Modeling Attention with Embodied Visual Behaviors

NATHAN SPRAGUE
Kalamazoo College
and
DANA BALLARD and AL ROBINSON
University of Rochester

Most experimental investigations of visual attention essentially measure a subject's differential performance with respect to an attended condition and a control. This design makes it difficult to integrate the results of different attentional studies, as they are typically measured under very different experimental conditions. One way of accomplishing such an integration is to create a model of a human that has a useful amount of complexity. Essentially, one is faced with proposing an embodied "operating system" model that can be tested against human performance. Recently technological advances have been made that allow progress to be made in this direction. Graphics models that simulate extensive human capabilities can be used as platforms from which to develop synthetic models of visuo-motor behavior. Currently such models can capture only a small portion of a full behavioral repertoire, but for the behaviors that they do model, they can describe complete visuo-motor subsystems at a level of detail that can be tested against human performance in realistic environments. This paper outlines one such model and shows both that it can produce interesting new hypotheses as to the role of vision and also that it can enhance our understanding of visual attention.

Categories and Subject Descriptors: I.2.10 [**Vision and Scene Understanding**]: Perceptual reasoning

General Terms: Theory

Additional Key Words and Phrases: attention, reinforcement learning, vision

1. INTRODUCTION

All brain operations are situated in the body [Clark 1997]. Even when the operations are purely mental, they reflect a developmental path through symbols that are grounded in concrete interactions in the world. The genesis of this view is attributed to the philosopher Merleau-Ponty [Merleau-Ponty 1962], but more recently it has been taken as a tenet of research programs for the reason that tremendous computational economies result. Essentially the body does a large part of the necessary computation, leaving the brain with much less to do.

Research programs that focus on embodiment have been facilitated by the development of virtual reality (VR) graphics environments. These VR environments can now run in real time on standard computing platforms. The value of VR environments is that they allow the creation of virtual agents that implement complete visuo-motor control loops. Visual input can be captured from the rendered virtual scene, and motor commands can be used to direct the graphical representation of the virtual agent's body. Terzoupolous and Rabie [Terzopoulos and Rabie 1997] pioneered the use virtual reality as a platform for the study of visually guided control. Embodied control has been studied for many years in the robotics domain, but virtual agents have enormous advantages over physical robots in the areas of experimental reproducibility, hardware requirements, flexibility, and ease of programming.

Embodied models can now be tested using new instrumentation. Linking mental processing to visually-guided body movements at a millisecond timescale would have been impractical just a decade ago, but recently a wealth of high resolution monitoring equipment has been developed for tracking body movements in the course of everyday behavior, particularly head, hand and eye movements (e.g. [Pelz et al. 2001]). This allows for research into everyday tasks that typically have relatively elementary cognitive demands but require elaborate and comprehensive physical monitoring. In these tasks, overt body signals provide a direct indication of mental processing.

The goal of this paper is to introduce the use of virtual humans as a platform for developing models of human visually guided control. We are particularly interested in developing models of visual attention. A tremendous amount of energy has been devoted to the study of human visual attention. A major difficulty in evaluating this body of work is that a wide range of experimental phenomena are placed under the heading of attentional effects, from perceptual learning and visual search experiments whose effects can be seen on a time scale of tens of milliseconds, to dual-task experiments, whose effects can take hundreds of milliseconds or longer. However there is so far no coherent model that relates all these phenomena. Our own view is that it can be misleading to think of attention as something distinct that may or may not be applied to a particular sensori-motor task. It is more informative to recognize that during the course of normal behavior humans engage in a wide variety of tasks, each of which requires certain perceptual and motor resources. Thus there must be mechanisms that allocate resources to tasks. Under this view, understanding attention requires an understanding of the ongoing demands of behavior, as well as the nature of the resources available to the human sensori-motor system. The interaction of these factors is complex, and that is where the virtual human platform can be of value. It allows us to imbue our artificial human with a particular set of resource constraints. We may then design a control architecture that allocates those resources in response to task demands. The result is a model of human behavior in temporally extended tasks that may be tested against human performance.

We refer to our own virtual human model as 'Walter.' Walter has physical extent and programmable kinematic degrees of freedom that closely mimic those of real humans. His graphical representation and kinematics are provided by the DI-guy package developed by Boston Dynamics. This is augmented by the Vortex package

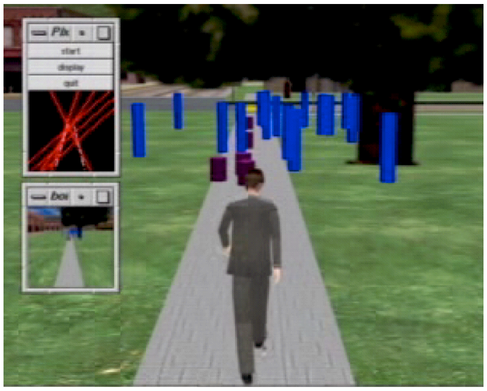

Fig. 1. The Walter simulation. The insets show the use of vision to guide the humanoid through a complex environment. The upper inset shows the particular visual routine that is running at any instant. The lower insert shows the visual field in a head-centered frame.

developed by CMLabs for modeling the physics of collisions. The crux of the model is a control architecture for managing the extraction of information from visual input that is in turn mapped onto a library of motor commands. The model is illustrated on a simple sidewalk navigation task that requires the virtual human to walk down a sidewalk and cross a street while avoiding obstacles and collecting litter. The movie frame in Figure 1 shows Walter in the act of negotiating the sidewalk which is strewn with obstacles (blue objects) and litter (purple objects) on the way to crossing a street.

The body of this paper is divided into two parts. First we describe Walter's control architecture and resource allocation mechanisms in detail. In the second part of the paper we present eye tracking data collected from a human subject engaged in the same sidewalk navigation task, and compare this to the output of the virtual human model.

2. BEHAVIOR BASED CONTROL

As pointed out by Newell [Newell 1990], any system that must operate in a complex and changing environment must be compositional, that is It has to have elemental pieces that can be composed to create its more complex structures. Figure 2 illustrates two broad compositional approaches that have been pursued in theories of cognition, as well as in robotics. The first decomposition works on the assumption that the agent has a central repository of symbolic knowledge. The purpose of perception is to translate sensory information into symbolic form. Actions are selected that result in symbolic transformations that bring the agent closer to goal states. This sense-plan-act approach is typified in the robotics community by early work on Shakey the robot [Nilsson 1984], and in the cognitive science community by the theories of David Marr [Marr 1982]. In principle, the symbolic planning approach is very attractive, since it suggests that sensation, cognition and action can be studied independently, but in practice each step of the process turns out to be difficult to characterize in isolation. It is hard to convert sensory information

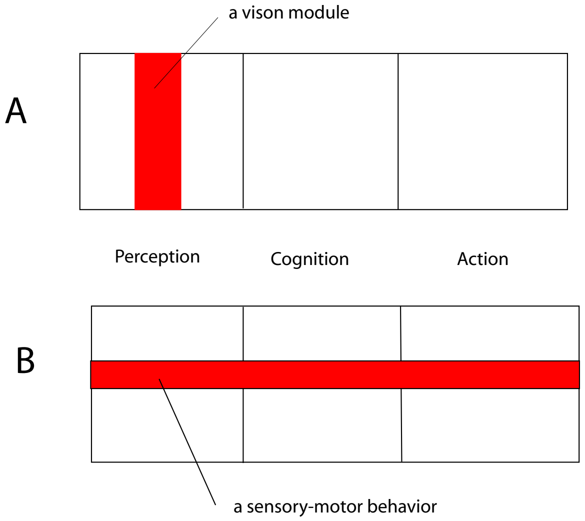

Fig. 2. Two approaches to behavioral research contrasted. A) In the Marr paradigm individual components of vision are understood as units. B) In the Brooks paradigm the primitive unit is an entire behavior.

into general purpose symbolic knowledge, it is hard to use symbolic knowledge to plan sequences of actions, and it is hard to maintain a consistent and up to date knowledge base.

The difficulties with the symbolic planning approach have led to alternate proposals. In the robotics community Brooks [Brooks 1986] has attempted to overcome these difficulties by suggesting a radically different decomposition, illustrated in Figure 2B. Brooks' alternate approach is to attempt to describe whole visuo-motor behaviors that have very specific goals. Behavior-based control involves a different approach to composition than planning-based architectures: simple microbehaviors are sequenced and combined to solve arbitrarily complex problems. The best approach to attaining this sort of behavioral composition is an active area of research. Brooks' own *subsumption* architecture worked by organizing behaviors into fixed hierarchies, where higher level behaviors influenced lower level behaviors by overwriting their inputs. Subsumption works spectacularly well for trophic, low-level tasks, but generally fails to scale to handle more complex problems [Hartley and Pipitone 1991]. For that reason we have chosen a more flexible control architecture.

Our version of Brooks' behavior-based control centers around primitives that we term *microbehaviors*. A microbehavior is a complete sensory/motor routine that incorporates mechanisms for measuring the environment and acting on it to achieve specific goals. For example a collision avoidance microbehavior would have the goal of steering the agent to avoid collisions with objects in the environment. A microbehavior has the property that it cannot be usefully split into smaller subunits. Walter's microbehavior control architecture follows more recent work on behavior based control (e.g. [Firby et al. 1995; Bryson and Stein 2001]) that allows the agent to address changing goals and environmental conditions by dynamically

Abstraction Level	Problem Being Addressed	Role of Vision
Behavior	Need to get state information	Provide State Estimation
	The current state needs to be updated to reflect the actions of the body	None
Arbitration	Active behaviors may have competing demands for body, legs, eyes. Conflicts have to be resolved	Move gaze to the location that will minimize risk
Context	Current set of behaviors B is inadequate for the task. Have to find a new set	Test for off-agenda exigencies

Table I. The organization of human visual computation from the perspective of the microbehavior model.

activating a small set of appropriate behaviors. Each microbehavior is triggered by a template that has a pattern of internal and environmental conditions. The pattern-directed activation of microbehaviors provides a flexibility not found in the fixed subsumption architecture.

3. THE COMPUTATIONAL HIERARCHY

The central tenet of Walter's control architecture is that, although a large library of microbehaviors is available to address the goals of the agent, at any one time, only a small subset of those are actively engaged. Addressing the issues associated with this vantage point leads directly to an abstract computational hierarchy. The issues in modeling vision are different at each level of this hierarchy. Table 1 shows the basic elements of our hierarchy highlighting the different roles of vision at each level.

The behavior level of the hierarchy addresses the issues in running a microbehavior. These are each engaged in maintaining relevant state information and generating appropriate control signals. Microbehaviors are represented as state/action tables, so the main issue is that of computing state information needed to index the table. The arbitration level addresses the issue of managing competing behaviors. Since the set of active microbehaviors must share perceptual and motor resources, there must be some mechanism to arbitrate their needs when they make conflicting demands. The context level of the hierarchy maintains an appropriate set of active behaviors from a much larger library of possible behaviors, given the agents current goals and environmental conditions. The composition of this set is evaluated at every simulation interval, which we take to be 300 milliseconds.

The issues that arise for vision are very different at the different levels of the hierarchy. Moving up the levels:

(1) At the level of individual behaviors, vision provides its essential role of computing state information. The issue at this level is understanding how vision can be used to compute state information necessary for meeting behavioral goals.

Almost invariably, the visual computation needed in a task context is vastly simpler than that required general purpose vision and, as a consequence, can be done very quickly.

(2) At the arbitration level, the principal issue for vision is that the center of gaze is not easily shared and instead generally must be allocated sequentially to different locations. Eye tracking research increasingly is showing that all gaze allocations are purposeful and directed toward computing a specific result [Land et al. 1999; Hayhoe et al. 1998; Johansson et al. 1999]. Our own model [Sprague and Ballard 2003a] shows how gaze allocations may be selected to minimize the risk of losing reward in the set of running behaviors.

(3) At the context level, the focus is to maintain an appropriate set of microbehaviors to deal with internally generated goals. One of these goals is that the set of running behaviors be response to rapid environmental changes. Thus the issue for vision at this level is understanding the interplay between agenda-driven and environmentally-driven visual processing demands.

This hierarchy immediately presents us with a related set of questions: How do the microbehaviors get perceptual information? How is contention managed? How are sets of microbehaviors selected? In subsequent sections, we use the hierarchical structure to address each of these in turn, emphasizing implications for vision.

4. STATE ESTIMATION USING VISUAL ROUTINES

The first question that must be addressed is how individual microbehaviors map from sensory information to internal state descriptions. The position we adopt is that this information is gathered by deploying visual routines. These are a small library of special-purposed functions that can be composed. The arguments for visual routines have be made by [Ullman 1985; Roelfsema et al. 2000; Kosslyn and Shwartz 1977; Ballard et al. 1997]. The main one is that the representations of vision such as color and form, are problem-neutral in that they do not contain explicitly the data upon which control decisions are made.[1] and thus an additional processing step must be employed to make decisions. The number of potential decisions that must be made is too large to pre-code them all. Visual routines address this problem in two ways: 1) routines are composable and 2) routines process visual data in an as-needed fashion.

To illustrate the use of visual routines, we describe the ones that create the state information for three of Walter's microbehaviors: collision avoidance, sidewalk navigation and litter collection. Each of these requires specialized processing. This processing is distinct from that used to obtain the feature images of early vision even though it may use such images as data. The specific processing steps are visualized in Figure 3.

—Litter collection is based on color matching. Litter is signaled in our simulation by purple objects, so that potential litter must be isolated as being of the right color and also nearby. This requires combining and processing the hue image with depth information. The result of this processing is illustrated in Figure 3b.

[1] Marr recognized this difficulty of processing visual data prior to knowing what it will be needed for implicitly in his 'principle of least commitment' [Marr 1982].

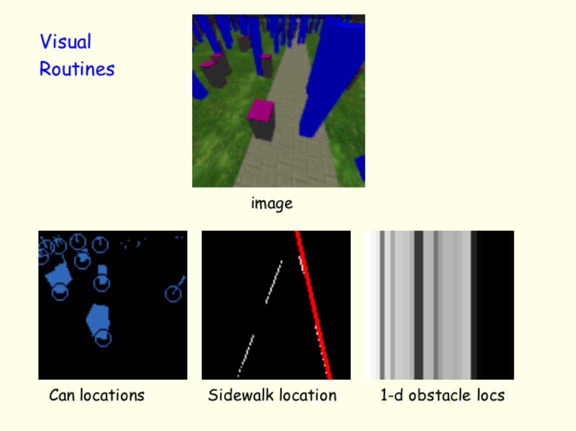

Fig. 3. The Visual Routines that compute state information. a) Input image from Walter's viewpoint. b) Regions that fit the litter color profile. Probable litter locations are marked with circles. c) Processed image for sidewalk following. Pixels are labeled in white if they border both sidewalk and grass color regions. The red line is the most prominent resulting line. b) One dimensional depth map used from obstacle avoidance (not computed directly from the rendered image).

—Sidewalk navigation uses color information to label pixels that border both sidewalk and grass regions. A line is fit to the resulting set of pixels which indicates the estimated edge of the sidewalk. The result of this processing is illustrated in Figure 3c.

—The collision detector uses a depth image. A depth image may be created by any of a number of cues, (stereo, kinetic depth, parallax depth, etc.) but for collisions, it must be processed to isolate potential colliders. The result of this processing is illustrated in Figure 3d. A study with human subjects shows that they are very good at this, integrating motion cues with depth to ignore close objects that are not on a collision course [Ballard and Sprague 2002].

Regardless of the specific methods of individual routines, each one outputs information in the same abstract form: the state needed to guide its encompassing microbehavior. The next section describes how Walter can learn to use this information to guide its parent microbehavior.

5. LEARNING MICROBEHAVIORS

Once state information has been computed, the next step is to find an appropriate action. Each microbehavior stores actions in a state/action table. Such tables can be learned by reward maximization algorithms: Walter tries out different actions in the course of behaving and remembers the ones that worked best in the table. The reward-based approach is are motivated by studies of human behavior that show that the extent to which humans make such trade-offs is very refined [Maloney and Landy pear] as well as studies using monkeys that reveal the use of reinforcement signals in a way that is consistent with reinforcement learning algorithms [Suri and

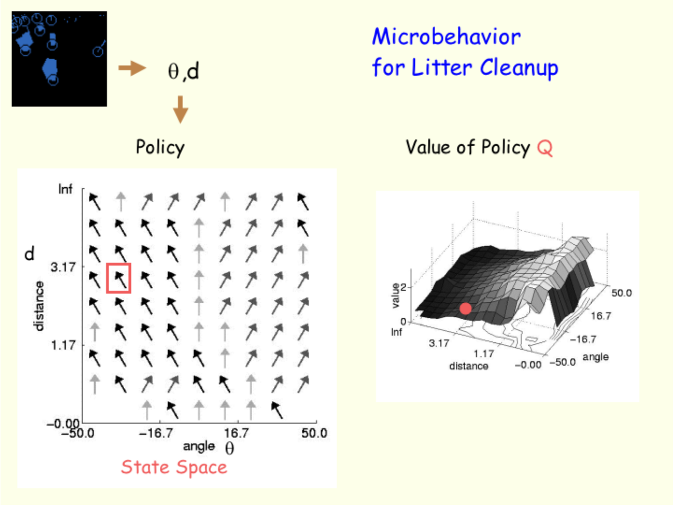

Fig. 4. The central portion of the litter cleanup microbehavior after it has been learned. The color image is used to identify the heading to the nearest litter object as a heading angle θ and distance d. Using this state information to index the table allows the recovery of the policy, in this case $heading = -45^o$, and its associated value. The fact that the model is embodied means that there is we can assume there is neural circuitry to translate this abstract heading into complex walking movements. This is true for the graphics figure that has a 'walk' command that takes a heading parameter.

Schultz 2001].

Formally, the task of each microbehavior is to map from an estimate of the relevant environmental state s, to one of a discrete set of actions, $a \in A$, so as to maximize the amount of reward received. For example the the obstacle avoidance behavior maps the distance and heading to the nearest obstacle $s = (d, \theta)$ to one of three possible turn angles, that is, $A = \{-15^o, 0^o, 15^o\}$. The *policy* is the action so prescribed for each state. The coarse action space simplifies the learning problem.

Our approach to computing the optimal policy for a particular behavior is based on a standard reinforcement learning algorithm, termed Q-learning[Watkins and Dayan 1992]. This algorithm learns a value function $Q(s, a)$ for all the state-action combinations in each microbehavior. The Q function denotes the expected discounted return if action a is taken in state s and the optimal policy is followed thereafter. If $Q(s, a)$ is known then the learning agent can behave optimally by always choosing $\arg\max_a Q(s, a)$(See Appendix for details). Figure 4 shows the table used by the litter collection microbehavior, as indexed by its state information.

Each of the three microbehaviors has a two-dimensional state space. The litter collection behavior uses the same parameterization as obstacle avoidance: $s = (d, \theta)$ where d is the distance to the nearest litter item, and θ is the angle. For the sidewalk following behavior the state space is $s = (\rho, \theta)$. Here θ is the angle of the center-line of the sidewalk relative to the agent, and ρ is the signed distance to the center of the

Outcome	**Immediate Reward**
Picked up a litter can	2
On sidewalk	1
Collision free	4

Table II. Walter's reward schedule

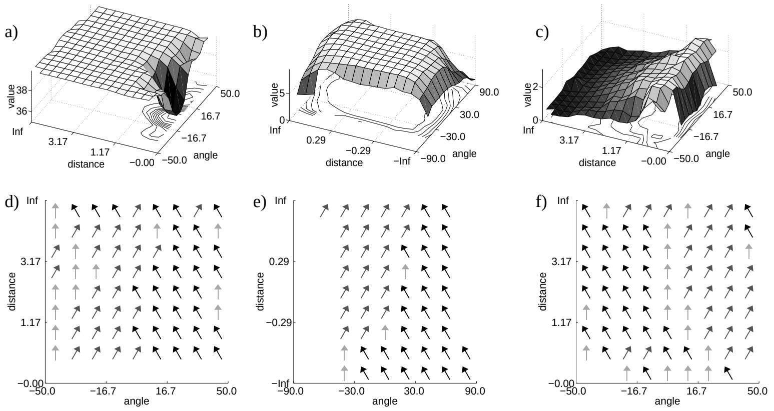

Fig. 5. Q-values and policies for the three microbehaviors. Figures a)-c) show $\max_a Q(s,a)$ for the three microbehaviors: a) obstacle avoidance, b) sidewalk following and c) litter collection. Figures d)-f) show the corresponding policies for the three microbehaviors. The obstacle avoidance value function shows a penalty for nearby obstacles and a policy of avoiding them. The sidewalk policy shows a benefit for staying in the center of the sidewalk $\theta = 0, \rho = 0$. The litter policy shows a benefit for picking up cans that decreases as the cans become more distant. The policy is to head toward them.

sidewalk, where positive values indicate that the agent is to the left of the center, and negative values indicate that the agent is to the right. All microbehaviors use the logarithm of distance in order to devote more of the state representation to areas near the agent. All these microbehaviors use the same three-heading action space described above. Table II shows Walter's reward contingencies. These are used to generate the Q-tables that serve as a basis for encoding a policy. Figure 5 shows a representation of the Q-functions and policies for the three microbehaviors.

When running the Walter simulation, the Q-table associated with each behavior is indexed every 300 milliseconds. The action that is the policy is selected and submitted for arbitration. The action chosen by the arbitration process is executed by Walter. This in turn results in a new Q-table index for each microbehavior and the process is repeated. The path through a Q-table thus evolves in time and can the visualized as a thread of control analogous to the use of the term thread in computer science.

6. MICROBEHAVIOR ARBITRATION

A central complication with the microbehavior approach is that concurrently active microbehaviors may prefer incompatible actions. Therefore an arbitration mechanism is required to map from the demands of the individual microbehaviors to final action choices. The arbitration problem arises in directing the physical control of the agent, as well as in handling gaze control and each of these requires a different solution. This is because in Walter's environment, his heading can be a compromise between the demands of different microbehaviors but his gaze location is not readily shared by them. A benefit of knowing the value function for each behavior is that the Q-values can be used to handle the physical arbitration problem in each of these cases.

Heading Arbitration Since in the walking environment each behavior shares the same action space Walter's heading arbitration is handled by making the assumption that the Q-function for the composite task is approximately equal to the sum of the Q-functions for the component microbehaviors:

$$Q(s, a) \approx \sum_{i=1}^{n} Q_i(s_i, a), \tag{1}$$

where $Q_i(s_i, a)$ represents the Q-function for the ith active behavior. Thus the action that is chosen is a compromise that attempts to maximize reward across the set of active microbehaviors. The idea of using Q-values for multiple goal arbitration was independently introduced in [Humphrys 1996] and [Karlsson 1997].

In order to simulate the fact that only one area of the visual field may be foveated at a time, only one microbehavior is allowed access to perceptual information during each 300ms simulation time step. That behavior is allowed to update its state information with a measurement, while the others propagate their estimates and suffer an increase in uncertainty. The mechanics of maintaining state estimates and tracking uncertainty are handled using Kalman filters - one for each microbehavior. In order to simulate noise in the estimators, the state estimates are corrupted with zero-mean normally distributed random noise at each time step. The noise has a standard deviation of .2m in both the x and y dimensions. When a behavior's state has just been updated by its visual routine's measurement, the variance of the state distribution will be small, but as we will demonstrate in simulation, in the absence of such a measurement the variance can grow significantly.

Since Walter may not have perfectly up to date state information, he must select the best action given his current estimates of the state. A reasonable way of selecting an action under uncertainty is to select the action with the highest expected return. Building on Equation (1) we have the following: $a_E = \arg\max_a E[\sum_{i=1}^{n} Q_i(s_i, a)]$, where the expectation is computed over the state variables for the microbehaviors. By distributing the expectation, and making a slight change to the notation we can write this as:

$$a_E = \arg\max_a \sum_{i=1}^{n} Q_i^E(s_i, a), \tag{2}$$

where Q_i^E refers to the expected Q-value of the ith behavior. In practice we estimate these expectations by sampling from the distributions provided by the Kalman filter.

Gaze Arbitration Arbitrating gaze requires a different approach than arbitrating control of the body. Reinforcement learning algorithms are best suited to handling actions that have direct consequences for a task. Actions such as eye movements are difficult to put in this framework because they have only indirect consequences: they do not change the physical state of the agent or the environment; they serve only to obtain information.

A much better strategy is to choose to use gaze to update the behavior that has *the most to lose* by not being updated. Thus, the approach taken here is to try to estimate the value of that information. Simply put, as time evolves the uncertainty of the state of a behavior grows, introducing the possibility of low rewards. Deploying gaze to measure that state reduces this risk. Estimating the cost of uncertainty is equivalent to estimating the expected cost of incorrect action choices that result from uncertainty. Given that the Q functions are known, and that the Kalman filters provide the necessary distributions over the state variables, it is straightforward to estimate, this factor, $loss_b$, for each behavior b by sampling (See Appendix). The maximum of these values is then used to select which behavior should be given control of gaze.

Figure 6 gives an example of seven consecutive steps of the sidewalk navigation task, the associated eye movements, and the corresponding state estimates. The eye movements are allocated to reduce the uncertainty where it has the greatest potential negative consequences for reward. For example, the agent fixates the obstacle as he draws close to it, and shifts perception to the other two microbehaviors when the obstacle has been safely passed. Note that the regions corresponding to state estimates are not ellipsoidal because they are being projected from world-space into the agents non-linear state space.

One possible objection to this model of eye movements is that it ignores the contribution of extra-foveal vision. One might assume that the pertinent question is not which microbehavior should direct the eye, but which location in the visual field should be targeted to best meet the perceptual needs of the whole ensemble of active microbehaviors. There are a number of reasons that we choose to emphasize foveal vision. First, eye tracking studies in natural tasks show little evidence of "compromise" fixations. That is, nearly all fixations are clearly directed to a particular item that is task relevant. Second, results in [Roelfsema et al. 2003] suggest that simple visual operations such as local search and line tracing require a minimum of 100-150ms to complete. This time scale roughly corresponds to the time required to make a fixation. This suggests that there is little to be gained by sharing fixations among multiple visual operations.

7. MICROBEHAVIOR SELECTION

The successful progress of Walter is based on having a running set of microbehaviors $B_i, i = 0, .., N$ that are appropriate for the current environmental and task context. The view that visual processing is mediated by a small set of microbehaviors immediately raises two questions: 1) What is the exact nature of the context switching mechanism? and 2) What should the limit on N be to realistically model the limitations of human visual processing?

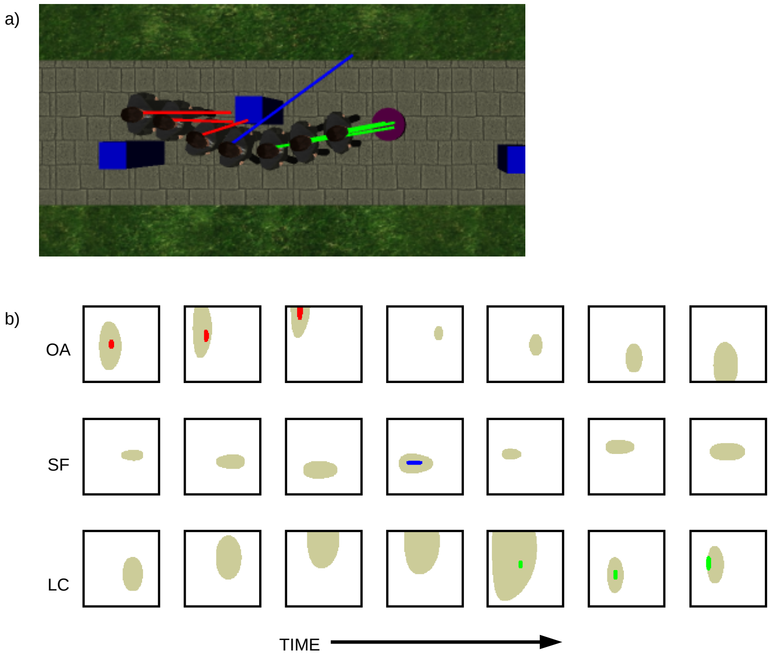

Fig. 6. a) An overhead view of the virtual agent during seven time steps of the sidewalk navigation task. The blue cubes are obstacles, and the purple cylinder is litter. The rays projecting from the agent represent eye movements; red correspond to obstacle avoidance, blue correspond to sidewalk following, and green correspond to litter collection. b) Corresponding state estimates. The top row shows the agent's estimates of the obstacle location. The axes here are the same as those presented in Figure 5. The beige regions correspond to the 90% confidence bounds before any perception has taken place. The red regions show the 90% confidence bounds after an eye movement has been made. The second and third rows show the corresponding information for sidewalk following and litter collection.

Answering the first question requires considering to what extent visual processing is driven in a top down fashion by internal goals, versus being driven by bottom up signals originating in the environment. Somewhat optimistically, some researchers have assumed that interrupts from dynamic scene cues can effortlessly and automatically attract the brain's "attentional system" in order to make the correct context switch e.g [Itti and Koch 2000]. However, a strategy of predominantly bottom-up interrupts seems unlikely in light of the fact that what constitutes a relevant cue is highly dependent on the current situation. On the other hand, there is a strong argument for some bottom up component: humans are clearly capable of responding appropriately to cues that are off the current agenda.

Our model of the switching mechanism is that it works as a state machine as shown in Figure 7. For planned tasks, certain microbehaviors keep track of the progress through the task and trigger new sets of behaviors at predefined junc-

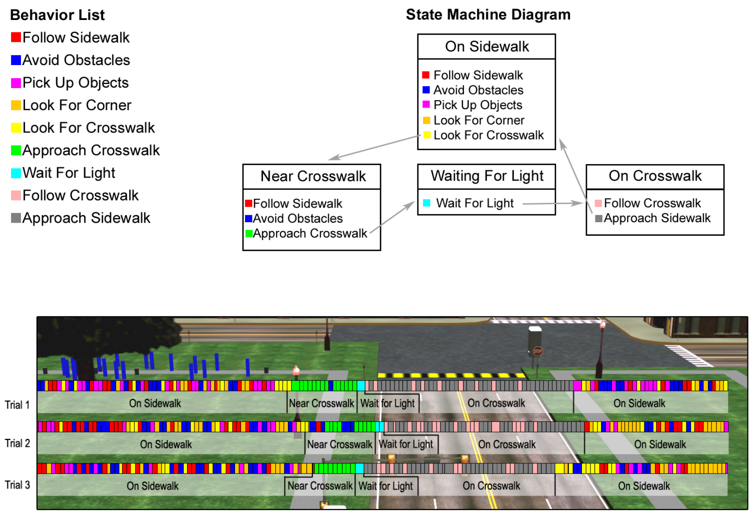

Fig. 7. (Top left) A list of microbehaviors used in Walter's overall navigation task. (Top right) The diagram for the programmable context switcher showing different states. These states are indicated in the bands underneath the colored bars below. Bottom) Context switching behavior in the sidewalk navigation simulation for three separate instances of Walter's stroll. The different colored bars denote different microbehaviors that are in control of the gaze at any instant.

tures. Thus the microbehavior "Look for Crosswalk" triggers the state NEAR-CROSSWALK which contains three microbehaviors: "FollowSidewalk", "Avoid Obstacles", and "Approach Crosswalk."

Figure 7B shows when the different states were triggered on three separate trials.

This model reflects our view that vision is predominantly a top-down process. The model is sufficient for handling simple planned tasks, but it does not provide a straightforward way of responding to off-plan contingencies. To be more realistic, the model requires some additions. First, microbehaviors should be designed to error-check their sensory input. In other words, if a microbehavior's inputs do not match expectations, it should be capable of passing control to a higher level procedure for resolution. Second, there should be a low latency mechanism for responding to certain unambiguously important signals such as rapid looming.

Regarding the second question of the possible number of active microbehaviors, there are at least two reasons to suspect that the maximum number that are simultaneously running might be modest. The first reason is the ubiquitous observation of the limitations of spatial working memory (SWM). The original capacity estimate by Miller was seven items plus or minus two [Miller 1956], but current estimates favor the lower bound [Luck and Vogel 1997]. We hypothesize that this limitation is tied to the number of independently running microbehaviors which we have termed threads. The identification of the referents of SWM has always been problematic, since the size of the referent can be arbitrary. This has lead to the denotation of the referent as a ‘chunk,’ a jargon word that postpones dealing with the issue of not being able to quantify the referents. The thread concept is clearer and more specific as it denotes exactly the state necessary to maintain a microbehavior.

The second factor limiting the number of running microbehaviors is that large numbers of active microbehaviors may not be possible given that they have to be implemented in a neural substrate. Cortical memory is organized into distinct areas that have a two-dimensional topography. Furthermore spatial information is usually segregated from feature based information so that the neurons representing the colors of two objects are typically segregated from the neurons representing their location. As a consequence there is no simple way of simultaneously associating one object’s color with its location together with another object’s association of similar properties (This difficulty is the so-called “binding problem” [von der Malsburg 1999]). Some proposals for resolving the binding problem hypothesize that the number of active microbehaviors is limited to one, but this seems very unlikely. However the demands of a binding mechanism may limit the number of simultaneous bindings that can be active. Thus it is possible that such a neural constraint may be the basis for the behavioral observation.

Although the number of active microbehaviors is limited there is reason to believe that it is greater than one. Consider the task of walking on a crowded sidewalk. Two fast walkers approaching each other close at the rate of 6 meters/second. Given that the main source of advanced warning for collisions is visual and that eye fixations typically need 0.3 seconds and that cortical processing typically needs 0.2-0.4 seconds, during the time needed to recognize an impending collision, the colliders have traveled about 3 meters, or about one and a half body lengths. In a crowded situation, this is insufficient advance warning for successful avoidance. What this means is that for successful evasions, the collision detection calculation has to be ongoing. But that in turn means that it has to share processing with the other tasks that an agent has to do. Remember that by sharing we mean that the microbehavior has to be simultaneously active over a considerable period, perhaps minutes. Several elegant experiments have shown that there can be severe interference when multiple tasks have to be done simultaneously, but these either restrict the input presentation time [VanRullen et al. 2004] or the output response time [Pashler 1998]. The crucial issue is what happens to the internal state when it has to be maintained for an extended period.

8. TESTING THE MODEL

Ultimately, an extensive experimental program would be required to establish all the structure of the model, however at this point we have conducted preliminary tests to gain evidence for some of its claims. To do this we introduced humans into the virtual environment and had them walk Walter’s walk. The humans wear a head-mounted binocular display that contains monocular eye tracking capability. in addition the rotational and translational degrees of freedom of their heads are monitored with a Hi-ball tracker. The head tracker has a latency of a few milliseconds so that the experience in the HMD has no detectable lags. One problem faced by the overall setup is that the linear track of Walter’s path is many times longer than the 7 meter width of the laboratory. Our solution to this discrepancy

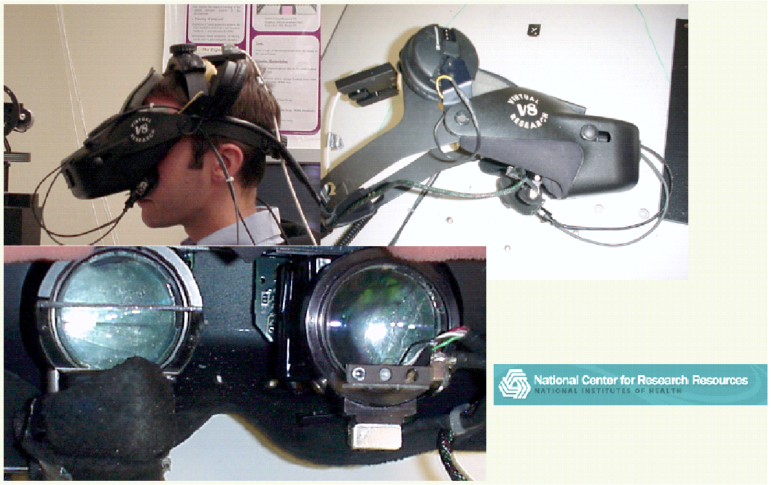
V8
National Center for Research Resources
NATIONAL INSTITUTES OF HEALTH

Fig. 8. The head mounted display worn by human subjects has eye tracking capability so that gaze can be tracked in virtual environments.)

was to map a curved path in motor space onto a linear path in visual space. That is, in order to experience a linear path in visual space, the subjects have to walk a circular path in the laboratory. A typical transit of Walter's path takes about four laps of this path. Eye movement data for each of six subjects is collected and scored on a frame-by-frame basis. As shown in Figure 7, Walter's path consists of a sidewalk portion where he has to handle staying on the sidewalk, obstacle avoidance and litter, and then a clear sidewalk segment followed by a crossing of a street. the crossing of the street is regulated with a large traffic light. Three subjects walked the sidewalk portion only and three additional subjects subsequently walked the entire segment.

The first claim of the model is that the use of fixation is to gather information for a small set of active microbehaviors. This claim is borne out in a number of ways in the subjects' data. In the first place over 95% of the fixations can be interpreted as gathering information for one of the ongoing tasks. For example in the initial segment, the fixations are invariably on the edge of the sidewalk or on a blue pillar (the obstacle) or a purple box (the litter). Figure 9 shows examples of the scored fixations.

A further point of comparison was obtained by sampling individual frames and running the Itti saliency computation on each frame[Itti and Koch 2000]. This software is a model for human eye fixations. Its central claim is that constellations of image features define locations of "saliency." Observed points of fixation in an image can be explained as being chosen from the most salient of these locations. In a limited sample, we compared the actual points chosen by human subjects to the points recommended by the Itti algorithm. Our comparison was generous: if one of the top five points recommended by the algorithm was on the same object as the human fixation it was scored as a match, otherwise it was denoted a non-match. Of eighteen points tested only eight matched under these criteria. The non-match

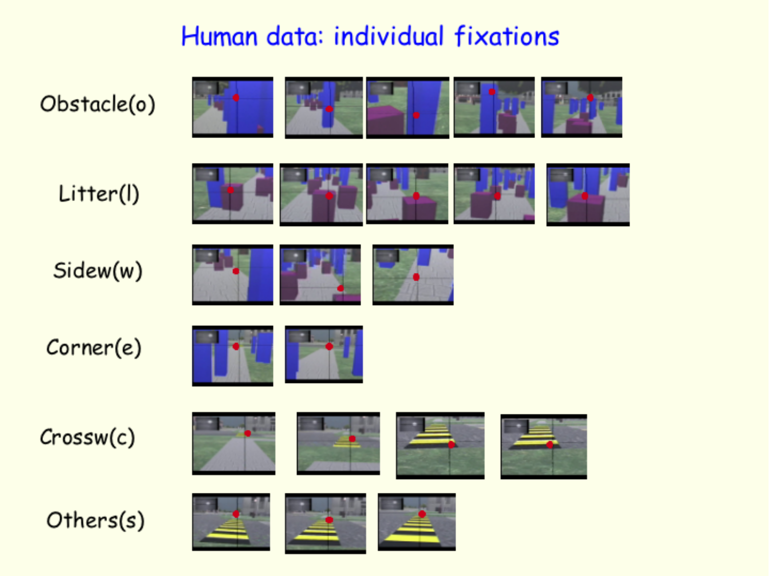

Fig. 9. Sample scored frames from the human video augmented with a red dot to highlight the fixation point.)

points were in the majority. As all of the human data could be readily interpreted as being directly relevant to one of the three tasks, we take this as evidence for task-directed visual routines. The second claim of the model is that in the course of natural behavior a small number of microbehaviors are active and these are competing for the gaze vector. On the initial sidewalk segment the fixation data was directed predominantly between one of the kinds of locations relevant to the three tasks. While our data cannot rule out all alternate interpretations of gaze control besides the 'most-to-lose' strategy, such as choosing fixation locations at random or looking at the nearest of the three objects, it can be used to rule out some models such as a simple fixed alternation strategy. Furthermore the data show similar behavior patterns to that of the human subjects. Figure 11 shows the histogram of fixations for three subjects in the initial sidewalk task compared to three runs of Walter over the same data. The figure shows that the subjects used more fixations than the model reflecting that Walter's walking speed was higher. More importantly it shows that the relative proportions of fixations on locations relevant to each of the three tasks was the same. Of course we chose the relative rewards in Table 5 to model the human data but the coarseness of values in that table shows that no extensive tuning was done. The one discrepancy in the table is that the humans use fewer sidewalk fixations than suggested by the model. Our explanation is that the human subjects make some litter and obstacle fixations do double duty. For example, if you are on the sidewalk and fixating an unobstructed litter can that is also on the sidewalk, you can confidently walk toward it knowing you will remain on the sidewalk.

The third claim of the model is that the mix of running behaviors is modulated by

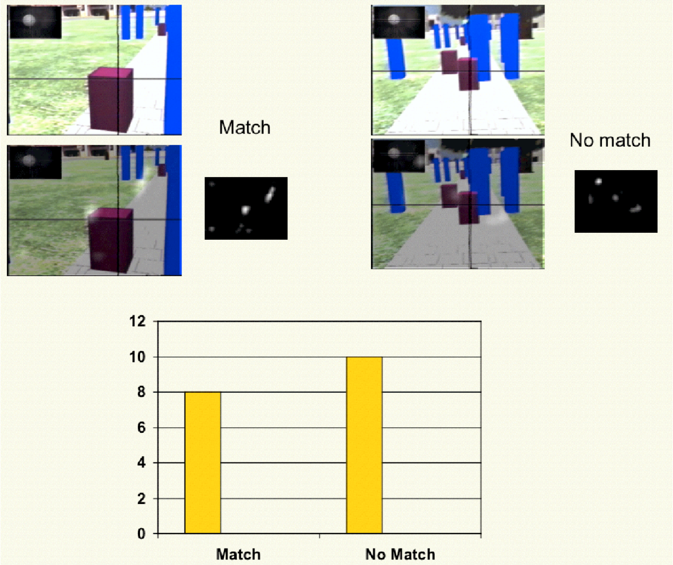

Fig. 10. Comparing human gaze locations to those found by the Itti saliency detector. The small inserts show the saliency maps that are overlaid as transparencies on the lower versions of the images. In a sample of 18 frames, more than half show fixation locations that are not detected by the maps. The saliency program was provided by Dr. Lawrence Itti at USC.

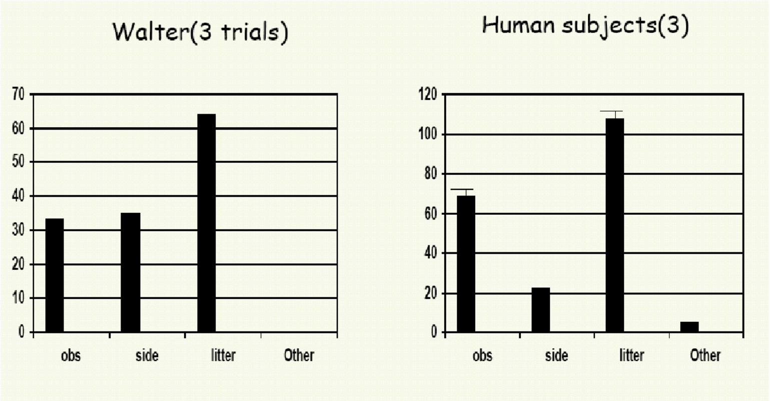

Fig. 11. Comparing the model and human subjects' fractional gaze allocation to different tasks)

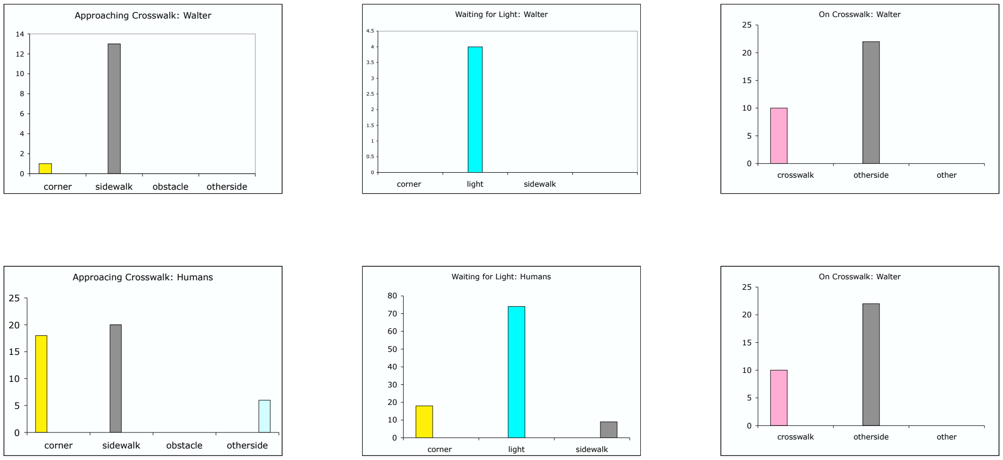

Fig. 12. Histograms for three subjects walking the entire path showing gaze allocations in different context. For human subjects the task being serviced is estimated based on the fixation point location.)

context. This is also borne out by our data. We had three subjects walk the entire course taken by Walter that included the street crossing and computed histograms for the fixations for the entire track. Once again the fixations could almost always be interpreted as task relevant as shown in Figure 12. Furthermore the relative proportions of fixations varied in different contexts in a very similar manner to Walter's fixation variations.

None of these comparisons are definitive and the detailed experiments that would test the model have yet to be done. Nonetheless the closeness of the correspondences between the human and Walter data suggest that the triage of structures used by the model have predictive value. Furthermore this data would be much more difficult to model given a more monolithic notion of attention.

9. CONCLUSIONS

The focus of this paper was to introduce the issues associated with using a graphical agent as a proto-theory of human visuo-motor behavior. One criticism of such a project is that, even though the system is vastly reduced from that needed to capture a substantial fraction of human behavior, the model as it stands is complicated and has enough free parameters so that any data from real human performance would be easy to fit. Although the system is complex, most of the constraints follow from the top-level assumption of composable microbehaviors. Once one decides to have a set of running microbehaviors, the questions of how many and when are they running are immediate. Furthermore they have ready answers in observations of human behavior in the classic observations of working memory and eye movements: Working memory suggests the number of simultaneous microbehaviors is small; eye movements suggest when a behavior is running as each fixation is an

Abstraction Level	**Attention**	**Working Memory**
Behavior	YES	The *contents* of working memory
Arbitration	YES	The referents of working memory or "chunks"
Context	YES	

Table III. The relationships between attention and working memory and the microbehavior model.

indication of the brain's instantaneous problem being updated. Table III summarizes the relationships between the hierarchy used by the model and the notions of attention and working memory.

The restricted number of active microbehaviors means that there must be a mechanism for making sure that a good behavioral subset has been chosen. Such a mechanism must interrogate the environment and 1) add needed microbehaviors as well as 2) drop microbehaviors if needed to meet the capacity constraint.

The essential description of microbehaviors is captured by reinforcement learning's Q-tables that relate the states determined by vision to actions for the motor system. Indeed the commands are in coded form, taking advantage of known structure in the body that carries them out. Assuming the existence of a table as is done at the reinforcement learning level finesses important details. Thus a more detailed model is necessary to account for how the table index is created.

The reinforcement learning venue provides a different perspective on gaze allocation. One of the original ideas was a bottom-up view that gaze should be drawn to the most salient locations in the scene as represented in the image, where salience was defined in terms of the spatial conjunction of many feature points. However recent measurements have shown that eye movements are much more agenda driven than that predicted by bottom-up saliency models. For example Henderson has shown that subjects examining urban scenes for people examine place where people might be even though these can have very low feature saliency [Ballard and Sprague 2002]. Walter's use of Q-tables suggest that to interpret gaze allocation, an additional level of indirection may be required. For example, the controller for sidewalk navigation uses gaze to update the estimate of the location of the sidewalk. In order to predict when gaze might be allocated to do this, in our model, requires knowing the uncertainty in the current estimate of the sidewalk location.

The most important benefit of the kind of model presented in this paper is that it encourages the modeler to frame experimental questions in the context of integrated natural behavior. There are dramatic differences between this perspective and traditional approaches to studying vision:

(1) The desired schedule of interrupts under normal behavior has a temporal distribution that is very different than worst-case laboratory situations. In the lab, subjects are typically in extremis with respect to reaction times, whereas natural behaviors typically allow flexibility in responding.

(2) In a multiple task situation, the most important task facing the deployment of

gaze is to choose the behavior being serviced. This problem is hardly considered in the search literature which concentrates on within-task saliency of individual targets.

(3) The natural timescale for studying microbehavior components is on the order of 100 to 200 milliseconds, the time to estimate state information. Below that one is studying the process of state formation, a level of detail is interesting in its own right but is below the central issues in human behavioral modeling.

(4) The context for the deployment of visual routines is reversed from a laboratory situation. In that situation the typical structure of a task forces a bottom-up description. The image is most often presented on a previously blank CRT screen. In a natural task, the particular test needed in a gaze deployment is known. Furthermore this test is known before the saccade is made. Thus in the natural case the situation is reversed, the test can be in place before the data is available. This has the result of making the test go as fast as possible. The speed of tests may account for the fact that fixation times in natural situations can be very short. Dwell times of 100 milliseconds are normal, less than half those observed in many laboratory studies.

All of these observations underline the importance of graphic simulation as a new tool in the study of human vision. While the model has extensive structure, each component of the structure serves a specific purpose and the whole combine to direct the performance of human behaviors. A competing performance model might look very different but would have to address these issues.

Perhaps the most important theme in recent vision research, is that no component of the visual system can be properly understood in isolation from the behavioral goals of the organism. Therefore, properly understanding vision will ultimately require modeling complete sensori-motor systems in behaving agents. The model presented in this paper is certainly not true in all of its particulars, and it leaves many details unspecified. However, it does provide a framework for thinking about action-oriented human vision. The fact that developing complete and correct models of human vision is such a difficult task should not stop us from trying to put as many of the pieces together as possible.

Appendix: Reinforcement Learning Details

Learning behaviors There are a number of algorithms for learning $Q(s, a)$ [Kaelbling et al. 1996; Sutton and Barto 1998] the simplest is to take random actions in the environment and use the Q-learning update rule [Watkins 1989]:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a'))$$

Here $\alpha \in (0, 1)$ is a learning rate parameter, $\gamma \in (0, 1)$ is a term that determines how much to discount future reward, and s' is the state that is reached after action a. As long as each state-action pair is visited infinitely often in the limit, this update rule is guaranteed to converge to the optimal value function. The Q-learning algorithm is guaranteed to converge only for discrete case tasks with Markovian transitions between states. Walter's tasks are more naturally described using continuous state variables. The theoretical foundations of continuous state reinforcement learning are not as well established as for the discrete state case. However empirical results suggest that good results can be obtained by using a function approximator such as a CMAC along with the Sarsa(0) learning rule: [Sutton 1996]

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r + \gamma Q(s',a'))$$

This rule is nearly identical to the Q-learning rule, except that the max action is replaced by the action that is actually observed on the next step. The Q-functions used throughout this paper are learned using this approach. A more detailed account of the learning procedure can be found in [Sprague and Ballard] and [Sprague and Ballard 2003b].

Choosing behaviors for a state update Whenever Walter chooses an action that is sub-optimal for the true state of the environment, he can expect to lose some return. We can estimate the expected loss as follows:

$$loss = E[\max_a \sum Q_i(s_i,a)] - E[\sum Q_i(s_i,a_E)]. \tag{3}$$

The term on the left-hand side of the minus sign expresses the expected return that Walter would receive if he were able to act with knowledge of the true state of the environment. The term on the right expresses the expected return if he is forced to choose an action based on his state estimate. The difference between the two can be thought of as the cost of the agent's current uncertainty. This value is guaranteed to be positive, and may be zero if all possible states would result in the same action choice.

The total expected loss does not help to select *which* of the microbehaviors should be given access to perception. To make this selection, the loss value can be broken down into the losses associated with the uncertainty for each particular behavior b:

$$loss_b = E\left[\max_a \left(Q_b(s_b,a) + \sum_{i \in B, i \neq b} Q_i^E(s_i,a)\right)\right] - \sum_i Q_i^E(s_i,a_E). \tag{4}$$

Here the expectation on the left is computed only over s_b. The value on the left is the expected return if s_b were known, but the other state variables were not. The value on the right is the expected return if none of the state variables are known. The difference is interpreted as the cost of the uncertainty associated with s_b.

REFERENCES

Ballard, D., Hayhoe, M., and Pook, P. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences 20*, 723–767.

Ballard, D. and Sprague, N. 2002. Attentional resource allocation in extended natural tasks [abstract]. *Journal of Vision 2,* 7, 568a.

Brooks, R. A. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation RA-2,* 1 (Apr.), 14–23.

Bryson, J. J. and Stein, L. A. 2001. Modularity and design in reactive intelligence. In *International Joint Conference on Artificial Intelligence.* Seattle, Washington.

Clark, A. 1997. *Being There: Putting Brain, Body, and World Together Again.* Cambridge, MA: MIT Press.

Firby, R. J., Kahn, R. E., Prokopowicz, P. N., and Swain, M. J. 1995. An architecture for vision and action. 72–79.

Hartley, R. and Pipitone, F. 1991. Experiments with the subsumption architecture. In *Proceedings of the International Converence on Robotics and Automation.*

Hayhoe, M. M., Bensinger, D., and Ballard, D. H. 1998. Task constraints in visual working memory. *Vision Research 38*, 125–137.

Humphrys, M. 1996. Action selection methods using reinforcement learning. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*.

Itti, L. and Koch, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40,* 10-12 (May), 1489–1506.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, J. R. 1999. Eye-hand coordination in object manipulation. *Perception 28*, 1311–1328.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. 1996. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research 4*, 237–285.

Karlsson, J. 1997. Learning to solve multiple goals. Ph.D. thesis, University of Rochester.

Kosslyn, S. M. and Shwartz, S. 1977. A simulation of visual imagery. *Cognitive Science 1*, 265–269.

Land, M., Mennie, N., and Rusted, J. 1999. The roles of vision and eye movements in the activities of daily living. *Perception 28*, 1311–1328.

Luck, S. J. and Vogel, E. K. 1997. The capacity of visual working memory for features and conjunctions. *Nature 390*, 279–281.

Maloney, L. and Landy, M. (to appear). When uncertainty matters: the selection of rapid goal-directed movements [abstract]. *Journal of Vision*.

Marr, D. 1982. *Vision*. W.H. Freeman and Co., Oxford.

Merleau-Ponty, M. 1962. *Phenomenology of Perception*. Routledge & Kegan Paul.

Miller, G. 1956. The magic number seven plus or minus two: Some limits on your capacity for processing information. *Psychological Review 63*, 81–96.

Newell, A. 1990. *Unified Theories of Cognition*. Harvard University Press.

Nilsson, N. 1984. Shakey the robot. Tech. Rep. 223, SRI International,.

Pashler, H. 1998. *The Psychology of Attention*. Cambridge, MA: MIT Press.

Pelz, J., Hayhoe, M., and Loeber, R. 2001. The coordination of eye, head, and hand movements in a natural task. *Experimental Brain Research 139,* 3, 166–177.

Roelfsema, P., Lamme, V., and Spekreijse, H. 2000. The implementation of visual routines. *Vision Research 40*, 1385–1411.

Roelfsema, P. R., P.S., K., and Spekreijse, H. 2003. Subtask sequencing in the primary visual cortex. *Proceedings of the National Academy of Sciences USA 100*, 5467–5472.

Sprague, N. and Ballard, D. Multiple goal learning for a virtual human. in preparation.

Sprague, N. and Ballard, D. 2003a. Eye movements for reward maximization. In *Advances in Neural Information Processing Systems 15*.

Sprague, N. and Ballard, D. 2003b. Multiple-goal reinforcement learning with modular sarsa(0). In *International Joint Conference on Artificial Intelligence*.

Suri, R. E. and Schultz, W. 2001. Temporal difference model reproduces anticipatory neural activity. *Neural Computation 13*, 841–862.

Sutton, R. 1996. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems*. Vol. 8.

Sutton, R. and Barto, A. 1998. *Reinforcement Learning: An Introduction*. MIT Press.

Terzopoulos, D. and Rabie, T. F. 1997. Animat vision: Active vision in artificial animals. *Videre: Journal of Computer Vision Research 1,* 1, 2–19.

Ullman, S. 1985. Visual routines. *Cognition 18*, 97–159.

VanRullen, R., Reddy, L., and Koch, C. 2004. Visual search and dual tasks reveal two distinct attentional resources. *Journal of Cognitive Neuroscience 16*.

von der Malsburg, C. 1999. The what and why of binding: the modeler's perspective. *Neuron 24*, 95–104.

Watkins, C. J. C. H. 1989. Learning from delayed rewards. Ph.D. thesis, King's College, Oxford.

Watkins, C. J. C. H. and Dayan, P. 1992. Q-learning. *Machine Learning Journal 8,* 3/4 (May).