point, giving more weight to newer points.

The second method of selecting attention on the SES involved performing attentional processing on the image reconstructed from the foveal windows posted on the SES. Less

information was available in this method since only one image determined the most salient locations in the scene as opposed to a sequence of overlapping images. Whether an attentional point has persisted in several adjacent images is not known. Therefore, the confidence level that a location deemed salient by this processing method is an actual salient feature in the environment is less than with the first processing method. Moreover, updating the salience of the SES as new images are made available is not easily done with this method without having to reconstruct the entire scene.

Future Work

There are several avenues of research to be explored for a practical implementation of this work. The most evident extension is to take advantage of the binocular vision system implemented on ISAC [36]. Image sequences in this work were taken from ISAC's left camera only. Eventually, images from both cameras could be taken and the corresponding stereopsis image derived from these images and used to populate the SES. The remainder of the future work to be completed can be divided into two areas; the first deals with making enhancement to the system so that it can be used in real time. The second area involves stabilizing attentional points and adding control so that the system returns areas task-relevant.

The amount of time necessary to obtain an image sequence and run each image through the attentional process is quite large and decreases the usefulness of this work. This system will ultimately be used on a robot interacting with a human-centered environment in real time. One option to speed up the process would be to set aside setup time to do the image sequence creation and population of the SES. Once this is done, images taken from the camera can be processed one at a time, in real time, and added to the SES at the correct

locations. A new image posted at a particular node could be compared to the image previously posted at the same node; differences between the images would indicate that the scene has changed in the particular location and salience could be assigned to the node based on the amount of change.

Another option would be to down-sample the number of images in the image sequence by a factor determined experimentally. Perhaps it would be the case that an image taken at nodes corresponding to the center of hexagonal or pentagonal regions would be enough to recreate the visual scene. The overlap between images taken at adjacent nodes is large enough to reduce the number of images in the image sequence significantly.

To speed up the attentional processing, the FeatureGate implementation, currently in Matlab, could be implemented in C++. The code could also be revised to maximize runtime speed and to run in real time.

Robustness experiments should be performed to test the stability of attentional points founds under different illumination levels. If points persist in visual scenes with different lighting conditions, then attention can be directed more robustly at areas of interest in the scene. The dense imagery sensory information would lead to more reliable results and this is likely necessary for grounding the robot in its environment. Incorporating top-down processing into the FeatureGate implementation and using it to guide attention to areas that are not only salient but also relevant to the task at hand will also lead to more useful results.

Once these improvements are made to the existing system, other sensory modalities such as sound, touch, and proprioception could be added to the SES. These dense sensory information types, in combination with the high-resolution imagery, could be used to guide a robot's attention.

BIBLIOGRAPHY

1. Brefczynski, J.A., DeYoe, E.A. A physiological correlate of the 'spotlight' of visual attention. *Nature neuroscience*, 2, 370-374 (1999).

2. Cave, K.R. The FeatureGate model of visual selection. *Psychological Research*, 62, 182-194 (1999).

3. Chawla, D., Rees, J., Friston, K.J. The physiological basis of attentional modulation in visual extrastriate areas. *Nature neuroscience*, 2, 671-676 (1999).

4. Clifton, C. SESDisplay. Vanderbilt University, Nashville, TN.

5. Corbetta, M. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proc. Natl Acad. Sci. USA*, 95, 831-838 (1998).

6. Corbetta, M. Positron emission tomography as a tool to study human vision and attention. *Proc. Natl. Acad. Sci. USA*, 90, 10901-10903 (1993).

7. Desimone, R. Neural mechanisms for visual memory and their role in attention. *Proc. Natl. Acad. Sci. USA*, 93, 13494-13499 (1996).

8. Directed Perception, Inc. Computer Controlled Pan-Tilt Units (PTU-D46). Burlingame, CA. www.dperception.com/products.html#PTU-D46-17

9. Driscoll, J.A., Peters, R.A., Cave, K.R. A Visual Attention Network for a Humanoid Robot. *Proc. 1998 IEEE/RSJ Int'l Conf. Intell. Robotic System*. (IROS'98)

10. Edmondson, A.C. *A Fuller Explanation: the Synergetic Geometry of R. Buckminster Fuller*. Boston: Birkhauser Verlag, January 1987.

11. Floridi, Luciano. Open problems in the philosophy of information. *Metaphilosophy*, Vol. 35, No.4, July 2004.

12. Gottlieb, J., Kusunoki, M., Goldberg, M. The representation of visual salience in monkey parietal cortex. *Nature*, 391, 481-484 (1998).

13. Hambuchen, K.A. "Multi-modal attention and event binding in humanoid robots using a sensory ego-sphere." PhD Dissertation, Vanderbilt University, 2004.

14. Harnad, Stevan. The Symbol Grounding Problem. *Physica D*, Vol. 42, pp. 335-346, 1990.

15. Itti, L. Modeling primate visual attention. *Computational Neuroscience: A Comprehensive Approach*. Boca Raton: CRC Press, 635-655, (2003).

16. Itti,L., Koch, C. A Comparison of Feature Combination Strategies for Saliency-Based Visual Attention Systems. *Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'99)*, 3644, 473-482 (1999).

17. Itti, L., Koch, C. Computational modeling of visual attention. *Nature neuroscience*, 2, 194-203 (2001).

18. Itti,L., Koch, C., Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE transactions on pattern analysis and machine intelligence*, 20, 1254-1259, (1998).

19. Kawamura, K., Peters, R.A. II, Bagchi, S., Iskarous, M., Bishay, M. Intelligent robotic systems in service of the disabled. *IEEE Transactions on Rehabilitation Engineering*, 3(1):14-21, March 1995.

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

21. Klein, R.M., MacInnes, W.J. Inhibition of return is a foraging facilitator in visual search. *Psychological Science*, 10, 346-352 (1999).

22. Kortmann, Rens. Embodied cognitive science. *Proceedings Of Robo Sapiens - The First Dutch Symposium On Embodied Intelligence*, (Eds. W. de Back, T. van der Zant, and L. Zwanepol), Artificial intelligence preprint series, vol. 24, Universiteit Utrecht, Utrecht, The Netherlands, 2001.

23. Lee, D.K., Itti, L., Koch, C., Braun, J. Attention activates winner-take-all competition among visual filters. *Nature Neuroscience*, 2, 375-381 (1999).

24. Marcelja, S. Mathematical Description of the Responses of Simple Cortical Cells. *J. Opt. Soc. Am.*, Vol. 70, 1980.

25. Matlab Camera Calibration Toolbox: http://www.vision.caltech.edu/bouguetj/calib_doc/

26. McIlwain, James T. *An Introduction to the Biology of Vision*. Cambridge University Press, 1998. pp. 11, 212

27. Navalpakkam, V., Itti, L. Modeling the influence of task on attention, *Vision Research,* Vol. **45**, No. 2, pp. 205-231, Jan 2005

28. Peters, R.A. II, Kawamura, K., Wilkes, D.M., Hambuchen, K.A., Rogers, T.E., et al. ISAC Humanoid: An Architecture for Learning and Emotion. *Proceedings of the IEEE-RAS International Conference on Humanoid Robots,* 2001.

29. Peters, R.A. II, Hambuchen, K.A., Kawamura, K., Wilkes, D.M. The Sensory Ego-Sphere as a Short-Term memory for Humanoids. *Proceedings of the IEEE-RAS Conference on Humanoid Robots*, 2001, pp 451-60.

30. Peters, R.A. II, Hambuchen, K.A., Bodenheimer, R.E. The Sensory Ego-Sphere: A Mediating Interface Between Sensors and Cognition. Submitted to *IEEE Transactions on Systems, Man, and Cybernetics*, December, 2003.

31. Pfeifer, R., Scheier, C. *Understanding Intelligence*. Cambridge, Mass. MIT Press, 1999.

32. Posner, M.I., Gilbert,C.D. Attention and primary visual cortex. Proc. *Natl. Acad. Sci. USA*, 96, 2585-2587 (1999).

33. Shapiro, Linda G., Stockman, G.C. *Computer Vision*. Prentice Hall, 2001.

34. Sony XC-999 Cigar Camera. http://www.donlinte.com/level2/sony_xc999.htm

35. Stewart, I. Circularly covering clatharin. *Nature*, vol. 351, p. 103, May 1991.

36. Sun, Li. A Binocular Vision System for a Humanoid Robot. MS Thesis, Vanderbilt University, 2004.

37. Tovée, Martin J. *An Introduction to the Visual System.* Cambridge University Press, 2001. pp. 66,126

38. Urner, K. The invention behind the inventions: synergetics in the 1990's. *The Synergetica Journal*, vol. 1, no. 1, 1991.

39. Weeks, Arthur J. Jr. *Fundamentals of Electronic Image Processing*. SPIE/IEEE Series on Image Science & Engineering, 1996.

40. Wolfe, J.M. Guided Search 2.0: A Revised Model of Visual Search. Psychonomic Bulletin & Review, 1(2): 202-238, 1994.

41. Wolfe, J.M., & Gancarz, G. Guided Search 3.0 Basic and Clinical Applications of Vision Science. Dordrecht, Netherlands: Kluwer Academic. 189-192, 1996.

42. Wolfe, J.M. Guided Search 4.0: A guided search model that does not require memory for rejected distractors [Abstract]. *Journal of Vision*, *1*(3), 349a, 2001. http://journalofvision.org/1/3/349/, doi:10.1167/1.3.349.

43. Wolfe, J.M.,Horowitz, T.S. What attributes guide the deployment of visual attention and how do they do it? *Nature Neuroscience*, Vol 5, June 2004.

42. http://philosophy.hku.hk/courses/cogsci/media/visionstreams.jpg

43. Diagrams courtesy of Dr. Richard Alan Peters, Vanderbilt University, Nashville, Tennessee.