Modular Models of Task Based Visually Guided Behavior

by

Constantin A. Rothkopf

Submitted in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

Supervised by

Professor Dana H. Ballard & Professor Mary M. Hayhoe

Department of Brain and Cognitive Sciences and Department of Computer Science Arts, Sciences and Engineering School of Arts and Sciences

University of Rochester Rochester, New York

2008
Curriculum Vitae

Constantin A. Rothkopf was born on May 2, 1969 in Kiel, Germany and grew up in Cologne, Germany and Florence, Italy. After coming back to Germany and obtaining his Highschool degree he took courses in Physics, Astronomy, Theater, Film, and Television Sciences, and Philosophy at the Universities of Cologne and Bonn. After successfully running his own company, he moved to Rochester, New York in 1999 where he obtained a Bachelor of Science in Imaging Science in 2003 with highest honor at the Rochester Institute of Technology and was awarded the Outstanding Undergraduate Scholar Award. He started pursuing a joint PhD degree in Brain and Cognitive Sciences and Computer Sciences at the University of Rochester in 2003. During this time he has continued exhibiting his artwork in galleries and museums.
Acknowledgments

I am deeply grateful to both my advisors Dana Ballard and Mary Hayhoe. The inspiration and the encouragement to critically evaluate ideas from a broad perspective have been invaluable and a source of inspiration throughout my studies. The caring and nurturing support has been in many ways beyond what one can expect from one’s graduate advisors. Thank you, Dana and Mary.

I would also like to thank David Knill for numerous conversations that have contributed to my understanding of the idea of tasks, which should still see their fulfillment in the collaborative work, that we have started. I also thank Alexandre Pouget for continuously challenging the participants of his lab meetings, which has been invaluable in developing ideas about neural coding.

I also have to thank the members of my committee, who not only played a role in the shaping of this thesis, but have also contributed to my education throughout my time as graduate student. Graduate school would not have been the same without the interaction with fellow students and postdocs. I would like to thank especially Bruno Averbeck, Neil Mennie, Jozsef Fiser, Melchi Michel, C. Shawn Green, Brian Sullivan, Jelena Jovancevic, Jason Droll, Rebecca Achtman, Jeff Beck, and Wei Ji Ma.

Finally I want to thank the Center for Visual Science and the Department of Brain and Cognitive Sciences at the University of Rochester. The collegiality, good mood, and scientific climate here have made graduate school a wonderful experience of exploration, collaboration, and fun.
Human behavior in extended visuomotor tasks is not well understood. This thesis considers the visuomotor task of navigating along a walkway while avoiding obstacles and approaching targets. Behavioral data of humans executing this task is presented together with a model of sidewalk navigation based on the reinforcement-learning framework. The connection between model and empirical data is made by using a new inverse reinforcement learning algorithm that estimates the parameters of the learning model so as to best match the observed human behavior. Thus, this work proposes to understand human visuomotor behavior in terms of learned solutions to specific tasks.

First, the analysis of behavioral data shows the limitations of current saliency-based models describing human gaze selection and quantifies the influences of task on gaze selection. The analysis furthermore demonstrates the similarity in walked trajectories and gaze patterns between subjects and how the similarities in behavior impose regularities on the input to the visual system, such as contrast and model simple cell response statistics.

If human vision is understood as an active process that has to learn how to select relevant information in time, then algorithms for the solution of complex visuomotor control tasks have to be developed. To this end, this thesis introduces a credit assignment algorithm for modular reinforcement learning that allows solving the same walking tasks in a virtual agent. The main feature of this algorithm is that it allows for multiple component tasks to learn their respective contributions to a single global reward signal.

Finally, in order to validate the proposed learning algorithm for the modeling of human side-
walk navigation an inverse reinforcement learning method is proposed that allows extracting
the relative reward weights implicitly attributed to the component tasks by the subjects. The
application of this method to the empirical data shows that this model can indeed be used to
describe well the human walking behavior in terms of the reinforcement learning model that
assumes a modular solution of the individual component tasks.
Table of Contents

Curriculum Vitae ii

Acknowledgments iii

Abstract iv

List of Tables ix

List of Figures x

1 Introduction 1

1.1 Contributions of the thesis .................................. 4

2 Background 8

2.1 Active vision .................................................. 8

2.2 Vision in natural tasks ......................................... 15

2.3 Models of natural behavior .................................. 17

2.4 Neural substrate of learning where to look ................. 19
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Task determines where you look</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>3.1 Introduction</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>3.2 Methods</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3.3 Results</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>3.4 Discussion</td>
<td>42</td>
</tr>
<tr>
<td>4</td>
<td>Task determines feature statistics at point of gaze</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>4.1 Introduction</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>4.2 Materials and Methods</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>4.3 Results</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>4.4 Discussion</td>
<td>68</td>
</tr>
<tr>
<td>5</td>
<td>A credit assignment algorithm for the learning of visuo-motor policies</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>5.1 Introduction</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>5.2 Background</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>5.3 Modular Behaviors</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>5.4 Credit Assignment with modular behaviors</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>5.5 Simulation results</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>5.6 Discussion and Conclusions</td>
<td>101</td>
</tr>
<tr>
<td>6</td>
<td>Extraction of human reward weighting</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>6.1 Introduction</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>6.2 Methods</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>6.3 Inverse Reinforcement Learning</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>6.4 Using prior knowledge of task structure</td>
<td>107</td>
</tr>
</tbody>
</table>
6.5 Biological constraints ...................................................... 109
6.6 Results ............................................................................ 110
6.7 Conclusion ........................................................................ 113

7 Conclusion ........................................................................ 117

7.1 Contributions of the thesis ............................................... 117

References ............................................................................ 123
List of Tables
List of Figures

3.1 Subjects view of the walkway from the starting position in three different conditions. Left: pickup purple litter and avoid blue obstacles condition with normal spatial distribution of objects on the walkway. Middle: same condition as in the left view but with a tighter distribution of objects. Right: normal spatial distribution of objects on the walkway with additional salient objects in the scene. .................................................. 23

3.2 Center: Typical view of the subject during execution of the task showing the walkway with blue obstacles and purple pickup objects, the cityscape in the background. Left: Schematic representation of the scene context represented by the color histogram of the scene, which shows the proportion of colors summed over the entire field of view. Right: Representation of the image patch at two fixation locations (see text). Note how gaze directed to the central region of a solid object is represented as a solid line whereas gaze directed towards an edge is represented as a succession of stripes. ............................................. 25
3.3 Visualization as introduced in figure 2 of the image patch at the center of gaze for subject ML in two different conditions. The x-axis corresponds to the normalized trial duration and the y-axis is foveal eccentricity in degrees. Top: pickup purple litter; Bottom: Avoid blue obstacles. The duration of the entire trial consists of the time the subject spent immersed in the environment and listening to the instructions, the time during the execution of the task, and the time during which the subject has finished the task and is waiting to start the next trial. The figure clearly shows the predominance of fixations on purple objects for the pickup task and a similar pre-dominance of fixations on blue objects in the obstacle avoidance task. ................................................................. 27

3.4 Comparison of the gaze over time for ten subjects labeled S1 to S10 using the visualization introduced in figure 2. Note that the sequences shown in figure 3 represent the entire trial duration while the sequences shown here are taken only from the time during the trial in which the subjects were moving. Left: Gaze visualization during the execution of the task Pickup purple objects. Right: Gaze visualization during the execution of the task Avoid blue obstacles. .................. 30

3.5 Proportion of fixation time spent on the object classes across subjects (n=19) for four different tasks. The shown proportions are colored according to the color convention depicted in figure 1, i.e. (1) purple represents pickup objects, (2) blue represents obstacles, (3) gray represents the walkway, (4) green represents the lawn and tree, (5) represents the background buildings and (6) light green represents the salient distractors. The diagrams were obtained by averaging over the two color conditions. From left to right: pickup purple litter, avoid blue obstacles, pickup purple litter and avoid blue obstacles, pickup purple litter and avoid blue obstacles with salient distractors in the scene. Error bars are 1 SEM across subjects. ................................................................. 32
3.6 Same representation of the fixation distribution of subject SG over time as in figure 3.3. Additionally, the color histogram of the scene in the field of view of the subject, as described in the methods section and shown in figure 2, has been plotted for each of the two trials. Top: The fixations and color histogram for pickup purple litter objects. Bottom: The fixations and color histogram for avoid blue obstacles.

3.7 Probability of fixating the pickup objects and the obstacles given the proportion of the respective object class in the current field of view. Left: proportion of fixation on a pickup-object (purple) and an obstacle (blue) given the proportion of pickup-objects in the current scene (0.25, 0.25-0.5, 0.5-0.75, 0.75-1.0). Right: proportion of fixation on a pickup-object (purple) and an obstacle (blue) given the proportion of obstacles in the current scene (0 1.0).

3.8 Comparison of proportion of fixations on object classes between the normal condition and the narrow condition in which the objects were placed only on the walkway. Top row: pickup and avoid in the normal condition. Bottom row: pickup and avoid in the narrow.

3.9 Visualization as introduced in figure 2 of the unrolled gaze for subject NM in the condition pickup litter objects and avoid obstacles when additionally a large number of salient objects are present in the scene.

3.10 Proportion of fixations on object classes for the condition in which a large number of additional salient objects were immersed into the scene. Left: Proportion of fixations during the execution of the task. Right: Proportion of fixations before and after executing the task.

3.11 Average fixation durations on object classes for the four conditions. From left to right: Pickup, Avoid, Combination Salient.
3.12 From left to right: An original scene from the sequence recorded from the interaction of the subject with the environment for which the saliency map was calculated. Conspicuity maps for color, intensity and orientation and the final saliency map after applying the winter-take-all competitive network to the sum of the conspicuity maps. 40

3.13 Comparison of gaze visualizations obtained from subject SG in condition pickup (top) with the visualizations obtained using a random gaze allocation (middle), and the described saliency model (bottom). 40

3.14 Comparison of proportions of fixation times spent on the object classes for human subjects (top), a random gaze allocation model weighted by the spatial gaze distribution (middle) and a saliency model (bottom). The four conditions are shown by column: pickup, avoid, combination, and salient conditions. 47

4.1 Nine representative views of the scene as seen by a subject during execution of the tasks. 52

4.2 The set of Gabor functions used in encoding the image. Left: The Unnormalized functions. Right: The same functions normalized individually for display. 57

4.3 Log-log plot of the rotational averages of the powerspectra of different image sets. The curves correspond to the slope of the rotational average for a set of natural images (green), the object texture in the VR experiments (purple), and multiple views from the subject's perspective in the VR experiments. Slopes of $-2$ corresponding to $1/f^2$ envelopes are shown as dashed lines. 58

4.4 Example average histograms of filter responses for a subset of the Gabor filters. The histograms were calculated for the filter responses to odd Gabors with spatial frequency of 1.6 cycles/degree on the set of 50 images obtained from the view of subjects walking down the walkway in the virtual environment. 59
4.5 RMS-contrast as function of image patch size for different image ensembles and tasks. Left: comparison of contrasts between image sets at the point of gaze for the conditions pickup, avoid, and the combination task. Right: comparison of contrast as a function of patch size for the avoid and pickup condition as in the left plot. Additionally, for the pickup condition the contrast for gaze falling on the litter objects and for gaze falling on the obstacles was plotted.

4.6 Histograms of model simple cell responses. Left: Normalized responses to vertical Gabor filter for condition pickup (purple) and avoid (blue). Right: Normalized responses to horizontal Gabor filter for condition pickup (purple) and avoid (blue). Note that the errorbars corresponding to the standard error around the mean are so small that they fall on top of the data points.

4.7 Histograms of model complex cell responses. Left: Normalized responses to vertical model complex cell for condition pickup (purple) and avoid (blue). Right: Normalized responses to horizontal model complex cell for condition pickup (purple) and avoid (blue). Note that the errorbars corresponding to the standard error around the mean are so small that they fall on top of the data points.

4.8 The set of 30 cluster centers obtained from the image representation using the Gabor filters for the set of 50 images obtained along the walkway.

4.9 The set of 25 cluster centers obtained from the image representation using the Gabor filters.

4.10 The set of 36 cluster centers obtained from the image representation using the derivative of Gaussians filter and their relative likelihood in the condition 'pickup' (top), 'avoid' (middle), and in all scenes (bottom).
4.11 Horizontal and vertical marginal distributions of gaze targets on litter objects in the pickup condition (left) and on obstacles in the avoid condition (right). These distributions were obtained using data from all 19 subjects. The plots with targets marked on the objects were obtained from the data of subject MB and are representative of the entire data set. 

5.1 The credit assignment algorithm exploits the constraint that during behavior, in any period there is only a subset of the total module set that is active. We term these periods episodes. In the timecourse of behavior, modules that are needed become active and those that are no longer needed become inactive. The diagram depicts two sequential episodes of three modules each \{3,4,7\} and \{2,8,10\}. The different modules are denoted with different shadings and numbers. The different lengths indicate that modules can exhibit have different numbers of states and finish at different times. The horizontal arrows denote the scheduler’s action in activating and deactivating modules. On the right is the large library of possible modules. Our formal results only depend on each module being chosen sufficiently often and not on the details of the selection strategy.

5.2 A fundamental problem for a biological agent using a modular architecture. At any given instant, shown with dotted lines, when multiple modules are active and only a global reward signal $G$ is available, the agents each have to be able to calculate how much of the rewards is due to their activation. This is known as the credit assignment problem. Our setting simplifies the problem by assuming that individual reinforcement learning modules are independent and communicate only their estimates of their reward values. The modules can be activated and deactivated asynchronously, and may each need different numbers of steps to complete, as suggested by the diagram.
5.3 Schematic representation of the grid world examples used in the simulations. From left to right: A) This example is taken directly from Sutton and Barto’s book where it is used in the single agent case. The agent is able to move in the four directions up, right, down, and left. Each transition results in a reward of 0 with the exception of the two states from where the arrows start. In those states any action transitions the agent to the shown states and a reward of either 10 or 20 can be collected. B) This grid world example follows Singh and Cohn An agent is located on a 5x5 grid and searches to find three different food sources f1 to f3 and tries to avoid the predator p, which moves every other time step towards the agent. C) This grid world problem is derived from the previous by increasing the size of the total area of the world within which the agent can only perceive food sources and predators within the perceptual range.

5.4 (Top left) Most of the Chang et al filtering agents converge to the optimal policy but a small number converges to a suboptimal policy and are not maximizing their average reward. (Top right) The root mean square error between the optimal policy found through value iteration and the policy found by the Chang et al filtering agents. Note that although the policy that an agent follows is optimal, the value function it found may have a very large error because of the random additive constant in the reward estimates. (Bottom left) Using the consumable rewards, agents converge to the optimal policies. (Bottom right) The sum of squared differences between the optimal and the found value function decrease to zero using consumable rewards.
5.5 Agents that assume that the variation in reward is noise and can be tracked with a Kalman filter have trouble when the mix of behaviors keeps changing. This is again evident in the collected rewards (left) and the RMS-error between the computed and the optimal value function (right). Consumable rewards has no such problem and in fact exploits the different mix of behaviors in order to track the reward accurately.

5.6 Effect of different learning rates on the accumulated reward and comparison to the variance-weighted learning as well as a Kalman Filter that has perfect state information about all component modules.

5.7 Average reward collected using consumable rewards. Simulations were run in which three food sources from a set of 75 and one predator from a set of 25 were always active. The three correspond to different learning rates $\beta$ in equation 5.15.

5.8 Sum of squared differences between the true rewards and the reward estimates of all behaviors over trials. The three curves correspond to different learning rates $\beta$ in equation 5.15.

5.9 The walkway navigation tasks. Left: typical view from the agent while navigating along the walkway. The three possible tasks are following the walkway, avoiding obstacles, which are the dark cylinders, and picking up litter, corresponding to the light cylinders. Right: Schematic representation of the state-space parameterization for the learners. Each module represents the distance to the closest object in the field of view and the angle between the current heading direction and the object center axis. The module learning the walkway behavior uses the signed distance to the midline of the walkway and the angle between the heading direction and the vector in the direction of the walkway.
5.10 Representations of rewards, value functions, and policies in the walkway navigation task for the three component behaviors at iteration 1 and iteration 1000. Top row: reward estimates \( \hat{r}(s) \). Middle row: value functions \( V(s) \). Bottom row: policies \( \pi(s) \). From left to right. First column: iteration 1 for obstacle avoidance. Second column: iteration 1 for litter pickup. Third column: iteration 1 for walkway following. Fourth column: iteration 1000 for obstacle avoidance. Fifth column: iteration 1000 for litter pickup. Sixth column: iteration 1000 for walkway following.

6.1 Comparison of the value functions for three component learners with different reward weights. The array of figures displays the value functions that were obtained through simulations in the walkway domain using reward weights of \( r=0.5 \) for the left column, \( r=1.0 \) for the central column, and \( r=2.0 \) for the right column. The three component value functions shown here are for the litter picking up behavior in the top row, the obstacle avoidance in the center row, and the walkway following in the bottom row.


6.3 Example data of a human subject executing the task ‘pickup litter and avoid obstacles’ using the parameterization of the state space for three component behaviors. Top: Representation of the distance and angle to the closest litter object during navigation along the walkway. Middle: Representation of the distance and angle to the closest obstacle. Bottom: Representation of the signed distance and angle between the subject’s walking direction and the midline of the walkway.
6.4 Comparison of the actual human trajectory (blue) for subject AF executing the combination task of following the walkway, avoiding obstacles, and picking up litter with four model trajectories obtained with different reward weights for the component tasks (black). The reward values are reported for the walkway following, the litter pickup, and the obstacle avoidance. From top to bottom: (0, 1, 0), (0, 0, 1), (1, 0, 0), (0.10, 0.12, 0.78) . . . . . . . . . . . . . . . . . . 115

6.5 Estimated normalized reward weights for the litter picking up behavior (purple) and the obstacle avoidance behavior (blue) for the 10 subjects in the task conditions 'pickup litter' and 'avoid obstacles'. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 116
1 Introduction

In order to address in a scientific way the question of how the brain accomplishes vision, research has branched out into a multitude of directions asking different specific questions and using different methodologies. Behavioral investigations have characterized human and animal activity in a vast variety of tasks ranging from visual navigation in ants (Collett et al., 1998) to visual strategies in observing other humans during task execution (Flanagan & Johansson, 2003). Neurophysiological studies have characterized responses of single neurons in primates during visual decision making (Platt & Glimcher, 1999) and of neuronal ensembles in monkeys while watching natural stimuli (Vinje & Gallant, 2000). Neuropharmacological research has identified neurotransmitters required for perceptual learning (Dinse et al., 2003) and discovered neuropeptides that modulate the level of trust in decision making between humans (Kosfeld et al., 2005). Computational models have been able to formulate models that extract the same visual features from the environment as primate visual areas (Olshausen & Field, 1997; Rao & Ballard, 1999) and demonstrated remarkable similarity with the neuronal signals involved in learning actions required for obtaining delayed rewards (Schultz et al., 1997).

Nevertheless, it is not the case that this gained knowledge has led to the development of machines that demonstrate the same agility and resourcefulness in executing visual tasks as humans. Despite recent successes in which autonomous agents were able to drive a car in desert and urban environments (Thrun et al., 2006), these solutions have been highly specific
to their domains, and inverse Turing tests in form of CAPTCHAS are still an effective way of testing for human visual capabilities. Furthermore, these artificial agents were equipped with a wealth of specialized sensors such as GPS and laser range finders which go far beyond human natural sensory equipment. As artificial computation is less and less expensive, such specialized solutions may become more available. But it is not clear to which degree these solutions are informative of the computations human brains actually execute, given that the constraints and the hardware for these agents are so different. In fact, it is still a matter of debate, how to ask the correct questions that address how humans accomplish vision. A good example for such a situation is how research addresses the selection of gaze targets. While not exclusive, a major distinction in the literature has been between research that has tried to characterize gaze as a reactive perceptual process driven by local properties of the image of the environment on one side and the emphasis of the active and predictive aspects of visual perception.

How these questions are asked is dependent on the conceptualization of what vision is, how it is implemented, and what it is for. Is perception a separate and independent process that results in the extraction of general purpose representations that can be used by 'later' stages in the processing stream such as decision-making stages? Or is perception essentially impossible without an embodied agent that acts in the world? While early research in artificial intelligence has emphasized the aspect of manipulating representations at an abstract level and the extraction of abstract general purpose representations that can be manipulated using entailment relationships, current work has been influenced by the statistical developments in artificial intelligence. See (Grush, 2002) for a historical elaboration on these developments.

A fundamental concept that has become essential in cognitive science research is that of computation. That the human brain and vision have to be understood in the framework of information processing has become the common ground in current research. By adopting this view, computational models have been successful at explaining behavioral data from a normative perspective. For example, the conceptualization of perception as Bayesian inference (Knill & Richards, 1996; Rao et al., 2002) is directed at explaining perceptual phenomena as the result
of vision attempting to infer the most probable causes of the sensory observation by integrating prior knowledge and the probabilistic relationships between such causes and image features. This framework has explained a large variety of phenomena ranging from motion perception (Simoncelli, 1993) to visuomotor learning (Körding & Wolpert, 2004) and has determined the search for how such inference could be implemented by the brain (Doya et al., 2007).

A second fundamental concept that has influenced research in vision is that of evolution. Environmental changes drive adaptation processes that decide whether a species will survive, and this development can be formalized as an optimization process. Depending on the plasticity of the nervous system, such optimization can take place at different rates both on the phylogenetic as well as on the ontogenetic level. This optimization process can be assumed to never have converged completely, not only because current research can only take a snapshot at the current stage in this process, but also because the continuous change in the environment changes the constraints of this optimization process (Rosen, 1970). Considerable work has gone into specifying concretely what may be optimized and how. The ideas range from the search for a general principle explaining sensory coding by information maximization strategies to the developments of heuristics that are computationally cheap and good enough for survival while the optimization models range from gain adjustments to genetic programming.

The consequences for the study of human vision are that it should be studied in the context of what it has evolved to do and should address computational principles that take into account, what vision is used for (Ballard, 1991). Thus, vision in this thesis is conceptualized as the perceptual process humans are engaged in when executing every-day extended visuomotor tasks such as walking along a walkaway, preparing a cup of tea, or making a sandwich. Under such circumstances they execute sequences of actions involving their eyes, their head, and their entire body. The psychophysical literature contains compelling evidence that such complex behavioral sequences are composed by stringing together elemental task solutions in space and time (Hayhoe, 2000; Land, 2004), and the differences in visuomotor performance between novices and experts in a variety of tasks has been demonstrated to be learned (Land & Furneaux,
Furthermore, there is evidence from neurophysiology in nonhuman primates, that the basal ganglia are involved in the representation and learning of sequential behavior (Graybiel, 1998) and that this learning is mediated by reward (Hollerman & Schultz, 1998). Such learning has been shown to agree well with models of reinforcement learning (Schultz et al., 1997). Based on these empirical results and theoretical models, this thesis follows this conceptualization of visual perception as an active process directed at extracting specific information from the environment as required by the ongoing task.

1.1 Contributions of the thesis

This thesis aims at better understanding how exactly tasks exert their control on gaze in natural tasks. Behavioral, computational, and theoretical methods have been employed in order to approach this question on different levels of description. The particular task that this thesis uses to investigate human visuomotor behavior is that of walking along a path while avoiding obstacles and approaching targets. These types of fundamental tasks are executed by people on a daily basis and can be regarded as evolutionarily relevant for survival.

First, this thesis presents results from psychophysical experiments further demonstrating that human gaze during natural behavior is highly task dependent. Participants navigated in a virtual environment and were instructed to follow a walkway and avoid obstacles as well as approach targets. This study revealed significant differences in behavior between tasks and high similarity across subjects within task conditions. This investigation is particularly relevant, because normative models of gaze have been previously proposed and can be tested in these experiments. The two most relevant normative models of gaze allocation tested here are the saliency model proposed by (Itti & Koch, 2000) and the model proposed by (Sprague et al., 2006). The saliency model is bottom-up model that does not incorporate any notion of the goal of a fixation and accordingly fails to model the observed data. While the model by (Sprague et al., 2006) incorporates explicitly the goal of behavior in human vision and gaze selection
it fail to explain the gaze distribution observed in the presented tasks quantitatively and the empirical results demonstrate new aspects of the regularities of gaze sequences across subjects, especially the influence of scene context.

Saliency models conceptualize vision as a passive process and postulate that gaze is attracted to regions in the image dependent solely on the image properties. This has lead to the analysis of the statistics of image features at fixation location in the past. Such statistics were obtained for the navigational tasks considered here. These feature statistics are quantified and are used to test previous models of gaze behavior. Results are reported that contradict previous conclusions based on free-view and search tasks, in that in one particular task condition of approaching targets, the contrast at fixation location was reduced compare to contrast at randomly selected parts of the scene. This demonstrates that it is not the case that particular visual features intrinsically always attract gaze, but that the statistics of features are task dependent.

If the analysis of visual features at fixation location is not sufficient to determine what is being processed, the question arises whether there are other possibilities to infer the cognitive state of the system during a behavioral sequence. This is particularly relevant, because a previous normative model by Spargue and Ballard (Sprague et al., 2006) made quantitative predictions about the internal states by observing which individual information gathering process had access to the perceptual system, but did not model which visual features would be used by vision. This means, that the proposed system makes predictions about which internal cognitive process is active at each moment in time, but not where gaze is directed to. While there is a vast literature in the computer science community about fitting behavioral data with a wide variety of statistical models, the focus in this thesis is on model selection. Here, crossvalidation and Bayesian model selection techniques are used to infer the most likely distribution of hidden internal cognitive states from the observed psychophysical data, such as the direction of gaze and from state variables describing the environment, such as the distance and angle between the subject’s heading direction and objects in the world during walking. It is shown, that a small number of such internal states can capture the observed data succinctly.
These behavioral results further demonstrate, that task based models of visuomotor behavior need to be constructed. One considerable issue is how map such task based behavior onto known learning and action selection substrate in the human brain. A contribution of this thesis is the development of a credit assignment algorithm for the learning of multiple tasks, when these are executed concurrently. Using the framework of reinforcement learning, it is shown how to learn component value functions when only a global reward signal is available, as is observed in primates and other mammals. This algorithm exploits the fact that different task combinations may be executed at different times and uses empirical data on the metabolic energy consumption in walking as the common currency for all learning aspects. It was applied to a formulation of the navigational task based on previous work by Spargue and Ballard (Sprague et al., 2006) and it is demonstrated that it is indeed feasible for learning to coordinate the visual behaviors.

In order to validate the proposed models based on reinforcement learning, it is important to reconnect the parameters of the models with the experimental data. This is not trivial when considering natural tasks, because participants are not restricted to a small number of actions but instead execute long sequences of actions which require high dimensional data for their description. Similarly, the statespace describing the state of the environment is very high dimensional. Using the proposed modular decomposition from the reinforcement algorithm, this thesis demonstrates how to extract parameters relevant for the reinforcement learning models from observing human subjects executing the navigational tasks, using a series of constraint that allow to reduce the complexity of the estimation problem. These results show that human trajectories while navigating through the environment are well described by the value functions found by the reinforcement learning algorithm. Furthermore, the extracted parameters of the model quantify the differences in behavior between tasks and the similarities across individual subjects.

In summary, this thesis investigates human visuomotor behavior of navigating towards targets, avoiding obstacles, and following a path. The human experiments are carried out in a virtual environment, which allows for the control of the task structure. Further evidence for
the strong influence of tasks on visual behavior is demonstrated, and it is shown that previous models of gaze allocation either fail to reasonably describe gaze or need further improvement. The visual features at the point of gaze are quantified in their statistical structure revealing that the input to the visual system undergoes significant changes on the timescale of minutes. This further demonstrates that eye movements are not directed reactively to points in the scene based on their low level features alone and this suggests that the representational processes in vision have to take this variations due to different tasks into account. How executing such tasks optimally could be learned in a biological agent is demonstrated through a proposed reinforcement learning algorithm that solves the credit assignment problem when multiple concurrent tasks have to be solved while being executed asynchronously. Finally, it is shown that this type of reinforcement learning based model can explain well the navigational behavior of the human subjects observed in the experiments and that the reward implicitly associated with the different tasks as executed by humans can be recovered. This demonstrates that reinforcement learning based algorithms are able to describe and predict human behavior in natural, goal-directed, ever-day visuo-motor tasks.
2 Background

2.1 Active vision

One of the essential properties of the human visual system is its active nature (O’Regan & Noe, 2001; Ballard, 1991; Ballard et al., 1997; Findlay & Gilchrist, 2003): humans shift gaze between locations in the visual world on the order of three times per second, resulting in more than 150,000 a day. This means that the visual system executes an action in order to select a target for perception. In order to understand vision, it is therefore essential to study gaze deployment. Where is the gaze directed to in a scene, what computations are performed at the gaze point?

The fact that human vision is foveated is often seen as a solution to the constraint imposed by the processing power of the brain. If the entire field of view had to be represented with the same acuity present in the fovea, the human brain would require a volume of over a cubic meter to store the neurons required for visual processing and the metabolic cost for processing would be enormous (Lennie, 2003). Because of the resulting foveation, gaze has to be moved to relevant parts of the scene. The animat vision framework put forward by Ballard (Ballard, 1991) extends this view by proposing to understand foveated vision as a solution to computational constraints: "Gaze control mechanisms fundamentally change computational models of vision. Without them the visual system must work in isolation, with the burden of solving difficult
problems with many degrees of freedom. With them a new paradigm emerges in which the visual calculations are embedded in a sensory-motor behavioral repertoire. Rather than thinking of visual processing as separate from cognitive or motor processing, they are interlinked in terms of integral behaviors.”

This fundamental distinction between the location of gaze and the computations being executed while directing gaze to a location, i.e. the difference between looking and seeing, are reflected in two fundamentally different approaches to vision and gaze control in particular. Explanations of the allocation of gaze within a scene in the behavioral literature have usually emphasized either bottom up saliency (Itti & Koch, 2000; Parkhurst & Niebur, 2003; Torralba et al., 2006), which is independent of the computational goal, or top-down cognitive control (Buswell, 1935; Yarbus, 1967; Land, 2004; Hayhoe & Ballard, 2005) in which the behavioral objective is stressed. These two approaches have emphasized different ways of thinking about vision and have therefore produced different experimental paradigms and different explanatory frameworks. While the former is directed more towards a mechanistic implementation the latter addresses the question of the goal of vision (Marr, 1982).

2.1.1 Stimulus driven vision

The bottom-up saliency assumption is based on the hypothesis that certain features of the visual scene inherently attract gaze, i.e. that vision is essentially reactive and stimulus driven. This view is in part based on psychophysical studies (Mannan et al., 1996; Reinagel & Zador, 1999; G. et al., 2000; Tatler et al., 2005) in which differences in image properties were observed between fixated and randomly chosen locations. The initial fixations when viewing a large number of 2-dimensional photographic images are recorded and statistics for several features are compared between the gaze locations chosen by the subjects with those of randomly chosen parts of the scene. Such studies have produced mixed results regarding which features are significantly different at fixation locations. While some re-searches found that luminance contrast was elevated at the point of gaze (Parkhurst & Niebur, 2003) other studies found the
edge density was significantly stronger at fixation locations (Mannan et al., 1996; Baddeley & Tatler, 2006). Subsequent models have been proposed that relate the bottom-up assumption to neuronal processing in cortical visual areas. Starting from the notion that ‘early’ visual areas represent low level features such as oriented edges, such models have extracted analogous features from images and proposed methods by which a scalar saliency map could be calculated (Koch & Ullman, 1985a; Itti & Koch, 2000). These methods apply different forms of center-surround competitive algorithms in order to find regions of across-scale contrast within single feature dimensions, and proceed to combine multiple maps to a single saliency map (Itti, 2000) by some weighting technique. Such feature saliency based models have a large number of free parameters, which have to be adjusted in order to obtain meaningful saliency maps. It is necessary to choose the number of filters, their respective parameters such as orientations and spatial frequencies as well as the spatial scales, the normalization functions, the summation rules, and the parameters of the network implementing the spatial competition within the saliency map.

Bottom-up models have recently implemented contextual effects and top-down effects (Itti, 2000; Navalpakkam & Itti, 2005; Oliva et al., 2003; Torralba et al., 2006), but these effects have been described as modulating the saliency map and have been restricted to the tasks of object detection or object recognition. (Navalpakkam & Itti, 2005) proposed such a model in object detection. Starting from key words specifying a task, their saliency map is biased towards known image features of the corresponding target object. Unfortunately, the authors do not specify how such a database is to be constructed or how such features could be determined reliably across different environmental states such as illumination, point of view, and clutter, just to name a few. Torralba also considered an object detection task and proposed a spatial modulation of the saliency map by an additional map representing the likely positions of target objects in a scene. Such a map was obtained by training a supervised algorithm on a database of images in which the likely positions of certain objects have been hand-labeled. Feature vectors were calculated for the image by obtaining the convolution with oriented filters at multiple
spatial scales and reducing the dimensionality of these feature vectors. The motivation for this approach is that the statistics of image features can provide a clue to more likely positions of particular object locations. The rationale is that there is a similarity in certain feature dimensions for locations in which a particular target is likely to be found and that humans extract such regularities through experience.

Experimental work has assessed how well such bottom-up models can predict human gaze. Most of the saliency modeling extracting features such as orientation, luminance, and color opponency have used the so called freeview task for the assessment of the accuracy of the prediction of such models (Parkhurst & Niebur, 2003; Itti, 2005; Tatler et al., 2005). The instructions for this task have subjects examining a scene without further instructions. The free-view task is unfortunately very uncontrolled. It is not clear what the subject is looking for or what the subject assumes about the setup that the experimenter designed. Given that most of these studies have used different metrics comparing the saliency maps with human fixations, it is difficult to assess how well the models describe human gaze. In summary, the first few gaze locations selected by humans when free-viewing fractal images or finding likely target locations in a set of similar images correlate somewhat with the predictions of bottom-up saliency models. However (Einhäuser & König, 2003) randomly altered parts of natural scenes in their luminance contrast and the same analysis of the features at fixation locations were executed. Their study showed that the bias to fixate regions of high contrast disappeared after such manipulations. Thus increased contrast at fixation location may be correlational in nature and not causal. Also, in a recent study by (Henderson et al., 2006) the locations selected by gaze correlated not only with high local contrast but also semantic content related to the task so that the authors concluded that the relation between gaze and bottom-up saliency is correlational and not causal. All these reservations with bottom-up modeling serve to motivate the study of top-down gaze control, which by itself has a significant history.
2.1.2 Task based vision

The influence of tasks on gaze control has been described since the experiments by Buswell and Yarbus (Buswell, 1935; Yarbus, 1967). These effects become even more prominent in tasks that are not based on picture viewing but study subjects actively involved in interacting with the environment when executing goal directed behavior. Task-based models are required to model which features are in fact more likely to be fixated dependent on the particular task at hand and how semantic content influences the gaze strategy. In fact, it is well established that the ongoing task influences the gaze strategy. As soon as the visual scenes become meaningful, the objects are related by semantic meaning, or prior knowledge can be utilized in order to find likely locations for targets, these factors become significantly better predictors of the direction of gaze (Henderson & Hollingworth, 1999). The allocation of gaze has been studied in tasks such as copying arrangements of blocks (Ballard et al., 1995), making tea (Land et al., 1999), making sandwiches (Hayhoe et al., 2003), driving (Land & Lee, 1994) and other goal directed behavior. See Land (Land, 2004) and Hayhoe & Ballard (Hayhoe & Ballard, 2005) for reviews.

A recurrent theme of task-based studies is the functional relation of gaze to the ongoing behavioral sequence. This means that the direction of gaze during the execution of such a task can be predicted significantly better by the phase of the action sequence than by local features of the visual scene. In the game of cricket (Land & McLeod, 2000), the gaze is temporally linked to relevant events in the sequence of occurrences that are relevant to hitting an approaching ball. Under such circumstances, gaze may be directed to parts of the scene that are not distinguishable from other parts by any features other than that the subject is expecting a ball to move towards that location. In a block-sorting task in which subjects watched a person stacking a set of blocks, gaze was predictively directed towards expected points of interaction (Flanagan & Johansson, 2003). In experiments carried out by Johansson (Johansson et al., 2001), subjects fixated relevant spatial locations as necessary for the ongoing hand movements with high temporal repeatability depending on the features of the manual interaction. In another block copying experiment (Ballard et al., 1995) subjects arranged colored blocks from
a resource area according to a pattern shown in a model area. The observed eye movements showed a regular sequential pattern within the evolution of the task that could be interpreted in terms of momentary information processing needs. Fixations were directed to the model to obtain a blocks color, then to a corresponding block of that color in the resource area, and then back to the model to get its position in the pattern, followed by a fixation towards the location in the work area to which the block was subsequently moved.

Gaze has also been interpreted functionally in simple visuo-motor tasks such as pointing and grasping. One commonly observed gaze strategy of subjects in picture viewing is to direct the initial fixation towards the center of the extended objects (He & Kowler, 1991) and to similarly target the center of gravity in cluttered scenes. In pointing, gaze and hand are often tightly correlated and directed towards the point of contact (Frens & Erkelens, 1991). Recently Brouwer et al. (submitted) have investigated the difference between where subjects gaze when looking at an object and where they gaze when grasping them and found that subjects tended to gaze at the future contact point of the index finger. Johannson (Johansson et al., 2001) described a complex gaze behavior in a more extended task in which subjects grasped objects, moved them to a target location, and sometimes avoided obstacles along the way. All these results point towards the fact that in extended behavior in which subjects execute actions, gaze is tightly linked to the ongoing demands of the executed task.

2.1.3 Visual attention and gaze

In the behavioral literature, gaze control is often described in conjunction with visual attention. An often noted problem with the concept of visual attention is that it has been used in the literature in many different ways, as it is already reflected by the statement of William James (James, 1890) that "everyone knows what attention is". The emphasis in this thesis is on gaze location and tasks and only implicitly on attention. The reason for this is, that the literature on visual attention is mostly tied to a particular methodology that results from a particular conceptualization of vision. Most attentional studies consist of at least two conditions in which
a task is executed. Differential performance on the task as assessed by a performance measure on this task is used as evidence of the presence or absence of attention. This thesis by contrast does not measure the performance of a participant in form of a reaction time, or the number of items tracked, or on a concurrent dual task in the traditional form. Therefore, attention is only implicitly addressed by the subject’s direction of gaze. Although overt shifts of attention involving eye movements can be dissociated from shifts of covert attention (Posner & Cohen, 1984), it has been shown that voluntary saccadic eye movements involve a preceding shift of attention (Kowler et al., 1995; Deubel et al., 1997; Henderson, 2003). Therefore, under most circumstances, the direction of gaze reflects ongoing computations and can be used to infer the moment-to-moment cognitive processing that subjects are engaged in (Tanenhaus et al., 1995; Liversedge & Findlay, 2000).

While it is clear what particular point gaze is directed to in a scene, the inference about what is processed at each moment in time is not so easily accessible. That is, if a subject fixates a particular object in a scene, it is not clear which features are being processed. Although human self-awareness seems to suggest a continuous perception of object identities and features, considerable evidence demonstrates that this may not be what is in fact represented (Triesch et al., 2003; Droll et al., 2005; Simons & Rensink, 2005; Paprotta et al., 1999).

2.1.4 Neural correlates of task based vision

In addition to the psychophysical results, an abundance of recent neurophysiological results demonstrate how cortical areas involved in representing signals involved in the planning and execution of eye movements are closely linked to the expected temporal delivery of a reward within the task (Glimcher, 2003; Schultz, 2000). Moreover, the functionality of so called 'early' perceptual areas have been demonstrated to reflect the behavioral state of the organism. Two recent reviews (Olshausen & Field, 2005; Kayser et al., 2004) collect a number of arguments and empirical results that demonstrate that characterizations of V1 cells with task-neutral white noise stimuli do not adequately predict their responses to natural stimuli in awake behaving
animals. In addition, several results show effects of the task on the activity of neurons as early as area V1. For example, in a series of experiments by (Li et al., 2004), monkeys were trained on a bisection-task and a Vernier acuity-task. The stimulus was the same in both tasks but the monkey was cued which task to execute. The activity of V1 neurons as assessed by obtaining their tuning curves was dependent on the task the monkey was involved in. Moreover, the tuning curve switched dependent on the task on a time scale of the duration of trials. This is relevant to the modeling of biological mechanisms of visual attention, because it demonstrates that the processing of visual stimuli in cortical regions that are often described as being 'early' stages is active and dependent on the current behavioral goal of the agent.

### 2.2 Vision in natural tasks

In view of this evidence for the central role of the ongoing task for the allocation of gaze, it is necessary to quantify the influence of tasks on vision and develop a theoretical framework describing where gaze is directed in complex extended tasks. A goal of such task-related models is to be able to formulate cost functions that can direct gaze in complex tasks. Trommershüser (Trommershäuser et al., 2003) have demonstrated that such cost functions can describe human visuo-motor behavior accurately in a manual reaching task in which the reward structure was made explicit through monetary rewards. Nelson (Nelson & Cottrell, 2007) showed how gaze is directed to most disambiguating parts of an object in a shape-learning task. Sprague and Ballard (Sprague et al., 2006) devised a model of gaze control for an extended visuo-motor task with multiple competing goals. They propose that control of gaze can be understood in the setting of a minimum-loss-strategy implemented using reinforcement learning. Starting from the notion that the visual system is actively seeking to extract specific information from the visual array, a model was proposed that gives an account of why eye movements could be directed towards certain parts of the scene, given the task demands. The agent is faced with a representation of the visual scene that enables it to extract relevant information about
the state of the objects in the world. This model incorporates specific representations of the locations of the relevant objects in the scene. The eyes are directed towards relevant parts of the environments in order to update these internal representations. The crucial point is that concurrent demands are imposed on the system because the uncertainty about the environment increases with time, if the relevant information is not updated. In the proposed model, the eyes are moved towards parts of the scene where information is available to minimize the loss of reward.

2.2.1 Quantifying task dependence

Gaze location does not uniquely specify the information being extracted. Note that attempts to consider task effects in the context of saliency models do not distinguish the ongoing computations whereas task-based models explicitly contain and represent the difference between directing gaze towards a particular part of the scene and the information being extracted. Thus there is the need to clarify and quantify to what degree these task-based models are accurate descriptions of human behavior and to study human gaze behavior in dependence of the ongoing tasks. If vision is an active process that is dependent on the behavioral goal of the organism, vision needs to be studied when humans are engaged in purposeful, goal directed behavior. Naturalistic tasks are therefore adequate to address the question how vision is used. Accordingly, two major challenges have to be addressed: 1) the design of goal directed experiments in natural environments and 2) techniques for measuring and describing goal-directed extended behavior.

First, natural task is a loose description for tasks that fulfill several criteria. It is important that subjects are involved in such activities in everyday life and are not confronted with such a task for the first time when executing an experiment in a laboratory. In a visual search task the stimuli may be artificial to probe feature searches (Wolfe, 1998), or may be more naturalistic, i.e. they may be photographs of natural scenes (Mannan:1996, Reinagel-E-P.:1999, Krieger-G:2000, Tatler:2005). But the task itself can be natural too such as in the
previously mentioned studies studying gaze in tea making (Land et al., 1999) or sandwich making (Hayhoe et al., 2003). In an analogous fashion, human subjects can be highly engaged in computer games that use only very crude geometric representations of the world. Another criterion is that natural tasks are often trained over a long time and therefore do not require a high degree of learning when the subjects start an experimental session. By using virtual environments it is possible to obtain a balance between the required complexity and enough control over the visual scene in order to change it parametrically.

The second major aspect is how to quantify and analyze natural behavior. Complex environments such as a kitchen or a sports court contain numerous objects and visual stimuli that change over time. Additionally, subjects may be interacting and changing the environment in real time. The dimensionality of the stimulus or the state space describing the environment is very large. Similarly, complex behavior such as movements through the world, gaze behavior and interactions with objects in the scene inherently are of a high dimension too. Accordingly, tools have to be developed in order to represent these high dimensional data for representation and analysis.

2.3 Models of natural behavior

Understanding the execution of extended every-day activities that humans engage in every day will most probably require developing models of high grade of complexity. The only successful path to handling this complexity seems to be to resort to some kind of hierarchical decomposition. In a hierarchy, a complex system has its overall functionality compartmentalized in different levels such that each level is tractable. Newell argued that such decompositions were essential to manage the complexity of biological systems (Newell, 1990) and (Churchland & Sejnowski, 1994) have argued that such a decomposition is essential to understanding brain complexity. Given the importance of hierarchies in silicon computing, it would seem likely that a computational understanding of the brain also may depend on some sort of hierarchical de-
composition. One general approach to finding a decomposition is to have it result from a mathematical optimization. While the ultimate model make take this form, so far this route has proven very difficult and progress has been modest.

Another route is to define the decomposition ab initio and explore its properties. Thus far however only a few such computational hierarchies have been proposed for the brain, despite its standard use in silicon computer software. Brooks (Brooks, 1986) has made one of the first proposals centered on the concept of subsumption wherein more abstract behaviors can be modeled as modifications to existing behaviors. This has worked impressively for small systems but so far has not successfully scaled to large primate-like systems. We have previously proposed a decomposition that contains an abstract behavior level where individual behaviors compete to be activated (Sprague et al., 2006). The hypothesis is that, during small temporal intervals, only a small number of behaviors need to be active to model human performance. Complex behavior is achieved by changing the composition of this small set of behaviors at the timescale of the duration of different tasks in order to meet demands.

If behaviors compete with each other, then there has to be a way of adjudicating the competition. This would be possible if there was a scalar measure of individual behaviors that could make them commensurate. Evidence from a great number of different studies suggests that this measure is provided in the form of the neurotransmitter dopamine. In particular, dopamine has been shown to mediate expectation of reward in primates (Schultz, 2000).

The use of scalar rewards is a central feature of reinforcement learning algorithms. These were pioneered by Sutton and Barto (Sutton & Barto, 1998). The version most applicable to biological systems is on-line and known as temporal difference learning. Algorithms based on temporal difference learning have been proposed as a model of bee learning (Montague et al., 1995), also in human systems (Schultz et al., 1997). Subsequently, richer models have explored how reinforcement learning algorithms could be implemented in the brain. Actor-critic models have played a particularly prominent role, because neuronal activity in the Basal Ganglia has been demonstrated
2.4 Neural substrate of learning where to look

An important aspect of gaze control in natural vision is recent work that reveals sensitivity to expectation of reward in many of the regions involved in saccade target selection and generation. This forms a critical substrate for explaining task-directed eye movement patterns. Saccade-related areas in the cortex (LIP, FEF, SEF and DLPF) all exhibit sensitivity to reward (Platt & Glimcher, 1999; Glimcher, 2003; Dorris & Glimcher, 2004; Platt; Sugrue et al, 2004; Stuphorn et al, 2000; Hikosaka). These areas converge on the caudate nucleus in the basal ganglia, and the cortical-basal ganglia-superior colliculus circuit appears to regulate the control of fixation and the timing of planned movements, by regulating tonic inhibition exerted by the substantia nigra pars reticulata on the superior colliculus. Such regulation is a critical component of task control of fixations. Hikosaka and colleagues have demonstrated that caudate cell responses reflect both the target of an upcoming saccade and the reward expected after making the movement (Hikosaka et al, 2000; Watanabe et al, 2003). Since some kind of sensitivity to reinforcement is necessary for learning, and saccadic eye movements demonstrate such sensitivity, the neural substrate for learning where to look in the context of a task is potentially available in the basal ganglia (Hikosaka 2006). The reward sensitivity of the eye movement circuitry provides a basis for reinforcement learning models that are necessary for understanding how these elemental processes are organized to compose the complex gaze patterns observed in everyday behaviors.

Sprague & Ballard (2007) show how a model of walking in a virtual environment can learn to allocate gaze sequentially to areas in the environment important for walking. To choose between ongoing competing tasks such as avoiding obstacles and controlling direction of locomotion, in their model, uncertainty increases (together with cost) when gaze is withheld from an informative scene location. Fixation is allocated to the task that would have the greatest cost if the relevant information were not updated. They show that such a cost is calculable within the reinforcement learning framework described by Schultz and others (Schultz, 2000; Montague et al, 2004). Although it has not been demonstrated directly that the acquisition
of information per se is directly rewarding (but see Deaner et al, 2005), all complex behaviors involve secondary reward of some kind, and the acquisition of information is always a critical step in achieving behavioral goals. When reward is externally defined, it has been shown that subjects making fast aiming movements learn a complicated spatially distributed target reward system and behave in a nearly optimal manner to maximize reward (Trommershauser et al, 2003). There is only limited evidence, however, of the sensitivity of natural eye movements to the reward structure of the environment (though saccade reaction times are known to be sensitive to stimulus probability).
3 Task determines where you look

3.1 Introduction

In order to better understand human behavior in natural visuomotor tasks, a set of experiment was devised in which human subjects executed combinations of navigational, approach, and avoidance behaviors.

3.1.1 Specific aims

This chapter presents psychophysical results that address both of the above issues. The tasks considered here are extended visuomotor tasks that require goal directed interaction with the environment. Human subjects navigate through a VR environment, approach and avoid objects while executing sequences of approximately 500 fixations per trial on average. By having subjects execute single tasks and the task combination, the differences in the allocation of gaze as a function of the respective gaze can be observed. These differences are then quantified more precisely. The experiments presented here were devised in order to compare how well task based models and bottom-up saliency models can explain the distribution of fixations observed in extended tasks, which can not be conceptualized as being a search task or an object detection task. Relevant parameters of behavior that can give clues about the ongoing processing include fixation location, fractional gaze allocation, and fixation durations. Related to this question, is
how fixations are dependent on object identities, visual features, and task features. Specifically, the experiments address whether the location of a fixation on an object depends on the task that is executed, how prominent a role distractors play when executing extended visuomotor tasks, how the fixation duration on objects depends on the task, and how global scene context affects fixation distributions. Moreover, by examining single tasks and their compositions, the question can be addressed, how behavior in the combined task is related to the component tasks.

3.2 Methods

3.2.1 Experimental setup

Subjects were immersed into a virtual reality environment consisting of a cityscape (Performer Town) created by SGI. They wore a Virtual Research V8 head mounted binocular display. The resolution of the stereo LCD screens in the headset was 640 by 480 pixels corresponding to a horizontal field of view of 52 degrees. The helmet also contained monocular eye tracking capability using an Applied Science Laboratory (ASL) 501 video-based eyetracker. The eye position was calibrated before each trial using a 9-point calibration target. Given the average trial duration across subjects, this calibration was carried out every 108 seconds. This frequent calibration was crucial in maintaining accuracy significantly below 1 degree of visual angle. In addition the rotational and translational degrees of freedom of head movements were monitored with a HiBall-3000 tracker. The head tracker had a latency of a few milliseconds so that the frame update in the HMD is between 30 and 50 msec. The scene was rendered using a Silicon Graphics Onyx 2 computer at a rate of 60 Hz. Three data streams were recorded simultaneously. First, the current position and orientation of the head mounted display, the gaze position relative to the display, and the current time code at which the sample was taken were written to a file. Sampling frequency was at 60 Hz. The second data stream contained Hi-8 video recordings of the scene as seen by the subject with superimposed crosshair representing the point of gaze and
Figure 3.1: Subjects view of the walkway from the starting position in three different conditions. Left: pickup purple litter and avoid blue obstacles condition with normal spatial distribution of objects on the walkway. Middle: same condition as in the left view but with a tighter distribution of objects. Right: normal spatial distribution of objects on the walkway with additional salient objects in the scene.

an image of the monitored eye. The third data stream consisted of a recording of the scene as seen by the subject onto digital video. The videos and the data stream were synchronized at the beginning and end of each trial by inserting visual markers into the video streams and tag data into the data file. Given that all systems used the same clock, single frames of the video could be directly related to the corresponding data obtained from the eye-tracker.

One problem in this environment was that the linear track of the path in the cityscape was many times longer than the 7 meters width of the laboratory. Our solution to this discrepancy 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 had to walk a circular path in the laboratory. The path that a subject walked along a single trial took about four laps in the laboratory space. A similar approach was used by Razzaque (Razzaque et al., 2002). Subjects were given enough practice in this environment until they reported being comfortable with the mapping before starting the experimental trials. With practice, subjects perceived the walkway track as being linear.
3.2.2 Experimental conditions

The environment in which subjects were immersed consisted of a linear walkway of length 40m and width 1.8m within the cityscape. At the end of this walkway subjects arrived at a road crossing where the trial ended. A total of 40 purple and 40 blue rectangular objects were placed along the walkway. These objects were placed randomly according to a uniform distribution that expanded 1.5m to both sides of the walkway on normal trials. On half of the trials the purple objects had a height of 1.5m and the blue objects had a height of 2m, while on the other half of the trials the heights were exchanged. Moreover, on half of the trials the purple objects were described to the subjects as being 'litter' and the blue objects were described as 'obstacles'. On the other half of the trials the blue objects were termed 'litter'. The random positions of the objects was different across the task conditions but the same across subjects, i.e. all objects were always at the same position in the 'pick up' condition across all subjects independent of the color. A typical trial consisted of the subject being immersed into the environment and standing still at the beginning of the walkway. Subjects then listened to the instructions describing the current task they were asked to carry out for approximately 15 seconds. After they had listened to the instructions, subjects proceeded to walk along the walkway and executed the task. At the end of the walk-way subjects arrived at the intersection and waited for approximately 15 additional seconds at which point a single trial ended. The overall duration of a single trial was one minute and 48 seconds on average. The time during which subjects were walking along the walkway and executing the instructed task was 80 seconds on average with a standard deviation of 25.0 seconds.

The task priorities were changed across trials by giving the subjects different verbal instructions. In condition 'pickup litter', subjects were instructed to pick up the litter objects, which were purple in one half of the conditions and blue in the other half. Picking up was achieved by approaching the litter object, which disappeared when the subjects body reached a distance of 30cm. The second condition consisted of the 'avoid obstacles' condition, in which the task was to avoid the obstacle objects. In the 'combination' condition, subjects both picked up the
litter objects and avoided the obstacle objects. The order in which individual subjects carried out these three tasks was randomized across subjects, but the spatial arrangements were the same across subjects. On additional salient trials, a large number of additional objects were introduced into the scene. These objects were multicolored, in part moving, or changing shape. Their common property was that they score high on common measures of visual saliency such as high luminosity contrasts, high color contrasts, or high edge density with respect to the entire scene. In the salient condition, subjects carried out the same task as in the 'combination' condition. These salient trials were presented to the subjects after they had carried out one

Figure 3.2: Center: Typical view of the subject during execution of the task showing the walkway with blue obstacles and purple pickup objects, the cityscape in the background. Left: Schematic representation of the scene context represented by the color histogram of the scene, which shows the proportion of colors summed over the entire field of view. Right: Representation of the image patch at two fixation locations (see text). Note how gaze directed to the central region of a solid object is represented as a solid line whereas gaze directed towards an edge is represented as a succession of stripes.
to three training trials in order to familiarize themselves with the environment and three more trials corresponding to the ‘pickup’, ‘avoid’, and ‘combination’ conditions. Accordingly, subjects were not expecting the new, additional, and surprising objects in the scene. Additionally, four subjects were presented with a narrow condition, in which the same four tasks including purple litter objects and blue obstacles as described above were executed but the placement of the objects was altered in such a way, that these objects were confined to the width of the walkway.

All subjects were undergraduates at the University of Rochester who were compensated for their participation. Subjects were nave with respect to the purpose of the experiment.

3.2.3 Analysis of experimental data

The data provided by the eye tracker was analyzed in order to segment the eye movements at saccades. These were determined using in-house Fixation Finder software, which implements an adaptive velocity-based algorithm. The algorithm obtains an estimate of the noise level present in the signal during the entire trial and then compares this estimate with a local estimate of the noise present in a window of one second around the current data sample. The algorithm then changes the current threshold depending on the current noise estimate. All recorded fixations needed to meet the additional criteria of having angular velocity less than 65 deg/sec for at least 60ms and occurring less than 30ms apart and being displaced by more than one degree of visual angle. Data collected during a track loss was excluded. The automated segmentation had previously been compared to manual frame-by-frame analysis obtained from three different experts on data recorded from other experiments and could not be distinguished from their classification. It should be noted that under the conditions described, human subjects executed sequences of complex head and eye movements including frequent vestibular ocular reflexes superimposed on fixations and saccades.

Data from a total of eight subjects was excluded for two reasons. First, three subjects needed considerably longer to finish at least one of the total of four trials. Data from these subjects,
Figure 3.3: Visualization as introduced in figure 2 of the image patch at the center of gaze for subject ML in two different conditions. The x-axis corresponds to the normalized trial duration and the y-axis is foveal eccentricity in degrees. Top: pickup purple litter; Bottom: Avoid blue obstacles. The duration of the entire trial consists of the time the subject spent immersed in the environment and listening to the instructions, the time during the execution of the task, and the time during which the subject has finished the task and is waiting to start the next trial. The figure clearly shows the predominance of fixations on purple objects for the pickup task and a similar pre-dominance of fixations on blue objects in the obstacle avoidance task.

who reported to feel 'uncomfortable in the virtual environment, was excluded. Second, because of the low quality of five of the obtained eye-tracks, these data could not be used for further analysis. All in all nineteen subjects were able to navigate the VR environment comfortably and provided good eye-tracking data. Given that the position and gaze direction are known in the scene, the intersection of the gaze vector can be determined for each moment in time in the VR environment. But this procedure is not very robust. If for example subjects consistently tended to fixate at the walkway close to the edge of an obstacle, utilizing only the object class at fixation location would not reveal this contingency. Therefore, the video sequence showing the scene from the point of view of the subject was utilized in order to extract the object class and the visual features at the point of gaze. In-house Matlab (Mathworks) functions were used to extract the central patch at fixation for each frame of the video sequence. This patch was the circular region of 3 degrees diameter around the point of gaze. This choice was motivated by two factors. First, although the density of cones in the retina falls off continuously, the central
rod free region of highest acuity is approximately 2 degrees in diameter. Secondly, data was only used if the inaccuracies due to the eye tracker were approximately less than one degree. Therefore a region of total size of three degrees was used.

The color at fixation was used in order to determine the object class at the point of gaze for each frame during the trial. Each fixation was then classified as being on the object class that was present at the point of gaze most of the time during each fixation. This took advantage of the fact that the virtual environment was designed in such a way that the colors of the objects clearly identified the object category. The classes utilized were 'litter', 'obstacle', 'walkway', 'lawn', 'other', and 'none'. The class 'none' included saccades, track losses from the eye tracker and fixations at points in the scene that were too close to the boundaries of the scene, so that a patch of one degree diameter could not be extracted. The classes 'litter' and 'obstacle' were assigned depending on the mapping of the colors purple and blue in the respective trial. The classes 'walkway' and 'grass' were the same for all conditions. Finally, the object class 'other' comprised all buildings and roads as well as the trees in the surroundings of the scene. Additionally, in the trials in which a large number of salient objects were introduced in the scene, the additional object category salient was used.

Furthermore, the properties of visual features at the point of gaze were investigated. The analysis of the data in terms of the local image properties was based on the three-dimensional image cube obtained by extracting the image patch at fixation location for each video frame. Here results are reported for the response of two derivative of Gaussians filters oriented horizontally and vertically. The central image patch was normalized and convolved with the two filters. Response distributions were obtained for the two filters for gaze directed to litter in the pickup condition and obstacles in the avoid condition. Although the image cube contains the relevant data for the analysis, this type of data structure is difficult to represent. In order to obtain an informative two-dimensional rendition of the scanpath, a visualization of a trial was developed, as shown in figure 2. The central fixation patch of 3 degrees diameter was unrolled in an outwards spiraling way for each time sample. This was achieved by transform-ing the
patch to polar coordinates and selecting all pixels over all angles for successively increasing discrete-valued radii. The obtained linear patch represents the fixated region with increasing eccentricity from the central fixation point mapped to a linear distance. The center of fixation is represented at the bottom of this linear strip and distance from the fovea increases upwards. As figure 2 demonstrates, this visualization shows gaze directed towards an edge of an object as a sequence of stripes whereas the central part of an object shows up as a solid line. Concatenating these vertical stripes over the duration of a trial results in a visualization of the gaze targets in two dimensions as shown in figure 3. Additionally, a global representation of the scene was obtained from the color histogram of the field of view for each frame. This color histogram of the scene in the field of view of a subject is shown on the left of figure 2. The color histogram is a general representation of the distribution of colors present in an image. It is obtained by counting the number of pixels of a particular color. Given that the color range in the virtual environment was controlled, the individual colors could be segmented without overlap and correspond to individual object classes. Finally, in order to compare the human gaze distributions to a random gaze allocation and saliency models of gaze allocation, fixation sequences were constructed for each subject and each task by selecting a particular location in the scene on the last frame before the subject had actually executed a saccade in the trial. The scene visible in the subjects field of view at each moment in time was recorded in the video stream so that the next fixation location according to different gaze distribution models could be determined. The gaze allocation was modeled 1) with a random gaze distribution in which a location weighted by the subjects distribution of gaze relative to the field of view and 2) according to the saliency model implemented by Itti and described in (Itti, 2000).

Given that the standard saliency model contains approximately 40 free parameters, these values had to be chosen. The parameters were adjusted so as to match the values mentioned in (Itti, 2000) and (Itti & Koch, 2000). Additionally, those parameters for which values were not given in the references or which had different values in these references, were chosen to reproduce the published saliency maps in (Itti, 2000) most faithfully.
Figure 3.4: Comparison of the gaze over time for ten subjects labeled S1 to S10 using the visualization introduced in figure 2. Note that the sequences shown in figure 3 represent the entire trial duration while the sequences shown here are taken only from the time during the trial in which the subjects were moving. Left: Gaze visualization during the execution of the task Pickup purple objects. Right: Gaze visualization during the execution of the task Avoid blue obstacles.
Although most of the discussion is based on the proportion of time spent on the different object classes, inferential statistics were not performed on the necessarily correlated percentages but on the original total fixation times. A two-factor within-subjects ANOVA with repeated measures was performed with the Task and Object Class factors. The durations of individual fixations were compared using a one-way ANOVA with repeated measures. The Greenhouse-Geisser correction was used in all cases in order to take possible violations of sphericity in the repeated measures data into account. The p-values for these tests, which all used a significance level of $\alpha = 0.05$, are reported. Because of the intricacies relating to multiple comparisons with repeated measures, paired t-tests with Bonferroni correction were used post-hoc in comparing individual looking times on object classes across tasks. This type of correction of the applicable p-values is quite conservative, because it takes all possible pairwise post-hoc tests into account.

3.3 Results

3.3.1 Effect of the task on fixation proportions

The proportion of fixations of subjects on the different object classes was highly dependent on the particular task they were executing. Figure 3.4 shows the representation of the gaze sequences described above for ten subjects while executing the tasks 'pickup purple litter' and 'avoid blue obstacles'. These representations are useful, because they provide a simple visualization of the similarities of the gaze sequences across subjects, the difference in gaze allocation across tasks, as well as revealing details about the features at the point of gaze. This representation already carries relevant information accessible to visual inspection. First, the proportion of colors in the two conditions shown is different and reflects the conditions of picking up purple litter versus avoiding blue obstacles. Secondly, subjects tend to fixate regions of uniform color distribution more than regions containing a high density of edges. The center of objects results in a solid stripe at each moment in time while regions with edges are
Figure 3.5: Proportion of fixation time spent on the object classes across subjects (n=19) for four different tasks. The shown proportions are colored according to the color convention depicted in figure 1, i.e. (1) purple represents pickup objects, (2) blue represents obstacles, (3) gray represents the walkway, (4) green represents the lawn and tree, (5) represents the background buildings and (6) light green represents the salient distractors. The diagrams were obtained by averaging over the two color conditions. From left to right: pickup purple litter, avoid blue obstacles, pickup purple litter and avoid blue obstacles, pickup purple litter and avoid blue obstacles with salient distractors in the scene. Error bars are 1 SEM across subjects. These observations can be quantified by obtaining the total times spent fixating the different object classes for the different conditions. This is shown in figure 5, averaged across subjects. The figure shows that the distribution of gaze across the different object classes varies in the four different conditions. This was tested with a two way ANOVA with repeated measures and Greenhouse-Geisser correction. The interaction between task and object class was significant ($p<0.001$). Subjects preferentially fixated the litter more in the pickup condition than in the avoid condition whereas fixations on the obstacles were larger in the avoid condition than in the pickup condition. Note also that the proportion of time that subjects fixate the walkway increased in the avoid condition. Indeed, the differences were significant for litter ($p<0.0002$), obstacles ($p=0.018$), and walkway, ($p<0.001$) when comparing the conditions pickup and avoid,
as revealed by post-hoc paired t-tests with Bonferroni correction.

The distributions shown in figure 3.5 are obtained from both color conditions and are therefore independent of the color identity of the respective object class. Note that the variability in the time spent on the different object classes across subjects is remarkably small. As an example, in the pickup condition, subjects spent an average of 18% of the time on blue objects with a SEM of only 0.05%. This demonstrates that the gaze deployment behavior was very similar across subjects. The distribution of gaze time on the different object classes in the combination condition is intermediate between the distributions observed in the component tasks. Although the differences in fixation time in the pickup and avoid conditions were statistically significant, the difference between each of these conditions and the combination condition were not statistically significant for litter, obstacles, and the walkway. This may be due to the fact that the variability in the looking times was still large due to the different trial durations across subjects and the fact that the Bonferroni correction results in conservative p-values due to the fifteen different pairings of object classes. Inspection of the individual subjects data nevertheless revealed consistent performance. In 18 of 19 subjects litter fixation times were intermediate in value for the combined condition and for 17 of 19 subjects obstacle fixation times were intermediate in the combined condition.

These distributions also demonstrate that the majority of fixations were directed towards objects relevant for the pickup and avoidance tasks and that the grass and the buildings in the background were fixated together across all conditions 15% of the time on average. Differences for gaze directed to the background objects were not significant \( (p=0.549) \). Thus, during execution of the task, subjects pre-dominantly fixated objects that were relevant for the ongoing task and changes in the task priorities resulted in changes in the proportion of gaze being directed to the different object classes. The effect of the salient distractors in the scene on the proportion of gaze during the execution of the task can be regarded as minimal. A comparison between the mean gaze proportions spent on the different object classes for the combination and salient conditions shows that the two distributions are almost identical. In order to compare...
the two conditions the fixation durations on salient objects were excluded, since no salient objects were present in the scene in the combination condition. The distribution of fixations in the two conditions was not significantly different (p=0.35, two-way ANOVA with repeated measures). Only 0.2% of the time was spent fixating the salient objects in the scene in the salient condition. Thus, the objects that have a high saliency value did not attract gaze during the execution of the walking, approaching, and avoiding tasks.

3.3.2 Effect of the scene context

As described above, subjects still fixated litter objects 26% of the time when instructed to avoid obstacle objects. What is responsible for these fixations? Figure 9 shows the fixation sequences of one subject in the two conditions ‘pickup litter’ and ‘avoid obstacles’ together with the color histogram of the scene visible in the field of view. Visual inspection of these plots suggests, that the proportion of fixations on the object classes depended on the task priority, but were

Figure 3.6: Same representation of the fixation distribution of subject SG over time as in figure 3.3. Additionally, the color histogram of the scene in the field of view of the subject, as described in the methods section and shown in figure 2, has been plotted for each of the two trials. Top: The fixations and color histogram for pickup purple litter objects. Bottom: The fixations and color histogram for avoid blue obstacles.
Figure 3.7: Probability of fixating the pickup objects and the obstacles given the proportion of the respective object class in the current field of view. Left: proportion of fixation on a pickup-object (purple) and an obstacle (blue) given the proportion of pickup-objects in the current scene (00.25, 0.25-0.5, 0.5-0.75, 0.75-1.0). Right: proportion of fixation on a pickup-object (purple) and an obstacle (blue) given the proportion of obstacles in the current scene (0 1.0).

also influenced by the proportion of the scene covered by objects of a particular class. In the ‘pickup purple object’ condition, subjects tended to fixate blue obstacles only if the proportion of purple objects within the field of view was small compared to the other object classes. Again, these observations were quantified by extracting the proportion of fixations on the respective object classes, given the distribution of litter objects in the current field of view of subjects. The proportion of the visual field covered by litter objects was extracted from the video sequence for each frame and then related to the current object class at the point of gaze. Figure 8 shows the proportion of time spent fixating litter and obstacles, depending on the proportion either litter obstacles in the filed of view, averaged across subjects. It therefore quantifies to what degree the scene context influences the object class selected by gaze. These graphs reveal that the more the scene is covered with objects that are relevant for the task, the higher the probability of a fixation on the task relevant object. Correspondingly, the fixation time spent on the task irrelevant objects decreases. Note that the small standard errors demonstrate that this effect is robust across subjects and color mappings. In summary, these histograms demonstrate, that
the context of the current scene is highly predictive of the type of object class that is fixated at each moment in time, given the particular task the subject is involved in.

3.3.3 Effect of the spatial distribution of objects

A model of this setting where all the individual tasks competed for the gaze vector produced many more fixations on the walkway than actually observed in our subjects (Sprague and Ballard 2003). We hypothesized that our subjects were taking advantage of the fact that litter was always on the walkway to stay on the walkway by heading to litter. We tested this hypothesis with an additional condition where four subjects were asked to pick up litter objects and avoid obstacles while staying on the walkway, but the distribution of the objects relative to the walkway was altered. Figure 1 shows the different spatial distributions of the objects on the walkway. In the normal width condition, the objects extended over the walkway into the region of the lawn. This was the common distribution that most subjects executed. The 'tight' condition restricted the position of the objects to the walkway. Figure 10 demonstrates that this affected the proportion of fixations directed to the walkway. While subjects fixated the walkway 18% of the time in the normal width condition, they only fixated it 4% of the time in the narrow condition (significance of condition: $p=0.004$ two-way ANOVA). This suggests, that subjects were able to reduce the number of fixations directed towards the walkway, because the information that is acquired during looking at the walkway in the other conditions can be obtained from the fixations directed towards the objects that are positioned on the walkway.

3.3.4 Effect of salient distractors

In the 'salient condition in which subjects executed the same task as in the 'combination' condition, the allocation of gaze to the different object classes was separated for the duration of the task itself and the time in which the subjects were immersed in the environment, but were not yet or no longer executing the task. Figure 3.9 shows a typical example of a sequence of fixations in this condition. At the
Figure 3.8: Comparison of proportion of fixations on object classes between the normal condition and the narrow condition in which the objects were placed only on the walkway. Top row: pickup and avoid in the normal condition. Bottom row: pickup and avoid in the narrow condition.

Beginning of the trial, the subject is immersed in the city-scape and is listening to the verbal instructions. During this time the subject is exploring the visual scene and a large number of the executed fixations fall on the salient objects and the background containing the buildings of the city-scape. While the subject is then executing the task of staying on the walkway, picking up litter objects, and avoiding obstacles, almost no fixations fall on those objects that score high on the saliency measure. Subjects then reach the road crossing at which point the task is finished and they are waiting for the trial to end. During this time a large proportion of fixations is again directed towards the salient objects in the background. Figure 11 quantifies these differences by obtaining the proportion of gaze time spent on the different object classes. The differences between looking times were significant for interactions between object classes and task execution ($p<0.001$ two factor ANOVA with repeated measurements). Thus, subjects fixated the salient objects when they were not executing a specific task. During the time in
Figure 3.9: Visualization as introduced in figure 2 of the unrolled gaze for subject NM in the condition pickup litter objects and avoid obstacles when additionally a large number of salient objects are present in the scene.

which they were listening to the current task instructions and when they were waiting for the trial to end, subjects divided their gaze much more evenly across the different object classes, including the salient objects.

Figure 3.10: Proportion of fixations on object classes for the condition in which a large number of additional salient objects were immersed into the scene. Left: Proportion of fixations during the execution of the task. Right: Proportion of fixations before and after executing the task.

3.3.5 Effect of the task on the fixation durations

Several studies in the past have demonstrated that differential fixation durations can be related to cognitive processes in executing tasks (Hayhoe, 2004; Pelz et al., 2000)(Hayhoe 2004, Pelz et al. 2000). Figure 3.11 shows the average fixation duration for the subjects separated
by task. First, a two-way ANOVA with repeated measures confirmed significant differences between fixation times (interaction between tasks and objects class: $p=0.0003$). The mean fixation durations were 0.35 seconds on average and comparable across tasks for the object classes of pick-up objects, obstacles, walkway, and grass. By contrast, the fixation duration for gaze directed to the background or the salient objects was significantly shorter with a mean duration of 0.17 seconds in all conditions ($p<0.01$ for all paired t-test with Bonferroni correction). Furthermore, fixation durations on litter were significantly longer than those for gaze directed towards obstacles (all paired t-tests with Bonferroni correction $p<0.01$). These results suggest that subjects executed different computations while looking at the background or the salient objects compared to when they looked at objects they were navigating towards or objects they were avoiding.

![Figure 3.11: Average fixation durations on object classes for the four conditions. From left to right: Pickup, Avoid, Combination Salient.](image-url)
3.3.6 Comparison to random gaze allocation and feature-saliency based models

As described above, fixation targets were calculated according to the saliency model described in (Itti & Koch 2000) for each subject whenever a saccade was executed. Figure 14 shows the three conspicuity maps and the resulting saliency map together with the location of most likely fixation after applying spatial competition to the saliency map for one particular scene. The obtained sequence of fixations according to the model was used to obtain the proportion of gaze on each of the object classes and compared to

Figure 3.12: From left to right: An original scene from the sequence recorded from the interaction of the subject with the environment for which the saliency map was calculated. Conspicuity maps for color, intensity and orientation and the final saliency map after applying the winner-take-all competitive network to the sum of the conspicuity maps.

Figure 3.13: Comparison of gaze visualizations obtained from subject SG in condition pickup (top) with the visualizations obtained using a random gaze allocation (middle), and the described saliency model (bottom).
the actually observe proportions resulting from the subjects gaze sequences. As a comparison, figure 15 shows the gaze visualizations for the original gaze sequence for one subject in the conditions 'pickup, when litter was purple and obstacles were blue. These are compared to the traces obtained by randomly choosing a fixation location within the field of view at the current location of the subject and the most salient location within the field of view at the current location of the subject.

Visual inspection of figure 3.4 points towards the similarity of the sequences across subjects and figure 3.13 hints at the differences between the human gaze targets and the two models. Neither the random selection of fixation points nor the saliency model describes the sequential order of fixations well. The dissimilarity with the saliency model is striking. One possible comparison between the predictions by the models and the observed gaze targets is again to obtain the proportions of fixations on the respective object classes. These proportions on the object classes obtained by averaging across subjects is shown in figure 16 for the random gaze model and for the saliency model. First, although inspection of the histograms describing the gaze allocation for the random model suggests a bias towards litter in the pickup condition and a bias towards obstacles in avoid condition, these differences were neither significant with respect to the task (p=0.749) nor to an interaction between task and object classes (p=0.503) as assessed by a two-way ANOVA with repeated measures. On the other hand, the proportion of fixations directed to the background objects is disproportionately higher in the random model. While subjects looked at the background objects 2% of the time, the random gaze allocation directed fixations to the back-ground 22.5

The differences in the allocation of gaze according to the saliency model are even more pronounced. First, the proportions of fixation times did not differ significantly for the conditions without salient objects with respect to task conditions (p=8.49 two-way ANOVA with repeated measurements) and no significant interaction with object classes was found (p=0.77 two-way ANOVA with repeated measurements). Given that the relevant objects for the task are uniformly colored, not textured, and not in contrast to the average luminosity of the scene, these objects
are especially non-salient. Furthermore, during the task execution subjects are close to the task relevant objects, which can cover a large part of the visual field. Such areas of homogeneous color are not deemed salient by the saliency model. Instead, regions in the scene are labeled salient if they have a high contrast in the edge channels and are in contrast to the colors that are dominating the scene. The background contains highly textured buildings and plants with strong contrasts due to the shadows in the scene. It is therefore not surprising that more than 70% of the fixations according to the saliency model are directed towards the background of the scene. Furthermore, in the salient condition 39% of gaze was directed towards the salient objects in the scene, while subjects only spent 0.2% of gaze on these objects.

3.4 Discussion

How does the brain select where to direct the gaze during active purposeful behavior? Most natural behaviors involve multiple tasks. We devised an environment in which the allocation of gaze in component tasks could be systematically investigated. In the context of walking we examined the allocation of gaze in path following, obstacle avoidance, and target approach. The overwhelming determinant of the subjects visual behavior was the ongoing component task. Several different measures of the behavior all demonstrated a consistent influence of the task across subjects.

3.4.1 Task-weighted fixation proportions

The proportion of time spent on the target, obstacle and walkway changes dependent on the task instructions. Previous work has demonstrated that fixations are directed to task relevant objects and locations within a scene (Ballard et al. 1995, Land 2004, Johansson 2000, Hayhoe et al. 2003) when these are needed. Shinoda et al. (2001) also observed that the distribution of fixations depends on the goal. It has also been observed, that at the beginning of a particular trial, when subjects have not yet inspected the scene, a number of fixations are used to explore
the entire environment (Land et al. 1999, Hayhoe et al. 2003), presumably to build up a spatial representation of the surroundings. Similarly, when subjects were immersed in the scene in the current experiments, they first executed a sequence of fixations that were distributed almost equally between the different object classes. But when walking along the walkway and interacting with the pick-up objects and obstacles, fixations were almost entirely directed towards the relevant objects necessary for the ongoing tasks.

It is of course not possible to control the allocation of attention to the different component tasks. For example, subjects usually avoided obstacles, as expected for over-learned behavior, and obstacles were always being fixated, even when the instructions were to pickup litter objects. Moreover, in all conditions subjects have to navigate down the walkway. It has been shown, that gaze in human loco-motion can utilize optic flow (Warren et al. 2001), land-marks, or scale changes (Shrater et al. 2000) for navigation, and that the degree to which these are utilized depends on the task and affects the distribution of gaze accordingly (Turano et al. 2005). Thus, a certain proportion of fixations in all conditions can be interpreted as being used to aid in navigating along the walkway and is therefore not only determined by the tasks of picking up or avoiding. An added complication in mapping fixations onto tasks is that the different tasks require gaze in different ways and for different durations. Consequently, a large part of the gaze time was spent on pickup objects, which may reflect that gaze was maintained on the object until contact with the body was made whereas gaze was departed from obstacles once avoidance could be assured.

3.4.2 Task sensitive feature statistics

The different interactions with the objects in the scene that are required by the tasks also determine the statistics of features at the fixation location. The density of horizontal edges was different for fixations directed towards objects that were approached versus objects that were avoided. It has been shown that image feature properties at fixation location are different from those obtained from randomly chosen image locations (Mannan 1996, Rinagel & Zador...
1999, Parkhurst & Niebur 2003, Tatler & Baddeley 2005), but it is likely that this results from object properties rather than from low-level image properties which are by them-selves attracting visual attention (Einhhuser & Knig 2003). The current experiments clearly show, that the majority of fixations is directed towards regions of the scene which are uniform in contrast, have low edge density, and contain no color contrast. More importantly, the density of horizontal edges is different depending on how subjects interacted with the object: if an object was approached, the edge filter response of the fixated image patch was significantly lower that when an obstacle was avoided. This al-most certainly reflected the information being extracted for the purpose of controlling the body. For example Johansson et al. (2001) showed that subjects fixate the edge of an obstacle to be avoided by a hand movement. Thus one cannot make conclusions about the significance of image properties at fixation in the absence of a known task con-text.

The primary influences of the task on the selection of fixation targets persisted also during the trials in which a large number of salient objects were introduced into the scene. While regions of the image that score high on saliency measures have been shown to be more likely to be selected during the first few fixations in picture viewing (Parkhurst & Niebur 2003), the current experiments show that this result cannot be generalized to extended navigation tasks such as the one considered here. Interestingly, subjects did fixate these objects of high color and luminosity contrast but only during the initial and the final part of the trial, i.e. during time in which they were no asked to execute a specific task. This result suggests, that previous studies in which human gaze was directed towards salient stimuli (Parkhurst et. al 2002) or to new objects (Hillstrom & Yantis, 1994) or in which attention was captured by transients (Remington, Johnston & Yantis, 1992) under passive viewing conditions while subjects viewed stimuli presented on a screen may be reconsidered under different task conditions such as extended visuomotor tasks.
3.4.3 Context dependence of gaze

A novel finding was the influence of context on gaze location as demonstrated in figure 8. Independently of the task instructions, subjects fixated the pickup and obstacle objects in proportion to their areal extent in the field of view. How can this be explained? In the present study a color histogram was used as a global descriptor of the scene. The color histogram is a statistical measure that quantifies the distribution of colors within the field of view of the subject at each moment in time. First, color has been demonstrated to be a powerful cue in natural scene recognition (Oliva & Schyns 2000, Wichmann et al. 2002) in the tasks of picture viewing. Perhaps the simplest explanation is that if a pickup object covers almost the entire field of view in the head mounted display, only limited alternatives as gaze targets exist. Secondly, because the colors in the virtual environment reflect object categories, the proportion of a color in the color histogram directly reflects the distance to the closest object, as well as the number of objects in the current field of view. Thus, if there are no litter objects close by, subjects will not be able to gaze at litter but will instead look at obstacles, even if the instructions are to pick up litter. In addition, more complex contextual effects could be influencing gaze e.g. (Chung & Jiang 1998). Torralba et al. (2006) recently demonstrated, that in a search task, contextual features can be extracted from a large number of labeled images in order to characterize the likelihood of objects being present at specific points in an image.

3.4.4 Gaze sharing

It was also observed that subjects take advantage of the layout of the scene in directing their path resulting in a decrease in the number of fixations directed towards the walkway when the target objects were confined to the walkway. It is possible that subjects might use peripheral vision to guide position on the walkway. However, these results argue instead that the fixations on objects do double duty and serve to aid navigation as well. If subjects used peripheral vision for path control, it would not be expected that the walkway fixations would depend on the
placement of the obstacles and the litter objects. Instead, subjects reduced the proportion of walkway fixations from 12% to 4% when objects were confined to the walkway. This demonstrates that the fixational time spent on the walkway in navigation depends on the layout of the scene and that subjects can take advantage of the fact that the walkway position can be inferred from the obstacle and target positions.

3.4.5 Comparison with gaze allocation models

The proportion of fixations observed was compared to a specific saliency model as proposed by Itti and Koch (2001). This saliency model did not predict the proportion of fixation on the different object classes. Moreover, in the experiments, it was observed that the target features were dependent on the ongoing task, i.e. the type of action the subject was executing. When navigating towards the object subjects fixated the center of the pickup objects resulting in a regions with low spatial frequency content being fixated whereas when navigating around an obstacle subjects tended to fixate the edge of the obstacle around which they were navigating. Current saliency models do not have sufficient complexity to tackle such subtle task-dependent differences in fixation strategy. An advantage of the present paradigm is that it can be directly compared to gaze allocation models, such as that introduced by Sprague and Ballard (2003), which describes a task-based model of gaze allocation in the identical environment. Quantifying human gaze allocation in multiple tasks can therefore be used to further develop such models.
<table>
<thead>
<tr>
<th>Subject SG</th>
<th>Random gaze model</th>
<th>Saliency model</th>
</tr>
</thead>
<tbody>
<tr>
<td>'pickup'</td>
<td>'avoid'</td>
<td>'combination'</td>
</tr>
<tr>
<td>'salient'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.14: Comparison of proportions of fixation times spent on the object classes for human subjects (top), a random gaze allocation model weighted by the spatial gaze distribution (middle) and a saliency model (bottom). The four conditions are shown by column: pickup, avoid, combination, and salient conditions.
4 Task determines feature statistics at point of gaze

4.1 Introduction

It has been proposed that a fundamental principle for the understanding of neuronal representations of sensory data is that they have been shaped on evolutionary and developmental timescales by the regularities in the environment (Attneave, 1954; Barlow, 1961). Theories of the encoding of natural stimuli based on their statistical regularities have been successful at reproducing a number of properties of neurons in early cortical sensory areas in humans and animals (for an overview see (Simoncelli & Olshausen, 2001; Dayan & Abbott, 2001)), but the dependence of these statistics on the task have been considered to a much lesser degree.

This paper analyzes statistics of the image at the point of gaze for human subjects involved in natural behavior. Although it may seem immediately plausible that the statistics of the sensory input to the visual system are dependent on the ongoing task, this possibility has far reaching consequences. The hypothesis that neurons have evolved to maximize mutual information between sensory stimuli and their neuronal representations (Bell & Sejnowski, 1997), or to build a sparse code (Olshausen & Field, 1997), or are encoding a difference signal between higher level predictions and actual input (Rao & Ballard, 1999) require the characterization of the input statistics. Learning of such representations has previously used static images or image sequences that have been sampled artificially. At the very least, if the efficient coding of
natural stimuli has shaped the responses of neurons, the changes in these statistics must have influenced the adaptation processes that adjust the encoding (see e.g. (Clifford et al., 2007)).

A different line of research has investigated properties of image regions that are fixated by human subjects in different picture viewing and search tasks (Mannan et al., 1996; Reinagel & Zador, 1999; Parkhurst & Niebur, 2003). This research has extensively used the notion of ‘saliency’ (Koch & Ullman, 1985b; Itti et al., 1998) in order to explain or attempt predicting human gaze selection. It is rather surprising, that the operational definition of saliency that has been promoted for a decade and is currently dominating the literature is that local image properties such as contrast in feature dimensions as luminosity, edge filter responses, and color attract gaze automatically. This is surprising, when one considers the comparison of observing the animal behavior of going to a river in order to drink water. The two possible explanations for the observed behavior could be that the water is attracting the animal or that the animal went to the water because it was thirsty. The current usage of the saliency concept corresponds mostly to the first interpretation and is almost entirely separated from the behavioral goal of gaze within visual behavior (Itti et al., 1998). This article provides further experimental evidence and quantifies that the features at fixation location are dominated by the ongoing task.

In contrast, extensive studies have shown that in picture viewing (Yarbus, 1967; Henderson & Hollingworth, 1999) and natural behavior (Land & Hayhoe, 2001; Johansson et al., 2001) image regions and visual features are targeted by gaze in dependence of the ongoing task and not only local image properties. The research on extended natural tasks has demonstrated that gaze is related to manual interactions with objects, predicted future locations of objects, retrieval of specific object features depending on the sequential order of tasks, and other parameters of the ongoing behavior (Hayhoe & Ballard, 2005). In terms of the analogy given in the previous paragraph, it may be behaviorally relevant to monitor the position of a predator that is moving in an environment for which its natural camouflage is optimized such that gaze is more likely to be directed towards regions that are low in luminance contrast, edge content, and color.
contrast. The question in this study is: How do the image statistics of the input to the visual system depend on where gaze is directed to in the extended visuomotor task of approaching and avoiding objects during sidewalk navigation?

Given that the visual system actively selects targets in the scene when involved in executing visuomotor tasks, it is of interest to characterize these features in dependence of the current behavioral goal. Two hypotheses in the animat vision framework (Ballard, 1991) are 1) that by directing gaze to specific targets in the environment during motion, visual computations based on the exocentric coordinate system can be done with less precision and 2) simpler control strategies based on servoing relative to the fixation frame can be used. Therefore, the gaze targets can reveal features pertaining to the visuomotor task that is being solved. This is different from a search task, where an optimal feature for the detection of a target embedded in a background can be arrived at purely in terms of the image, e.g. (Najemnik & Geisler, 2005; V. Navalpakkam, 2007). Here, a collection of features was analyzed in order to establish quantitatively how the task influences the gaze targets and thereby the statistics of the input to the visual system.

A further motivation for the following analysis of intermediate features at fixation locations in dependence of the ongoing task is the observation (Zhu, 2003), that rich vocabularies for features at different hierarchical levels exist for example in language, including features such as ‘phonemes’, ‘syllables’, ‘words’, and ‘sentences’, but that similar vocabularies are sparse for visual features. In terms of this analogy, the visual features that are pursued here, are features that relate to how vision is used in order to achieve a goal in visuomotor behavior and could be related to the field of pragmatics in linguistics. This paper contributes to the extraction of such visual features for the tasks of approaching and avoiding objects while navigating along a walkway.
4.2 Materials and Methods

4.2.1 Experimental setup

Subjects executed the two task conditions ‘pickup’ and ‘avoid’ while they were immersed in a virtual reality environment consisting of a cityscape (Performer Town) created by SGI. They wore a Virtual Research V8 head mounted binocular display having a resolution of 640 by 480 pixels corresponding to a horizontal field of view of 52°. The helmet also contained monocular eye tracking capability using an Applied Science Laboratory (ASL) 501 video-based eye tracker. The eye position was calibrated before each trial using a 9-point calibration target. This frequent calibration was crucial in maintaining accuracy below 1° of visual angle. In addition, the rotational and the translational degrees of freedom of head movements were monitored with a HiBall-3000 tracker. The head tracker had a latency of a few milliseconds so that the frame update in the HMD is between 30 and 50 ms. The scene was rendered using a Silicon Graphics Onyx 2 computer at rates above 60 Hz. The collected data consisted of the position of gaze, the current frame, and the current position and direction of the subject’s head.

The environment in which subjects were immersed consisted of a linear walkway of length 40 m and width 1.8 m within the cityscape. At the end of this walkway, subjects arrived at a road crossing where the trial ended. A total of 40 purple and 40 blue cylindrical objects were placed along the walkway. These objects were placed randomly according to a uniform distribution that expanded 1.5 m to both sides of the walkway. The purple cylinders were described to the subjects as ‘litter’, while the blue cylinders were termed ‘obstacles’. The random positions of the objects were different across the task conditions but the same across subjects; that is, all objects were always at the same position in each of the 3 conditions across all subjects. The overall duration of a single trial was 93 seconds on average. Both object types were textured with a image texture sampled from $1/f^2$ noise. Figure 4.1 shows nine representative views from the subject’s perspective during execution of a trial. One problem in this environment was that the linear track of the path in the cityscape was four times longer than the 10-m width of
the laboratory. Our solution to this discrepancy was to break up the linear path into five linear tracks of shorter distance. Subjects walked along the walkway until they reached the end of the laboratory. At this point the display turned black and subjects turned around and continued walking on the walkway while moving in the opposite direction within the laboratory. Subjects were initially given enough practice trials until they were familiar enough with this mapping.

Subjects were given different verbal instructions for the two experimental conditions in order to change the task priorities of approaching and avoiding objects. In the first condition ‘pickup’ subjects were instructed to pickup the purple litter objects. Picking up was achieved by approaching the litter object, which disappeared when the subject’s body reached a distance of 20 cm. The second condition consisted of the ‘avoid’ condition, in which subjects had to avoid the obstacles. The order in which individual subjects carried out these two tasks was randomized across subjects, but the spatial arrangements were the same across subjects. All subjects were undergraduates at the University of Texas who were compensated for their participation. Subjects were naive with respect to the purpose of the experiment.

Figure 4.1: Nine representative views of the scene as seen by a subject during execution of the tasks.
4.2.2 Analysis of experimental data

The goal of the present analysis is to quantify differences in the input to the visual system due to the execution of different visuomotor tasks. The two tasks considered here are approaching and avoiding objects while walking along a walkway. The features selected for the following analyses were chosen in order to compare the results with previous research that analyzed static natural images. First, the analysis of the image data required building different image sets. A first image set was build by randomly selecting fifty images along the entire walkway from the point of view of different subjects executing the tasks. The two task conditions relevant for the following analysis are conditions ‘pickup’ and ‘avoid’. Two more image sets were obtained by selecting the image data at the center of gaze for all subjects in the two conditions. An automated saccade extraction algorithm based on and adaptive velocity threshold was applied to this data in order to select only the frames during which gaze was relatively stable on a particular object in the scene. Summary statistics were obtained in order to exclude biases in the analysis due to the usage of artificially rendered images.

The features of luminosity, contrast, multi-scale filter responses, were calculated for each image patch in the image sets. These features were calculated according to common methods used in image processing. In the following paragraphs, the details of these calculations are reported together with brief descriptions of the motivations for such choices.

Luminance

The experimental setup was such that subjects wore the head mounted binocular display in which the virtual scene is shown. The luminance of the display screen was measured using a photometer and the relationship between pixel values in the image and luminance in units of candelas/m² was established. After calibration, the luminance was estimated in the displayed scene, and was measured in circularly symmetric image patches by using windowing functions
of the form:

\[ w_i(x, y) = \frac{1}{2} \left( \cos\left( \frac{\pi}{r} \sqrt{(x - x_i)^2 + (y - y_i)^2} \right) + 1 \right) \tag{4.1} \]

where \( r \) is the radius of the circular region centered around \((x_i, y_i)\) and values outside the circle of radius \( r \) are set to 0. Varying sizes of windowing functions were used with radii of 8, 16, 32, 64 pixels, corresponding to 1.2°, 2.4°, 4.8°, 9.6° of visual angle. This type of windowing functions was mainly chosen because it has been used previously making it easier to compare the results with previous work described in the literature (Frazor & Geisler, 2006). Luminance was then calculated using this windowing function as the weighted sum:

\[ \bar{L}_i = \frac{1}{\sum_x \sum_y w_i(x, y)} \sum_x \sum_y w_i(x, y) L(x, y) \tag{4.2} \]

**Contrast**

Contrast is regarded as one of the fundamental properties of the visual environment so that extensive analysis of its statistics in natural images has been collected. There is a variety of different definitions of contrast that have been used in the literature in order to quantify the variation in luminance in images. Here, the root-mean-squared definition of contrast was used and measured in circularly symmetric image patches by using the above windowing functions according to equation 4.1 in order to compare the results with previous work (Frazor & Geisler, 2006). Local contrast was then calculated using the average luminance \( \bar{L} \) as defined above in equation 4.2 within the window according to:

\[ C_{RMS_i} = \sqrt{\frac{1}{\sum_x \sum_y w_i(x, y)} \sum_x \sum_y w_i(x, y) \frac{(L(x, y) - \bar{L}_i)^2}{\bar{L}_i^2}} \tag{4.3} \]
Power spectrum

The power spectrum of the images was estimated by calculating the discrete Fourier transform (DFT) of an image with spatial dimensions $N$ by $N$ according to:

$$F(\xi, \eta) = \frac{1}{N^2} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} L(x,y) e^{-\frac{2\pi i}{N}(x\xi+y\eta)}$$

(4.4)

where the discrete spatial frequencies are $f_x = \xi/N$ and $f_y = \eta/N$, $i$ is not an index but the imaginary unit, and then obtaining the modulus of the complex valued transform according to:

$$|F(\xi, \eta)|^2 = F(\xi, \eta)\overline{F(\xi, \eta)}.$$

(4.5)

The rotationally invariant powerspectrum was then obtained by calculating the average one dimensional powerspectrum at regularly spaced orientations every $15^\circ$ starting at $0^\circ$.

Multi-scale filter responses

Several functional forms for multiscale filter representations of images have been used to approximate simple and complex cell responses encountered in primary visual cortex. Each set has particular properties that distinguishes it from the other sets, such as biological plausibility, compactness, steerability, separability, and more. One set with particularly nice computational properties is the set of oriented derivatives of Gaussians (Freeman & Adelson, 1991). This set has the advantage of steerability while still resembling the profile of simple cells. Other sets that are less biologically plausible have been used to decompose images, but one of the recurrent findings is that the distributional properties associated with representing natural images is obtained with a wide set of different filters (Simoncelli & Olshausen, 2001). Another often used set are Gabor functions. This set has been shown to emerge as the solution to a number of theoretical optimality formulations (Daugman, 1985) and has been often used in the literature in order to asses the distribution of parameters describing simple cells in animals (Ringach, 2002).
In principle the set of derivative of Gaussians has a good mixture of properties, but the only disadvantage is that it is not straightforward to obtain model complex cell responses using this filter set. In order for a set of filters to be used to obtain monocular energy responses, the two filters used as quadrature pair have to be Hilbert transforms of each other, see e.g. (Mallot, 2000). Unfortunately there is no closed form expression for the Hilbert transform of first and second order derivatives of Gaussian functions, although numerical approximations can be obtained in the frequency domain. Instead of using these filters, it was decided to instead use Gabor filters, because a pair of filters consisting of an even and an odd Gabor function are a Hilbert transform pair of the respective other filter. Accordingly, the set of filter functions for the computation of the model single and complex cell responses were obtained as:

\begin{align}
G(\sigma_x, \sigma_y, \omega) &= \frac{1}{2\pi \sigma_x \sigma_y} e^{-\frac{1}{2} \left( \frac{x^2 + y^2}{\sigma_x^2 \sigma_y^2} \right)} \cos \left( 2\pi \omega (x + y) \right) \\
G(\sigma_x, \sigma_y, \omega) &= \frac{1}{2\pi \sigma_x \sigma_y} e^{-\frac{1}{2} \left( \frac{x^2 + y^2}{\sigma_x^2 \sigma_y^2} \right)} \sin \left( 2\pi \omega (x + y) \right)
\end{align}

where \( \sigma_x \) and \( \sigma_y \) are parameters regulating the width of the spatial windowing function and \( \omega \) determines the spatial frequency of the sinusoid. The size of the windowing function was chosen to be \( \sigma_x = 3/\omega \) and either \( \sigma_y = 3/\omega \) or \( \sigma_y = 5/\omega \) which is in accordance with the literature. The four levels of spatial frequencies of the sinusoids were 0.4, 0.8, 1.6, 3.3 cycles per degree. The above functions were calculated at four different orientations by rotating the coordinate systems about the angle \( \theta \in \{-\pi/4, 0, \pi/4, \pi/2\} \) according to:

\begin{align}
x' &= x \cos \theta + y \sin \theta \\
y' &= -x \sin \theta + y \cos \theta
\end{align}

4.3 Results

It is necessary to compare the image set corresponding to the view of the subjects during task execution in the virtual environment to natural images obtained in natural environments.
Figure 4.2: The set of Gabor functions used in encoding the image. Left: The Unnormalized functions. Right: The same functions normalized individually for display.

This comparison confirms the validity of the results and avoiding biases that may be due to the virtual environment, the textures applied to the objects therein, and the rendering of the scenes. As no full statistical description of natural images is available, summary statistics of subsets of images from within the virtual environment were obtained and compared to known properties of natural image ensembles.

The second order statistics of natural images can be characterized by the power spectrum of the spatial luminance function. It is well known, that the power spectrum for natural images demonstrates scale invariance by following a $1/f^\alpha$ envelope, where $\alpha$ is close to 2, e.g. (Ruderman & Bialek, 1994). Different natural as well as man-made environments have been shown to follow such power laws, with small variations in the exponent around 2, e.g. (Balboa & Grzywacz, 2003). For the virtual environment a set of 50 images was selected randomly from different subjects at different locations along the entire walkway. This set contains typical views of a subject as those shown in figure 4.1. The resulting rotational average of the power
spectrum for these scenes, the individual textures on the objects placed on the walkway, and a set of natural images are shown in figure 4.3. The plot demonstrates that the second order statistics of the views in the virtual environment are close to those reported for natural image sets.

![Log-log plot of the rotational averages of the powerspectra of different image sets.](image)

**Figure 4.3:** Log-log plot of the rotational averages of the power spectra of different image sets. The curves correspond to the slope of the rotational average for a set of natural images (green), the object texture in the VR experiments (purple), and multiple views from the subjects perspective in the VR experiments. Slopes of $-2$ corresponding to $1/f^2$ envelopes are shown as dashed lines.

It is furthermore well known that the natural images contain statistical regularities that go beyond second order. Although there are no models describing these higher moments such as the skewness or kurtosis in a general form, one common observation is that histograms of responses to local filters have highly kurtotic distributions (Simoncelli & Olshausen, 2001). If the natural image statistics were only of second order, they would follow Gaussian distributions and therefore all marginal distributions would also be Gaussian. Therefore, the kurtotic distributions are further evidence for the higher order dependencies. Here, the histograms of the responses of the filters defined in equation 4.7 were obtained and the histograms for the subset corresponding
to the odd filters with spatial frequencies of 1.6 cycles/degree shown in the sixth column of figure 4.2 are shown in figure 4.3. These histograms demonstrate the familiar kurtotic distributions known from natural image sets.

![Figure 4.4: Example average histograms of filter responses for a subset of the Gabor filters. The histograms were calculated for the filter responses to odd Gabors with spatial frequency of 1.6 cycles/degree on the set of 50 images obtained from the view of subjects walking down the walkway in the virtual environment.](image)

4.3.1 Contrast statistics at point of gaze

Previous work on the distribution of contrast in natural images has investigated how contrast statistics depend on the size of the area over which they are calculated for different environments and different object classes such as foliage and backlit areas (Frazor & Geisler, 2006). While these results were obtained on static images, other investigations have measured these statistics at fixation location (Reinagel & Zador, 1999). Here the same statistics as in (Frazor & Geisler, 2006) were obtained, but separately for the two task conditions of approaching and avoiding objects in the walkway tasks. Furthermore, this analysis was repeated separately for gaze falling on litter in the ‘pickup’ condition and gaze falling on obstacle in the ‘avoid’ condition. The respective plots of the contrast for the different image ensembles are shown in figure 4.5.

The overall contrast level for the set of images obtained in VR is lower than that obtained in natural environments (Frazor & Geisler, 2006). This is to be expected, given that the contrast in displays as the one used in this study is smaller than the contrast that can be recorded with
Figure 4.5: RMS-contrast as function of image patch size for different image ensembles and tasks. Left: comparison of contrasts between image sets at the point of gaze for the conditions pickup, avoid, and the combination task. Right: comparison of contrast as a function of patch size for the avoid and pickup condition as in the left plot. Additionally, for the pickup condition the contrast for gaze falling on the litter objects and and contrast for gaze falling on the obstacles was plotted.

cameras such as the one used for the study (Frazor & Geisler, 2006). Moreover, the scenes in VR are rendered with omnidirectional diffuse lighting so that no hard shadows are present in the image set. Nevertheless, local contrast was between 0.08 and 0.13 for patch sizes of 1.2 degrees radius versus a contrast of 0.3 for images of ground areas have been reported earlier (Frazor & Geisler, 2006). The overall pattern of an increase in contrast with patch size observed here is similar to previous results (Frazor & Geisler, 2006). The contrast for the combination of the two conditions pickup and avoid matches the one for the set with randomized positions remarkably closely. By comparison, the corresponding plots for the tasks ‘pickup’ demonstrate a mostly shifted version of this trace towards lower contrasts, while the curve for the ‘avoid’ condition are shifted towards higher contrasts. Note, that these are the results of averaging
over the entire task duration for the respective conditions, i.e. these curves reflect the average contrast resulting from gaze being directed to all object classes.

The common finding in all studies that have investigated the contrast at fixation location versus randomly chosen locations is that contrast is significantly higher at the locations targeted by gaze (Mannan et al., 1996; Reinagel & Zador, 1999; Parkhurst & Niebur, 2003). In contradiction to those studies, the above results show, that contrast can be reduced fixation location when subjects navigate towards objects of low contrast. While previous studies used the free-view or different types of search tasks, in the current study subjects were approaching homogeneously textured objects placed in the environment and the data was collected over a larger time as in the previous studies, where typically only the first dozen or so fixations are considered (Mannan et al., 1996; Reinagel & Zador, 1999; Parkhurst & Niebur, 2003).

In order to further investigate the source of this variation, the RMS-contrast was separately calculated for gaze directed to litter in the ‘pickup’ condition and to obstacles in the ‘avoid’ condition. In order to obtain these average contrasts, a criterion had to be applied to the gaze data in order to classify a fixation as being directed towards either object class. The selection of this criterion has to take into account, that the eye tracker was calibrated for each trial so that the positional error was help below one degree of visual angle. The criterion chosen was to classify gaze to be directed to litter or an obstacle if the area within one degree of the fixation location reported by the eye tracker was entirely on one of the object classes. Note that this is a very conservative criterion which excludes many frames during which subjects directed their gaze directly to the edge of an object. This tended to underestimate the contrast especially for small windowing functions. Nevertheless this criterion was necessary to exclude the confounds resulting from gaze being directed to edges between litter and obstacles.

The contrast for these image sets are shown in figure 4.5. These plots demonstrate, that contrast at fixation location was even lower when directed to litter in the pickup condition and further elevated when directed to obstacles in the avoid condition. This results demonstrates that gaze was directed towards different regions depending on the ongoing task.
4.3.2 Model simple cell response statistics at point of gaze

The higher order statistical dependencies in natural images have been described repeatedly and reported by demonstrating that the response distributions to linear filters similar to those found in primary visual cortices of the mammalian brain have high kurtosis, e.g. (Simoncelli & Olshausen, 2001). While differences in kurtosis may depend on the type of image region and have been used to analyze and synthesize different texture pattern, the question here was, whether there are significant differences in these statistics over extended periods of time due to the execution of tasks.

Figure 4.6 shows the histograms of the model simple cell filter responses collected across subjects separately for the tasks pickup and avoid. The histograms depicted are for two filters only, where the left image is for the vertically aligned Gabor function of highest spatial frequency and the right histogram is for the horizontally aligned Gabor of highest spatial frequency. These plots demonstrate, that the responses for the vertical filter are significantly higher in the pickup condition, while differences are not significantly different between the two tasks for the horizontal filter. The standard error over the mean in the histograms are very small, owing to the large set of image patches these histograms are based on. Although the differences between histograms may seem small, note that the probability of a normalized filter response of 0.3 or higher for the vertical Gabor was twice as high in the pickup condition compared to the avoid condition.

4.3.3 Model complex cell response statistics at point of gaze

While significant differences were already visible for the single cell responses, it is more natural to ask, whether significant differences are also observable for model complex cell responses, where here monocular energy was computed. The main characteristic of these filters is that they are phase invariant for edges, i.e. the model energy cell responds to edges in its entire field and the response will depend less on the polarity of the visual stimulus. Given that the
task required subjects to navigate the environment containing the litter and obstacles it should be expected that these responses are more relevant to the task at hand and that therefore significant differences in these responses should be observable across the task conditions.

Indeed, figure 4.7 shows the responses to a horizontal and a vertical set of model energy neurons separately for the tasks pickup and avoid. The parameters for the component cells were the same as for the previously used model simple cells. The histograms again demonstrate that significant differences in responses between the two orientations can be found across a timescale of a few minutes. These differences reflect the fact that gaze was more likely to fall on edges in the avoid condition compared to the pickup condition.
Figure 4.7: Histograms of model complex cell responses. Left: Normalized responses to vertical model complex cell for condition pickup (purple) and avoid (blue). Right: Normalized responses to horizontal model complex cell for condition pickup (purple) and avoid (blue). Note that the errorbars corresponding to the standard error around the mean are so small that they fall on top of the data points.

4.3.4 Extraction of intermediate features

The above results demonstrated that the statistics of filter responses are different depending on the ongoing task. Here the question is investigated as to what type of intermediate features are used by the visual system during the navigation tasks. The method employed is to extract features of intermediate complexity from the obtained filter responses. Previously, such features have been analyzed in the setting of object recognition and classification (Ullman et al., 2002), whereas here these features are relevant because gaze uses them for servoing.

There are a variety of methods that can be used to extract such intermediate level features. The k-nearest neighbors clustering was used to extract clusters of feature vectors from the set of 50 images obtained along the entire walkway path. Figure 4.8 shows the respective reconstructions in image space. From these clusters that were obtained on the individual images, a set of 25 clusters was extracted. These are shown in figure 4.9. These image patches
Figure 4.8: The set of 30 cluster centers obtained from the image representation using the Gabor filters for the set of 50 images obtained along the walkway.

can be thought of as a small set of image patches that can be used to reconstruct all the images that subjects encounter along the walkway.

The interesting question here is, how subjects distributed their gaze among these different intermediate features during execution of the two tasks pickup and avoid. In order to answer this question the patches at fixation location across all subjects and tasks were obtained, and each feature vector representing the fixation patch was assigned to one of the 25 clusters. The resulting histograms are shown in figure 4.10.

The histograms together with the cluster centers demonstrate, that the vast majority of fixations was directed towards image locations that are well represented by clusters 13 and 14, which represent the solid purple and blue areas of the objects on the walkway. The likelihood of
a fixation being on image regions closes to these to clusters was almost $P=0.85$. This shows the remarkable selectivity of gaze during task execution. Closer examination of these histograms shows, that in the avoid condition, subjects were twice as likely ($P=0.016$ versus $P=0.008$) to fixate an image region that was best characterized best by cluster center 16, which is a blue versus gray edge.

4.3.5 Relating the feature statistics to the gaze targets

The above analyses demonstrate, that the statistics of image features at the point of gaze are significantly different depending on the ongoing visuomotor task. What these analyses do not reveal is how such differences come about. In order to establish the difference in gaze allocation between tasks that are the cause of these differences, the landing position of the first saccade executed towards obstacles and targets can be marked on these objects respectively.
for the different tasks. The distribution of fixation targets relative to the objects is obtained by averaging across subjects. The colored bars in Figure 4.11 show the results for a single subject. The saccade target patterns suggest that this subject was more likely to target the edge of the obstacles while the center of the litter objects were more likely to be targeted in the pickup condition. The marginal distributions were obtained by normalizing the target distribution of all subjects after convolution with a Gaussian kernel of standard deviation of 0.5 degrees, which reflects the uncertainty in the position of the eye tracking measurement.

The distributions of gaze targets across subjects demonstrate, that subjects are indeed more likely targeting the edges of the obstacles when avoiding them versus looking at the centers of the targets when approaching them. The differences in the observed statistics of image features such as contrast, model simple and model complex cell responses can therefore be attributed to the differences in gaze targets used by the subjects in the different visuomotor tasks.
Figure 4.11: Horizontal and vertical marginal distributions of gaze targets on litter objects in the pickup condition (left) and on obstacles in the avoid condition (right). These distributions were obtained using data from all 19 subjects. The plots with targets marked on the objects were obtained from the data of subject MB and are representative of the entire data set.

4.4 Discussion

If the representation of sensory information is shaped by the regularities of the input to the visual system, the characterization of this input should aid in the understanding of the neuronal representations encountered in the brain. These regularities are not only due to the regularities in the environment but depend also the regularities in the selection by the visual system. Therefore, this input needs to be characterized in dependence of the executed eye movements, which in turn depend on the ongoing task. Furthermore, adaptation processes that operate on different time scales and adjust the the output statistics of neuronal activity in dependence of the statistics of the stimulus input have to deal with the input that is in turn dependent on the task effects. The present study analyzed and quantified the influence of this active selection
process in the visual scene by the oculomotor system on contrast, model simple and complex cell responses, and on features of intermediate complexity, which are all known to be important carriers of information about the visual world.

4.4.1 Contrast statistics

Luminance and RMS-contrast obtained in the virtual environment are comparable to those encountered in natural images over several aperture sizes as reported in (Frazor & Geisler, 2006). A bias due to the differences in the images obtained by rendering the scene were excluded by obtaining a powerspectrum and edge-filter response distributions comparable to natural images. Also, contrast increased with larger aperture sizes as previously described for natural image ensembles that were sampled at simulated scan paths.

The difference with previous studies was, that the subjects executed visuomotor behavior in natural tasks with different task priorities. This allowed to analyzed the feature statistics as a function of the task. It was shown, that the contrast statistics of RMS-contrast, model simple cell responses, model complex cell responses, and features of intermediate complexity were significantly different over timescales of tens of seconds depending on the ongoing task. When subjects approached targets, the contrast at fixation was reduced significantly while in the case of avoiding objects, contrast was significantly elevated when compared to random selection of gaze in the scene.

The results also challenge the common notion of saliency as a task independent selection process. The hypothesis, that contrast or high edge density automatically attract gaze is clearly not valid for the data presented. When subjects approached object by walking towards them, contrast at fixation location was reduced significantly wen compared to randomly selected regions in the visual field.
4.4.2 Model simple and complex cell responses

The typical kurtotic distribution of linear filter responses to natural images has been reported repeatedly (Simoncelli & Olshausen, 2001) and is evidence for the non Gaussianity of the distribution of luminance values in natural images. Here the responses to model simple cells at multiple orientations and spatial scales were obtained and compared for the two tasks of approaching and avoiding target objects during walking. Significant differences were found for these filter responses over timescales of ten of seconds. When approaching targets by walking towards them, horizontal edge filter responses were significantly lower than when avoiding obstacles, while the edge filter responses at vertical orientations did not show these differences. Similar to the model simple cell responses, the analysis of the complex cell responses was significantly different for horizontal orientations and did not show this difference for filters oriented horizontally.

4.4.3 Selectivity of gaze targets

In order to interpret the differences in the contrast, simple and complex cell responses described above, the simple cell responses were clustered and reprojected in the image space. These features of intermediate complexity demonstrated that the vast majority of fixations were directed to only five feature clusters. The first two such features were a solid areas for the litter objects and the obstacles. The next three clusters were edges the blue object in different directions and orientations. This result demonstrates, that from the large number of possible features present in the environment, only very few were selected most of the time.

The fact that subjects tended to direct their gaze closer to the edge of obstacles when avoiding them can therefore explain why the proportion of gaze directed towards the walkway increased from the pick-up to the avoid condition, as noted when comparing the proportion of gaze times on object classes. When avoiding obstacles, gaze landed often close to the edge of the obstacle resulting in the fixation being classified as falling onto the walkway. The
additional analysis of the specific features at fixation location together with the map of the fixations relative to the pickup objects and the obstacles shows that the increased proportion directed towards the walkway reflects the difference in features targeted depending on whether subjects walked towards or around the object.

4.4.4 Feature statistics and task execution

All the feature dimensions of the visual input that were analyzed in the current study showed sensitivity to the ongoing task as determined by the task priorities given through verbal instructions and manifest in the different walking trajectories that human subjects chose to solve the respective task. These differences in the feature statistics can be explained in terms of the gaze targets that human subjects chose in the execution of the visuomotor tasks of avoiding and approaching objects during walkway navigation. Accordingly, it was found that the ongoing task determines these feature statistics and not that certain features such as high contrast reflexively attract gaze. Furthermore, these results demonstrate that simulating human gaze selection in natural images according to the distributions of eye movement parameters observed in humans can give a good indication of the average properties that feature statistics at fixation location can have, but that the variability of such statistics can be considerable, if the ongoing task is taken into account.
5 A credit assignment algorithm for the learning of visuo-motor policies

5.1 Introduction

The promise of reinforcement learning (RL) is to have an agent or a collection of agents learn how to solve a task in some optimal way based on the experience they accumulate while interacting with the environment. Although there are several different RL algorithms that can do this for small problems, they all have the drawback that the state spaces required to represent moderately realistic problems grow exponentially in the number of state variables. This problem is the so-called curse of dimensionality. This has made it difficult to apply RL to realistic settings and typically the statespaces considered are small and similarly the number of agents considered in multiagent problems is limited.

In order to overcome the scaling problem, two basic approaches have been developed. One way of addressing this problem is to try to exploit the structure of the problem and factor the problem in some way automatically (Dayan & Hinton, 1992; Barto & Mahadevan, 2003; Sutton et al., 1999; Parr & Russell, 1997). A more recent avenue of factoring the problem is to use graphical models in order to express the conditional independencies in such models so as to reduce the size of the variables necessary for a full description of the problem at hand (Craig Boutilier & Goldszmidt, 2000; Guestrin et al., 2003). The approach by Sallans and Hinton (Sallans & Hinton, 2004) can also be conceptualized as exploiting the statistical
structures of the state and action spaces. Doya et al. in (Doya et al., 2002) and (Samejima et al., 2003) use a number of actor-critic modules and learn a linear combination of the controllers for the local approximation of the policy.

The second approach to the scaling problem assumes that the overall reinforcement learning problem can be separated into independent tasks ab initio. Individual modules can be obtained by solving individual problems or by creating a complex from separate modules that each have learned what to do independently (Singh & Cohn, 1998). The main idea is that there will be a useful range of operation whereby the different modules can be used concurrently (Sprague & Ballard, 2003). This venue reintroduces a classical problem and that is the credit assignment problem. In a biological agent it is reasonable to assume that only a global estimate of reward $G$ is available. Thus different active reinforcement learning modules have the problem of dividing it up between them. The problem is compounded by the fact that at different times, different sets of modules may be active and also that, at any moment in time, the different active modules may be in different stages of completion.

Using a predetermined decomposition is a natural model for a multi-agent setting where global reward is given out and the different agents must discover how much each one has earned. This case was treated by (Chang et al., 2004). Their solution is based on the idea of modeling the individual reward that each agent should use in order to learn it’s policy as a local reward signal corrupted by Gaussian noise. The noise in this case is due to the rewards collected by the other agents at each moment in time. The algorithm uses a Kalman filter in order to obtain optimal estimates of the current reward pertaining to a particular agent from the global reward signal. Its convergence properties were analyzed by (Bagnell & Ng, 2006) for the case where the reward is collected only by some neighbors of each agent.

This paper describes a general solution that is valid for both the single-agent setting where multiple modules are simultaneously active and the multi-agent setting where each agent has a separate behavior. The assumptions are only that 1) the overall behavior of modules (or agents in the multi-agent venue) is such that they work together in different subsets and 2)
that the sum of the current estimates of the reward across all active modules is accessible to each individual module. Under these conditions we derive formulas for the estimates of reward that converge rapidly to their true values. Simulations of the method show that it is superior to (Chang et al., 2004). Furthermore, it is shown that such learning can be implemented in a multi-agent setting where individual modules communicate as in (Bagnell & Ng, 2006), and also in a single-agent case where individual modules consume the global available reward at a rate which is proportional to the individual current reward estimates. The latter model could be the basis for a biological implementation.

5.2 Background

The problem setting can be formalized as a Markov Decision Processes. An MDP consists of a 4-tuple \((S, A, T, R)\) with \(S\) being the set of possible states, \(A\) the set of possible actions, \(T\) the transition model describing the probabilities \(P(s_{t+1}|s_t, a_t)\) of reaching a state \(s_{t+1}\) when being in state \(s_t\) at time \(t\) and executing action \(a_t\), and \(R\) is a reward model that describes the expected value of the reward \(r_t\), which is distributed according to \(P(r_t|s_t, a_t)\) and is associated with the transition from state \(s_t\) to some state \(s_{t+1}\) when executing action \(a_t\).

Reinforcement learning attempts to find a policy \(\pi\) that maps from the set of states \(S\) to actions \(A\) so as to maximize the expected total discounted future reward through some form of learning. The dynamics of the environment \(T\) and the reward function \(R\) are not known in advance and an explicit reward function \(R\) is learned from experience. This method is usually associated with the model-based learning. The central idea for sequential tasks is to assign a value \(V^\pi(s)\) to each state, which represents this expected total discounted reward obtainable when starting from the particular state \(s\) and following the policy \(\pi\) thereafter:

\[
V^\pi(s) = E^\pi \left( \sum_{t=0}^{\infty} \gamma^t r_t \right)
\]

Alternatively, the values can be parametrized by state and action pairs, leading to the so-called “Q” values \(Q^\pi(s, a)\). The optimal achievable reward from a state \(s\) is then expressed as
V^*(s) = \max_a Q^*(s, a)  
and where the Bellman optimality equations for the quality values can be formulated as:

\begin{equation}
Q^*(s, a) = \sum_r r P(r|s, a) + \gamma \sum_{s' \in S} P(s'|s, a) \max_{a'} Q^*(s', a')
\end{equation}

Temporal difference learning (Sutton & Barto, 1998), uses the error between the current estimated values of states and the observed reward to drive learning. In a related Q-learning form, the estimate of the quality value of a state-action pair is adjusted by this error \( \delta Q \) using a learning rate \( \alpha \):

\begin{equation}
Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta Q
\end{equation}

The two fundamental learning rules in reinforcement learning are Q-learning (Watkins, 1989) and SARSA (Rummery & Niranjan, 1994). While Q-learning is an off-policy learning rule, i.e. it uses errors between current observations and estimates of the values for following an optimal policy, while actually following a potentially suboptimal policy during learning, SARSA\(^1\) is an on-policy learning rule, i.e. the updates of the state and action values reflect the current policy derived from these value estimates. Formally, in the case of Q-learning, the the temporal difference is:

\begin{equation}
\delta Q = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)
\end{equation}

while in the case of SARSA it is:

\begin{equation}
\delta Q = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)
\end{equation}

In the case of SARSA one substitutes equation (6) into (4) obtaining the weighted sum:

\begin{equation}
Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left( r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right)
\end{equation}

\begin{equation}
= (1 - \alpha) Q(s_t, a_t) + \alpha (r_t + \gamma Q(s_{t+1}, a_{t+1}))
\end{equation}

\(^1\)SARSA is an acronym for the quintuple \( s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1} \) denoting the actual trajectory followed.
5.3 Modular Behaviors

In the full description of the problem within the full joint composite state space $S$, a global policy $\pi$ maps from states $s$ to actions $a$. The relationship between the optimal value functions for each of the individual component tasks and the global task in which multiple objectives are pursued depends on the structure of the problem and can be very complex. Our main task is to define a restricted venue where the required behavior can be realized with separate RL modules. The main assumptions are that 1) modules are independent of each other and 2) are activated in subsets. The second premise is that momentary situations can be handled by selecting the right small set of modules (Sprague et al., 2006; Guestrin et al., 2003).

Specifying the detailed description of a module and its selection processes may be formidable problems in themselves, but for the purposes of explication here, we finesse this difficulty by picking problems from the literature. Other simulations e.g. (Sprague et al., 2006) hint that the strategy of designing modules and activation strategies in specific cases may be practical.

A module is defined as a MDP i.e. the $i$-th module is given by

$$ M_i = \{ S_i, A_i, T_i, R_i \} $$

where the subscripts reflect that the information is referred to the particular MDP. The states $s^{(i)}$ of each module $i$ are assumed all non-overlapping. In this decomposed formulation, each module can follow its own policy $\pi^i$, mapping from the local states $s^i$ to the local actions $a^i$. If in such a case, the optimal value function is readily expressible in terms of the component value functions, if the states and actions are fully factored so that there is no overlap and additionally the following two conditions hold:

$$ P(s_{t+1}^{(1)}, \ldots, s_{t+1}^{(N)}, s_t^{(1)}, \ldots, s_t^{(N)}, a_t^{(1)}, \ldots, a_t^{(N)}) = \prod_{i=1}^{N} P(s_{t+1}^{(i)}, s_t^{(i)}, a_t^{(i)}) $$

$$ P(r_t^{(1)}, \ldots, r_t^{(N)} | s_t^{(1)}, \ldots, s_t^{(N)}, a_t^{(1)}, \ldots, a_t^{(N)}) = \prod_{i=1}^{N} P(r_t^{(i)} | s_t^{(i)}, a_t^{(i)}) $$
Then these two conditions can be used together with equation 5.2 in order to arrive at the result:

\[ Q(s, a) = \sum_{i=1}^{N} Q(s^{(i)}, a^{(i)}) \]  

(5.10)

If equations 5.8 and 5.9 hold, and all the rewards are known, the action maximizing equation 5.10 can be selected, and is guaranteed to be optimal. However the case of interest is that where the individual rewards are not known, but only the global reward is supplied to the agent at each time step.

An important sub-case is that of a single agent pursuing multiple goals, which are divided up between multiple independent modules (Humphrys, 1996; Karlsson, 1997; Singh & Cohn, 1998; Sprague & Ballard, 2003). The main modification in this instance is that the action space is shared such that \( a^{(i)} = a \), for all \( i \), so the \( i \)-th module is now:

\[ M_i = \{S_i, A, T_i, R_i\} \]  

(5.11)

This case requires some form of action selection in order to mediate the competition between actions proposed by individual modules. Once the action has been selected, all modules execute the same selected action.

5.3.1 Module activation

At any instant, it is assumed that a good set of modules can cover an agent’s needs. Each module is potentially activated by a scheduler. The consequence of a module being activated is that:

1. Some associated procedure, such as a visual routine (Ullman, 1984; Ballard et al., 1997), computes the initial state the module is in.

2. Its Q-values are included in the sum indicated in Eq. 5.10, and

3. It influences the global reward that is handed out at every time step.
Figure 5.1: The credit assignment algorithm exploits the constraint that during behavior, in any period there is only a subset of the total module set that is active. We term these periods episodes. In the timecourse of behavior, modules that are needed become active and those that are no longer needed become inactive. The diagram depicts two sequential episodes of three modules each \{3,4,7\} and \{2,8,10\}. The different modules are denoted with different shadings and numbers. The different lengths indicate that modules can exhibit have different numbers of states and finish at different times. The horizontal arrows denote the scheduler’s action in activating and deactivating modules. On the right is the large library of possible modules. Our formal results only depend on each module being chosen sufficiently often and not on the details of the selection strategy.
To implement module activation, we introduce the constraint of an episode of fixed length $\Delta T$. During each episode, only a subset of the total module set is active. The guiding hypothesis is that in the timecourse of behavior a certain set of goals is pursued and therefore the corresponding set of modules that are needed to achieve these goals become active and those that correspond to tasks that are not pursued become inactive (Sprague et al., 2006). During an episode the composition of a particular module set is assumed to not change. Given this constraint, the pivotal idea is that a module can refine its own reward estimates by having access to the sum of the reward estimates of the other active modules.

The scheduler chooses the set of modules every episode based on some activation policy (See Figure 5.1). Two activation policies that we use are: 1) To show the general properties we have sets of identical modules and pick them at random such that the subset is less than some small maximum limit, typically $\leq 4$ (If the number of modules is greater than the maximum limit, the choice has to be resolved in a way dictated by the problem requirements), 2) For the problems from the literature, we choose subsets of modules that can address the particular problem situations.

5.4 Credit Assignment with modular behaviors

Although each of the individual modules may represent some portion of the entire state space and may execute some part of the composite action, a global performance measure is optimized across all agents. While some problems have been formulated so that the individual sources of reward are kept separate and are accessible to the modules (Sprague & Ballard, 2003; Singh & Cohn, 1998), a more natural performance measure is the total discounted return collected by the individual modules (Chang et al., 2004). This is achieved by using a reward function, which is based on the sum of the rewards obtained by the individual modules, i.e. summing
the individual rewards collected by all of the $N$ modules at each time step to a global reward:

$$G_t = \sum_{i=1}^{N} r^{(i)}_{t}.$$  

(5.12)

The central problem that we tackle is how to learn the composite quality values $Q^{(i)}(s^{(i)}, a^{(i)})$ if only global rewards $G_t$ are directly observed, but not the individual values $\{r^{(i)}_{t}\}$.

The key constraint is the assumption that the system can use the reward information from the modules that are active at any instant. This knowledge leads to the idea to use the different sets to estimate the difference between the total observed reward $g_t$ and the sum of the current estimates of the individual rewards of the concurrently running behaviors. Credit assignment is achieved by bootstrapping these estimates over multiple task combinations, during which different subsets of behaviors are active. This reasoning can be formalized as requiring the individual behaviors to learn independent reward models $R^{(i)}(s^{(i)}, a^{(i)})$. The current reward estimate for one particular behavior $i$, dropping the temporal subscript, is obtained as

$$\hat{r}^{(i)} = \hat{r}^{(i)} + \beta \delta_{r}^{(i)}$$  

(5.13)

where the error on the reward estimates $\delta_{r}$ is calculated as the difference between the global reward and the sum of the component estimates:

$$\delta_{r}^{(i)} = G - \sum_{j} \hat{r}^{(j)}$$  

(5.14)

so that equation 5.13 becomes:

$$\hat{r}^{(i)} = \hat{r}^{(i)} + \beta \left( G - \sum_{j} \hat{r}^{(j)} \right)$$

(5.15)

Together with the module activation protocol, Eq. 5.15 represents the core of our solution to the credit assignment problem. When one particular subset of tasks is pursued, each active behavior adjusts the current reward estimates $\hat{r}$ in the individual reward functions according to
Figure 5.2: A fundamental problem for a biological agent using a modular architecture. At any given instant, shown with dotted lines, when multiple modules are active and only a global reward signal $G$ is available, the agents each have to be able to calculate how much of the rewards is due to their activation. This is known as the credit assignment problem. Our setting simplifies the problem by assuming that individual reinforcement learning modules are independent and communicate only their estimates of their reward values. The modules can be activated and deactivated asynchronously, and may each need different numbers of steps to complete, as suggested by the diagram.
equation 5.15 at each time step. Over time, the set of tasks that have to be solved will change, resulting in a different set of behaviors being active, so that a new adjustment is applied to the reward functions according to equation 5.15. This bootstrapping process therefore relies on the assertion that the subsets of active behaviors visits all component behaviors.

The component quality values for the state-action pairs of the individual behaviors are learned using the above estimates of the individual reward functions. Given the current reward estimates obtained through repeated application of equation 5.15, the SARSA algorithm is used to learn the component Q-functions:

\[
Q_i(s_t^{(i)}, a_t^{(i)}) \leftarrow Q_i(s_t^{(i)}, a_t^{(i)}) + \alpha \delta_{Q_i}
\]

where \(\delta_{Q_i}\) is given by:

\[
\delta_{Q_i} = \hat{r}_t^{(i)} + \gamma Q_i(s_{t+1}^{(i)}, a_{t+1}^{(i)}) - Q_i(s_t^{(i)}, a_t^{(i)})
\]

The usage of an on-policy learning rule such as SARSA is natural, as noted in (Sprague & Ballard, 2003; Russell & Zimdars, 2003), because the arbitration process may select actions that are suboptimal for one particular module. A feature of the SARSA algorithm is that it makes use of suboptimal policy decisions.

### 5.4.1 Proof of Convergence

Assume that if the modules are seeded with their correct rewards, that the setting is such that they learn their correct Q-values. Now we ask: in such a setting, will our algorithm for computing rewards, given only global reward together with an arbitrary initial setting of the rewards, converge also? Equation 5.15 represents essentially a version of a standard iterative method for solving an overdetermined set of linear equations. Thus the convergence of such a system is quadratic provided all the potential variables are included in the equations. In Q-learning, there is potentially one reward for each action so that the total number of rewards \(N_r\) is bounded by \(N_m N_A\) where \(N_m\) is the number of modules and \(N_A\) is the largest number...
of actions in any module. Suppose that the modular system is run for sufficient iterations so that the linear system converges. During this time further suppose that no actual updating of the Q values occurs but simply the action space is sampled sufficiently often so that the linear system of reward equations converges. At this juncture, the rewards are correct and Q-learning can be used. Since by supposition, the individual modules converge, the system of modules will now also converge as they all will be using the correct rewards.

The above method is of course very wasteful because our simulations show that the rewards can be solved simultaneously with Q-learning.

5.4.2 A Biological Possibility for Implementing Consumable Rewards

Equation 5.15 describes learning that does not require an individual behavior to ‘know’ exactly which other behaviors are concurrently running but only the sum of their reward estimates. Also, it is not necessary to communicate the current reward estimates between modules. Instead, using the consumable rewards, each module only needs to observe the difference between the global reward value and the sum of all the current estimates of the other modules. The use of the name consumable reflects the fact, that under the assumption of rewards being absorbed at a rate proportional to the total amount of reward $G$ and the current reward estimate $t_i$ of each individual behavior, one obtains a temporal difference error in the total consumption as in equation 5.15. The following section makes this notion more precise.

The consumable rewards model is motivated by the fact, that biological systems use the phasic signal of dopaminergic neurons to encode reward-related signals in different tasks (Schultz, 2000). This leads to the assumption that this activity is not dependent on the individual task the organism is executing, but is related to the reward prediction error. Accordingly, if several tasks are executed concurrently, this implies that some combination of these rewards must be represented. Here it is assumed that the total reward is the linear sum of the individual rewards obtained by all individual modules together, as expressed in equation 6.1.
Under the assumption that some global reward $G$ is available at some moment in time and each of the $N$ behaviors consumes its share of reward at a rate $c_i$, one can set up a system of linear differential equations describing the amount of consumed reward as a function of time. Let $x_i(t)$ be the amount of reward consumed up to time $t$ by process $i$, which absorbs reward at a rate $c_i$. The latter term is assumed proportional to its current estimate of its share of the global reward $r_i$. Furthermore, the consumption rate is proportional to the amount of still available reward, i.e. the global reward $G$ minus the sum of already consumed reward.

For the processes consuming reward, it is proposed that the respective rates (using continuous notation) $\{\dot{x}_i\}$ at which reward is consumed are given by:

$$\dot{x}_i = c_i \left(G - \sum_{i=1}^{N} x_i(t)\right), \ i = 1, \ldots, N$$

(5.18)

This system of differential equations is straightforward to solve yielding the result:

$$x_i(t) = -\frac{c_i}{\sum_{j=1}^{N} c_j} G \left( -1 + e^{-\left(\sum_{j=1}^{N} c_j\right)t} \right)$$

(5.19)

Assuming that the time constants determined by the absorption rates are much shorter than those of the reinforcement dynamics, this leads to the steady state solution:

$$\lim_{t \to \infty} x_i(t) = \frac{c_i}{\sum_{j=1}^{N} c_j} G$$

(5.20)

which states that each individual process absorbs an amount that is proportional to the total global available reward and the individual proportion of the absorption rates. Now, assume that the absorption rate $c_i$ for each individual learning process is proportional to the current reward estimate, i.e.:

$$c_i = k \hat{r}^{(i)}$$

(5.21)

where a constant of proportionality $k$ has been introduced in order to convert from absorption rates to amounts of reward absorbed. It is assumed that this rate is the same for all participating
processes. The global reward $G$ is independent of the current estimates and given by equation 6.1. Accordingly, the steady state solution in equation 5.20 becomes:

\[
\lim_{t \to \infty} x_i(t) = \frac{\hat{r}^{(i)}}{\sum_{j=1}^{N} \hat{r}^{(j)}} \sum_{j=1}^{N} r^{(j)}
\]

(5.22)

Finally, the difference between the actually absorbed reward and the expected reward for each module $i$ is given by:

\[
\delta r_i = \left( \frac{\hat{r}^{(i)}}{\sum_{j=1}^{N} \hat{r}^{(j)}} \sum_{j=1}^{N} r^{(j)} \right) - \hat{r}^{(i)}
\]

(5.23)

Note that this corresponds to the expression given in equation 5.15 only up to the leading factor.

Although the connection is still tenuous, the phasic activity of dopaminergic neurons in rodents and primates during reward related learning has been shown to agree well with the temporal prediction error in reinforcement learning (Schultz et al., 1997). The relationship between this spiking activity and the release of dopamine has started to emerge by using new electrochemical techniques (R.M. & D.L., 2002). But it is still not fully understood how all the various factors such as, the activity of midbrain dopaminergic neurons, the resulting levels of dopamine released into the intracellular medium, and the rates of dopamine uptake, are all related to the variables describing the task related reward. Recent advances have made it possible to measure transient changes in dopamine concentration during food seeking (Roitman et al., 2004), while freely moving (Montague et al., 2004), and during self-administration of drugs (P.E. et al., 2003). Using fast-scan cyclic voltammetry, which allows measuring these levels with subsecond temporal resolution, it was established (Roitman et al., 2004) that dopamine in the nucleus accumbens of rodents is modulated at timescales of 0.1 seconds. Such fast changes in the dopamine levels are in accordance to the small time constants necessary in
the above algorithm, which requires the consumable rewards step to be much faster than the reinforcement learning step.

5.4.3 Dealing with uncertainty

By assuming that each module observes the global reward according to equation 6.1, the inference task is to find the individual rewards for all participating modules. Under the assumption of Gaussian observation noise, this problem can be cast as a Kalman filter. However, this would require each module to track or know by some form of communication, which state all other modules are in. This problem is therefore only interesting as a limiting case in order to establish how well one could do with perfect knowledge about the states all agents are in.

Previously, Chang, Ho, and Kaelbling (Chang et al., 2004) used a Kalman filter in the multiagent setting in such a way, that each individual agent had knowledge about its state and the sum of the reward collected by all other agents was modeled as noise with a Gaussian distribution. Although in their setting it was possible to learn approximate policies for most agents, the crucial problem with that approach is that the Kalman filter is not able to recover the individual rewards. As described in (Chang et al., 2004), all reward estimates of each individual agent are only accurate up to an additive constant. This will not affect the policy the agents learn, but it affects the value function. It is well known that adding a constant to the reward function changes the value function but does not change the policy (Ng et al., 1999). But a small additive constant $c$ in the reward estimates results in a large difference in the corresponding value function when $\gamma$ is near unity as shown by the following:

$$V'\pi(s) = E_\pi \{ \sum_{i=0}^{\infty} \gamma^i (r_{t+1+i} + c) \}$$

$$= \frac{c}{1 - \gamma} + E_\pi \{ \sum_{i=0}^{\infty} \gamma^i r_{t+1+i} \}$$

$$= \frac{c}{1 - \gamma} + V^\pi(s)$$

(5.24)
Equation 5.24 means that although it may be possible to learn some of the policies for component modules for one particular task combination, the value functions will be corrupted by a large bias, which will be especially problematic, when new task combinations are to be solved. The reward estimates will also be biased such that they have to be relearned as well, but will again be biased. Furthermore, it has been shown (Sutton, 1992) that although the Kalman filter is Bayes optimal by providing the estimate which is the expected value of the latent variable given all previous observations, the situation is quite different, if the process noise is not known exactly a priori, which is most commonly the case in RL.

The consumable rewards algorithm overcomes the bias problem by assuming a venue where it can use modules in different combinations. However this strategy has its own drawback. When one particular subset of goals is pursued in any particular episode, the corresponding behaviors are active, and the estimates of the respective rewards is updated according to equation 5.15 for all component behaviors. This means, that on a particular task combination, all component behaviors update their respective reward estimates in the same direction, independent of the fact, how well component behaviors have already estimated their share. Thus with the above scheme it is possible for a behavior to unlearn good reward estimates if it is combined with other behaviors whose reward estimates are far from the true values. Accordingly, learning can be made much more efficient, by considering the respective uncertainties in the estimates of the respective rewards.

In order to integrate uncertainties into the estimation of the reward values using only local computations, the following two approximations can be used. First, when limiting the covariance of the estimate in the Kalman filter setting to be a diagonal matrix, one can express the gain for each reward estimate in terms of the individual uncertainties in the respective reward estimates $(\sigma(i))^2$:

\[
\beta_i = \frac{(\sigma(i))^2}{\sum_{j=1}^{N} (\sigma(j))^2}
\]

\[
= \frac{(\sigma(i))^2}{\sum_{j \neq i} (\sigma(j))^2 + (\sigma(i))^2}
\]

(5.25)
where the last term in the denominator is variance in the observation noise. The diagonal covariance matrix approximation reflects the observation that the variation in reward estimations are primarily due to within-module fluctuations, as opposed to between-module fluctuations, which tend to be uncorrelated.

Expanding the sum in the denominator in equation 5.25 suggests the second approximation, in which each individual agent uses an online estimate for the variance \( \sum_{j \neq i} (\sigma^{(j)})^2 \) by tracking the variance in the difference between the global reward \( G \) and the sum of the reward estimates of the other agents \( \sum_{j \neq i} \hat{r}^{(j)} \). In the following simulations each agent tracked this difference using a recursive least squares estimator with exponential forgetting, and maintained the uncertainty about the rewards of individual state action pairs \( (\sigma^{(i)})^2 \) locally.

### 5.5 Simulation results

We illustrate the advantages of the algorithm on two standard problems from the literature. The first is a grid world problem described by Sutton and Barto (Sutton & Barto, 1998) (See Figure 5.3). The reason for picking this multi-agent problem is that it was chosen by (Chang et al., 2004) to illustrate their reward estimation algorithm. Thus the various advantages of the consumable rewards algorithm can be illustrated with a replication of their previous results as a baseline of comparison. Unlike Chang et al, the consumable rewards algorithm is able to compute the correct Q-values as part of the reward estimation. We demonstrate that, unlike the Chang et al algorithm, consumable rewards can tolerate the mix of modules changing in the course of the computation. Finally we show that the Bayesian method of weighting estimates by the inverse of their variances produces the best convergence results.

The second set of comparisons use variations of problems by Singh and Cohn (Singh & Cohn, 1998). These single-agent problems have additional structure which comes closer to representing some of the problems that would have to be addressed by a more elaborate biological agent. Unlike the multi-agent problem, here the action space is shared by the modules.
Modules are indexed by food source types and predator types. That is, we assume that when either a food source or predator is within range of the sensors, its appropriate behavior is made active, and when out of range, the respective behavior is made inactive. Given the motion of the agent and the related appearance and disappearance of food sources and predators, the mix of active modules is constantly changing.

Finally, the third set of simulations uses a model of the walkway navigation task that was first considered by Sprague and Ballard (Sprague et al., 2006) and which was also the basis for the psychophysical experiments described in the previous chapters. The task is separated into three component behaviors of avoiding obstacles, approaching targets, and following the walkway. The simulations presented here are mainly intended to demonstrate that the approach of consumable rewards is in principle feasible for such a problem. As such, the parameterization of the statespace, the reward values, and most parameters of the simulation follow those described in (Sprague et al., 2006).

For all these simulations, the RL learning parameter was 0.1. The first set of experiments use both set \( \beta \) values as indicated as well as the \( \beta \) computed with the variance formula. The second set of experiments use set \( \beta \) values as indicated.

5.5.1 The correct computation of Q-values

In the multi-agent version of the Barto and Sutton Problem each agent is placed on a 5 × 5 grid and is able to move in 4 directions, i.e. North, East, South, and West. A transition between positions results in a reward of 0. If a movement is made toward the walls of the grid, the agent obtains a reward of -1 and stays at the same location. There are two locations on the grid, which result in a transition to a new state for all selected actions. On one of these a reward of 20 units is obtained while in the second case a reward of 10 units is obtained. The problem is set up in such a way that the optimal policy successively collects the reward of 20 units, given a discount factor of 0.9. As demonstrated by (Chang et al., 2004), this grid world problem can
be used to test the multi agent setting by having a number of agents learn an optimal policy independently of each other from a global reward signal.

Figure 5.4 demonstrates the result reported by (Chang et al., 2004) in which the agents are able to converge to policies that are optimal or near optimal using a Kalman filter to estimate their contribution to the global reward. The top right panel in the figure demonstrates the principal problem with this approach: the bias in the reward estimates present in all reward functions for each agent lead to value functions with a sum of squared errors to the optimal value function of more than $10^5$.

The consumable rewards algorithm applied to this problem did not exhibit this bias. For the simulation, after each $\Delta T$ of 30 iterations, a new subset of agents of sizes between 3 and 7 were chosen randomly. The scheduler chose these subsets according to a uniform distribution over all modules. These agents were then run on the grid world and learned their respective Q-values with reward estimates that were updated according to equation 5.15. The bottom left panel of Figure 5.4 shows that the agents learn to divide up the global reward and are therefore able to learn their respective policies.

A variation on the problem simulation demonstrates another problem with modeling the contribution of the other agents to the global reward as noise. Figure 5.5 shows an experiment in which the filtering agents try to estimate their respective rewards from the global reward, but different subsets of tasks have to be executed at each moment in time. By switching to different subsets of active agents, the estimates of the noise die to the other agents changes with each change in subtasks. Accordingly, the Kalman filter performs poorly and most agents do not converge to the optimal policy. Moreover, most agents often unlearn the reward estimates over time when new task combinations have to be executed resulting in the oscillations in the collected reward as seen in Fig. 5.5. Thus the natural problem setting of episodes, which is a boon for consumable rewards, is devastating for the standard Kalman filter algorithm.

Focusing again on the consumable rewards algorithm, we ran experiments to compare the weighting of reward estimates by their variance estimates with weighting using a more standard
Figure 5.3: Schematic representation of the grid world examples used in the simulations. From left to right: A) This example is taken directly from Sutton and Barto’s book where it is used in the single agent case. The agent is able to move in the four directions up, right, down, and left. Each transition results in a reward of 0 with the exception of the two states from where the arrows start. In those states any action transitions the agent to the shown states and a reward of either 10 or 20 can be collected. B) This grid world example follows Singh and Cohn. An agent is located on a 5x5 grid and searches to find three different food sources f1 to f3 and tries to avoid the predator p, which moves every other time step towards the agent. C) This grid world problem is derived from the previous by increasing the size of the total area of the world within which the agent can only perceive food sources and predators within the perceptual range.
Figure 5.4: (Top left) Most of the Chang et al filtering agents converge to the optimal policy but a small number converges to a suboptimal policy and are not maximizing their average reward. (Top right) The root mean square error between the optimal policy found through value iteration and the policy found by the Chang et al filtering agents. Note that although the policy that an agent follows is optimal, the value function it found may have a very large error because of the random additive constant in the reward estimates. (Bottom left) Using the consumable rewards, agents converge to the optimal policies. (Bottom right) The sum of squared differences between the optimal and the found value function decrease to zero using consumable rewards.
Figure 5.5: Agents that assume that the variation in reward is noise and can be tracked with a Kalman filter have trouble when the mix of behaviors keeps changing. This is again evident in the collected rewards (left) and the RMS-error between the computed and the optimal value function (right). Consumable rewards has no such problem and in fact exploits the different mix of behaviors in order to track the reward accurately.

5.5.2 Multiple sources and reactive scheduling

The Singh and Cohn problems are two learning problems described in (Singh & Cohn, 1998) in which the authors explore the use of multiple tasks in a grid-world problem. In the original formulation, an agent moves on a a five by five grid. Possible actions move the agent in the eight compass directions. Moves at the edge of the grid-world which would result in the agent leaving the grid result in the agent staying in the current position. The grid is populated by three food items and a predator. The picking up of a food item results in a reward of one unit. After reaching a food item, this is repositioned at a random location on the grid. The world is also populated by a predator, which moves every other time unit towards the current position of the agent. The agent obtains a reward of 0.5 units for every time step in which it does not collide with the predator. Each learner represents the position of the respective
Figure 5.6: Effect of different learning rates on the accumulated reward and comparison to the variance-weighted learning as well as a Kalman Filter that has perfect state information about all component modules.
food item or predator, i.e. there are 25 states for each of the component learner, where in the original problem a total of four learners were always active in order to solve the four component tasks. Singh and Cohn (Singh & Cohn, 1998) and Sprague and Ballard (Sprague & Ballard, 2003) used this task in multi-goal learning but both papers used individual rewards that were delivered for each task as separate reward signals.

5.5.3 Learning many tasks

The above problem was modified so that the reward each behavior sees is only the global sum of the individual rewards. Moreover, the three food seeking behaviors are sampled randomly for each episode lasting 100 iterations from a pool of 75 different food sources. Therefore, the food seeking agents have to solve the credit assignment problem for each of the new combinations of foods present during an episode. Similarly, at each moment in time, one predator is present from a pool of 25 different predators. This means that at each moment in time, a total of three food sources and one predator are present in the gridworld, as in the original Singh and Cohn problem, but that these come from a pool of 75 and 25 respectively. Given that the scheduler starts the appropriate module for each food source and predator reactively, the subset of five modules from the total of 100 modules is active during each episode lasting 100 iterations. The rewards for each food source and each predator are then bootstrapped across different episodes. Figure 5.7 shows the average reward collected using different values for $\beta$, the learning rate in the credit assignment algorithm. As can be seen in the figure, individual behaviors can learn their respective contributions to the global reward from different task combinations if the learning rate $\beta$ has intermediate values between 0.01 and 0.1.

5.5.4 Learning tasks with environmentally driven switching

A second task was devised, in which the agent is moving on a grid-world of size $20 \times 20$. The agent has a perceptual radius of 5 by 5 units, i.e. the agent only represents the objects
Figure 5.7: Average reward collected using consumable rewards. Simulations were run in which three food sources from a set of 75 and one predator from a set of 25 were always active. The three correspond to different learning rates $\beta$ in equation 5.15.
present within that proximity. The agent activates the necessary behaviors depending on the objects present in this area. If for example only two different food sources are present within this 5 by 5 area, only behaviors for those food sources are active. The agent therefore learns asynchronously, because the set of active behaviors is dependent on the structure of the state space an how the agent moves through this space. For each of the three food types there were 16 foods present on the 20 by 20 grid resulting in a total of 48 foods of three types. Similarly, a total of 10 predators were always present on the grid. The state representations and reward structure for each individual task were again as in the original formulation by Singh and Cohn. This problem was mainly introduced to demonstrate that the proposed algorithm works under a reactive scheduling mechanism that leads to random episode lengths that are started and ended asynchronously for individual modules. Figure 5.8 shows the sum of squared differences between the true reward values for all behaviors and the current estimates over trials for different learning rates $\beta$. The average reward collected is less informative in this case, because it can not be compared in a direct way to the reward collected in the previous examples. Note, that in this case, even with learning rates of $\beta = 0.001$ and $\beta = 0.5$ the reward values are learned.
Figure 5.9: The walkway navigation tasks. Left: typical view from the agent while navigating along the walkway. The three possible tasks are following the walkway, avoiding obstacles, which are the dark cylinders, and picking up litter, corresponding to the light cylinders. Right: Schematic representation of the statespace parameterization for the learners. Each module represents the distance to the closest object in the field of view and the angle between the current heading direction and the object center axis. The module learning the walkway behavior uses the signed distance to the midline of the walkway and the angle between the heading direction and the vector in the direction of the walkway.

correctly. The reason for this is that due to the structure of the task and the distribution of the food sources and predators sometimes the agent is in a part of the state space in which only a single task has to be solved. In such a case the agent observes global reward whose only component is the reward for that specific task and can learn the reward for that behavior directly form the observed reward.

5.5.5 Learning walkway navigation

The walkway navigation task was first considered by Sprague and Ballard (Sprague et al., 2006) where a factorized solution was presented. However, that solution was obtained by delivering each of the individual learners their respective reward, i.e. the agent received three separate
Figure 5.10: Representations of rewards, value functions, and policies in the walkway navigation task for the three component behaviors at iteration 1 and iteration 1000. Top row: reward estimates $\hat{r}(s)$. Middle row: value functions $V(s)$. Bottom row: policies $\pi(s)$. From left to right. First column: iteration 1 for obstacle avoidance. Second column: iteration 1 for litter pickup. Third column: iteration 1 for walkway following. Fourth column: iteration 1000 for obstacle avoidance. Fifth column: iteration 1000 for litter pickup. Sixth column: iteration 1000 for walkway following.
rewards, one for the walkway following module, one for the obstacle avoidance module and one for the litter picking up module. This problem was considered here with only a global reward value being observed by all modules in each task combination. The global reward was always the linear sum of the rewards obtained by the individual modules according to equation 6.1. The parameterization of the statespace is shown in figure 5.9. Each module represents the states with a two dimensional vector containing a distance and an angle. For the picking up and the avoidance behaviors, these are the distance to the closest litter object and obstacle respectively and the signed angle between the current heading direction and the direction towards the object. The distance is scaled logarithmically as in the original setup (Sprague et al., 2006) using the equation:

\[
d_i = \lceil (1.0 - 2^{(-0.5/\rho_i)})n \rceil
\]

and the resulting distance \(d_i\) is then discretized into one \(n = 21\) possible values between 0 and infinite distance. The angles within the field of view i.e. with a magnitude smaller than 50 degrees are similarly discretized to 21 values. The walkway statespace is slightly different from (Sprague et al., 2006) in that it represents all positions of the agent relative to the walkway for all possible walking directions. Finally, instead of 3 possible actions as in (Sprague et al., 2006) the current simulations use 5 actions corresponding to steering at one of the five angles \{-15, -7.5, 0, 7.5, 15\} with additive Gaussian noise of variance \(\sigma = 1\).

The resulting reward estimates, value functions, and policies of these simulations are shown in figure 5.10 at the first iteration with random initial values and after learning, when the agent has walked the walkway for 1000 episodes. As can be seen from the representation of the reward estimates, the individual behaviors have learned the true rewards of their respective tasks, where not intersecting with an obstacle results in a reward of one unit, intersecting a litter object gives four units of reward, and staying on the walkway results in a reward of 0.8 units. The figures of the reward estimates also demonstrate that a function approximation scheme should be better at capturing structure in the reward space such as smooth reward landscapes, reward functions with only one state being rewarded, or separate areas with discrete rewards.
The figures of the Value functions demonstrate qualitatively that the correct functions have been learned. Finally, the policies demonstrate the subdivisions of the statespace into regions in which different actions are selected. For example, the obstacle avoidance policy is to walk to the right when the obstacle is close and to the left and vice versa for obstacles on the other side. For obstacles at larger distances the differences in the values are so small, that the policy is almost indifferent with respect to the actions.

5.6 Discussion and Conclusions

We have described a way that individual behaviors can estimate reward given only the momentary global reward. The key constraint, motivated by considerations of how to structure modular systems, is to use modularized behaviors in different combinations. This allows the rewards obtained by reinforcement to be estimated iteratively. In addition this formulation lends itself to the use of Bayesian techniques for combining reward estimates amongst modules that speeds convergence of the estimating process.

The present work is related to other approaches that start out with compositional solutions to individual problems and then devise methods in order to combine a large number of such elemental solutions e.g. Singh and Cohn (Singh & Cohn, 1998) and Meuleau et al. (Meuleau et al., 1998). Both approaches are concerned with learning solutions to large MDPs by utilizing solutions or partial solutions to smaller component MDPs. In (Meuleau et al., 1998) the solutions to such components are heuristically combined to find an approximate solution to the composite MDP by exploiting assumptions on the structure of the joint action space. A way of learning a composite MDP from individual component MDPs by merging has been described in (Singh & Cohn, 1998). First the policies for elemental component MDPs are learned by separate modules. In their formulation, when combined, the different modules share a common action space, i.e. one single action is chosen in the composite problem. This is often used as the criterion to distinguish between single agent modular and multiagent problems. Because
of this selection of a single action, each individual action is not necessarily optimal for all the component MDPs. Given that the component value function is not necessarily obtained as linear sum of the component policies, the authors propose a way of initializing the value iteration process of the composite problem using bounds on the state values derived from the state values of the individual component MDPs.

The credit assignment problem is an important problem for multi-agent systems, but even more so for biology as a behaving animal in a complex environment has to solve this problem. The solution here is very compatible with the numerous experiments that suggest that such reward is represented internally as the neurotransmitter dopamine. Individual processes could potentially realize the algorithm here by communicating with dopamine absorption rates.
6 Extraction of human reward weighting

6.1 Introduction

Considerable empirical data collected with primates and humans under constrained experimental conditions has shown that the primate brain represents rewards in multiple brain regions, and especially in areas related to the visuomotor system (Platt & Glimcher, 1999; Glimcher, 2003; Dorris & Glimcher, 2004; Platt; Sugrue et al, 2004; Stuphorn et al, 2000; Hikosaka). Such data has been obtained by recording behavioral and brain activity data while participants were executing a task, that was designed specifically by the experimenter in order to test a specific reward related question. Participants first have to learn the reward structure of the environment imposed by the experimental design. But how do these results relate to the extended actions sequences humans execute in every day life such as walking along a walkway, picking up objects, and avoiding obstacles? Here we present experimental results and an analysis of this data that demonstrates that reinforcement learning algorithms are able to describe such human behavior and that the rewards implicitly underlying the executed behavior can be recovered. In this way, support is given to the hypothesis, that the human brain may learn the programming of action sequences through algorithms that are able to solve the reinforcement learning problem and which have been shown to describe human learning in a variety of other circumstances under externally imposed reward schemes.
In previous chapters it was argued, that human vision is tightly linked to the ongoing goal directed behavior of the organism. Vision was though of as being used in order to obtain information from the environment when this is necessary during the sequence of actions and from the point in the scene at which relevant information can be obtained. Here it is assumed that the navigation behavior of subjects can be modeled by the previously described multi-module reinforcement learning algorithm and that therefore human subjects optimize their trajectory when navigating the walkway environment in such a way that some internal representation of a reinforcer is maximized. Under this premise, the current chapter suggests how to extract these reward values solely from the observed walking data of individual subjects in the different task conditions.

An important question when applying reinforcement learning algorithm to the modeling of human behavior is that of the units of the reward or cost respectively. While some experiments have used tasks with explicit rewards such as monetary rewards (Trommershäuser et al., 2003), liquid rewards (Seitz et al., submitted), abstract points (Paulus et al., 2004), or time (Gray et al., 2005), in natural tasks it is not so clear which basic unit of reward to use. Here the metabolic cost of walking was used as the common unit of reward in order to compare the weighting of individual tasks. The reason for this choice is that it is the only cost that can be extracted from the trajectories that subjects used in the tasks and that can be related to the ample empirical data quantifying this cost for walking at different speeds. This is only one particular solution to the problem of using a common currency and it does not argue that this is necessarily the currency used by animal brains. Especially in the case of humans and primates it is an important question, which non-primary rewards may work as reinforcer.

6.2 Methods

The goal of this chapter is to describe the observed human navigation behavior in terms of the proposed reinforcement learning algorithm, which uses a factorization of the individual compo-
nent tasks. This type of general problem is called the inverse-reinforcement-learning problem (Ng & Russell, 2000) and is usually framed in terms of recovering the reward function underlying an agent’s behavior, given observations of state action pairs and the full transition function describing the dynamics of the environment. Several algorithms for solving this task have been proposed recently (Ng & Russell, 2000; Abbeel & Ng, 2004; Ramachandran & Amir, 2007; Neu & Szepesvári, 2007), but their applicability for inferring the reward structure in complex human behavior has not been well established. The current situation is that these algorithms seem to work well on restricted types of reinforcement learning problems such as those considered in the cited papers. Such tasks comprise random Markov Decision Processes (MDPs) in which the rewards are obtained from a known prior probability distribution or navigation of a car on a highway in a computer game in which only three possible actions executed through key presses were allowed. Testing of these algorithms with human reaching data has not resulted in consistent estimation of reward functions with sufficient accuracy (Paul Schrater, personal communication).

The method for inverse reinforcement learning proposed here is based on the algorithm previously proposed by (Ramachandran & Amir, 2007). This method was altered in order to exploit as much previous knowledge about the considered problem and include as many available constraints as possible. To this end, several simplifications needed to be made. These simplifications are described in the following sections, after first briefly reviewing the method by (Ramachandran & Amir, 2007).

### 6.3 Inverse Reinforcement Learning

The basic setup of the IRL problem used here follows previous formulations with the difference being that the state representation is factorized, as in previous chapters. This type of modularization is especially convenient here, because it allows to efficiently estimate the unknown reward function by obtaining samples for each individual component module. Again,
the starting point of the factorization is to assume that the global reward obtained is compose of individual rewards from the component tasks:

\[ G_t = \sum_{i=1}^{N} r_t^{(i)}. \]

(6.1)

In the inverse RL formulation by Ramachandran and Amir (Ramachandran & Amir, 2007) the goal is to obtain the mean of the posterior distribution over reward functions given the observed state-action pairs. It is assumed that the observed state-action pairs are independent, identically distributed samples of a distribution over Q-values of the form:

\[ P((s_j, a_j)|R) = \frac{1}{Z} e^{\alpha Q^*(s_j, a_j, R)} \]

(6.2)

A prior distribution over reward values has to be chosen. As previously proposed, a Laplacian distribution over rewards was chosen according to the assumption, that in often considered tasks a large number of states exist that yield very small or no rewards whereas a few states deliver large rewards:

\[ P(R) = \frac{1}{2\sigma} e^{-\frac{|r|}{2\sigma}} \]

(6.3)

The inference problem is then to calculate the posterior probability of a reward function given the observed state-action pairs. Using Bayes rule one obtains:

\[ P(R|(s_j, a_j)) = \frac{1}{Z} P((s_j, a_j)|R)P(R) \]

(6.4)

As no closed for solution to this expression is readily available, (Ramachandran & Amir, 2007) describe a MCMC method that allows to sample from the posterior over reward functions by obtaining the corresponding value functions and policies for a given MDP associated with a particular reward function sample.
Figure 6.1: Comparison of the value functions for three component learners with different reward weights. The array of figures displays the value functions that were obtained through simulations in the walkway domain using reward weights of $r=0.5$ for the left column, $r=1.0$ for the central column, and $r=2.0$ for the right column. The three component value functions shown here are for the litter picking up behavior in the top row, the obstacle avoidance in the center row, and the walkway following in the bottom row.

6.4 Using prior knowledge of task structure

6.4.1 Structure of modular Q-functions

A first simplification comes from the fact that the modular Q-functions are scaled when their respective reward weight changes. This is a straightforward consequence of the fact that given
a policy \( \pi \), a transition function \( T \), a reward function which is the scalar multiple of some other reward function \( R = cR_0 \), and a discount factor \( \gamma \), the value function can be expressed as:

\[
V^\pi = R + \gamma T^\pi V^\pi \\
= (I - \gamma T^\pi)^{-1} R \\
= c(I - \gamma T^\pi)^{-1} R_0
\]

(6.5)

which is related to the matrix form of the Bellmann equations and where \( T^\pi \) and \( V^\pi \) are the transition function and reward function that are induced on the MDP by the policy \( \pi \). This has the consequence, that the value function for one particular component task has to be calculated only once and that the value functions that are the result of a multiplicative change in the reward can be calculated without reevaluating the MDP. Figure 6.3 shows through simulation results that this is indeed the case for the considered tasks.

### 6.4.2 Prior probability of reward functions

In the inverse reinforcement learning algorithm proposed by (Ramachandran & Amir, 2007) the authors considered MDPs for which reward functions were sampled from a given probability distribution. While this is a useful assumption when no specific domain knowledge is available, in the current setting the reward function was designed in such a way, that reward was only obtained in very specific states of the statespace of each component MDP. Reward is only obtained when picking up, i.e. intersecting with a litter object, negative reward was obtained when intersecting with an obstacle, negative reward was also obtained when leaving the walkway. The setup was chosen so that the reward was always 0 for task neutral actions. This allowed to include an additional, fourth learner that obtained a negative reward during each time step of the simulation. This last behavior corresponds to the implicit cost of walking.
6.5 Biological constraints

6.5.1 Metabolic cost in human walking

The literature on mechanical work and metabolic work in walking has determined a number of empirical relationships between walking speed and the metabolic cost in walking for humans. The data reported is mostly obtained from humans walking straight ahead on a treadmill. A recent study (Hall et al., 2004) has established how well the predictions of these models match observed energy consumption using indirect calorimetry both for walking and running on a treadmill as well as on a track. These results were used to obtain a value for the metabolic cost of walking straight ahead at the average walking speed of 1.6$m/s$ observed in the current experiments.

The metabolic cost for forward walking was found to agree best with the models of the ACSM (of Sports Medicine, 2000) and of Pandolf and colleagues (Pandolf et al., 1977) according to (Hall et al., 2004). Using those empirical models, the metabolic cost at an average speed of 0.6$ms^{-1}$ were calculated. Given that here the interest is in the metabolic cost of walking only, the resting component was subtracted from the total metabolic expenditure. The average of the two values obtained from the two models was used together with an estimated average mass of 70$kg$ for the subjects. The resulting metabolic cost expressed as power is 80$W$.

6.5.2 Variability in human walking

The problem of Inverse Reinforcement Learning as commonly formulated assumes that the transition function that describes the dynamics of the environment in which the agent is acting is known in advance. It is in principle possible to obtain this data specifically for the task at hand and it could have been obtained for the current navigational experiments, but previous research has produced data, that can be reused. For the navigational tasks, the experimental protocol would require subjects to walk to a particular target and obtain data on the variation of the endpoint depending on the distance traveled. This type of data could be obtained for
each of the tasks separately, corresponding to the proposed factorization. Recent experiments on this variability by (Kallie et al., 2007) in which subjects had to walk without visual feedback established very small variability for traveling distances covered within 300 ms, the timescale used as single time interval in the current simulations. It was therefore assumed, that the transitions of the reinforcement learning agent can be approximated as being deterministic.

6.6 Results

Figure 6.2: Comparison of the observed human trajectories along the walkway in the three task conditions. Top: ‘Pickup purple litter’. Middle: ‘Avoid blue obstacles’. Bottom: ‘Pickup and avoid’.

6.6.1 Human trajectories

The data of the subjects navigating the walkway environment and executing the task combinations of following the walkway while picking up targets, following the walkway while avoiding
obstacles, and the combination condition were considered. Graphical representations of the corresponding trajectories are depicted in figure 6.2 as top views for the individual task conditions. In all three figures, subjects started the itinerary at the left end of the walkway and walked down the 40m long distance. It is notable that the variability is quite modest between subjects especially for the first half of the trajectories and especially for the task of avoiding obstacles. The largest variability can be seen in the condition in which subjects were instructed to pickup litter objects. Further inspection of this conditions shows, that the main sources of these differences come from two sources. First, not all subjects picked up the same total number of litter objects, i.e. while some subjects tended to approach many more objects, others did not pick up as many. The second source is the fact that some subjects implicitly also avoided the blue obstacles, while other subjects walked right through them. The raw position data of the human subjects was transformed into the parameterization of the statespace used in the RL experiments described in the previous chapter, i.e. by obtaining angle and distance between the closest object of interest for each learner and the current walking direction of the subjects. Representation of typical such data for subject AF in the three conditions is given in figure 6.3.

6.6.2 Model trajectories

In order to obtain an intuition about the expressiveness of the used model a number of simulations were run with different relative reward weights and the resulting trajectories are shown in figure 6.4. The four top views were obtained by learning the respective value functions in separate simulations and then rescaling them according to equation 6.5. The reinforcement learning agent was then started at the same position on the walkway at which subject AF started the trial and the directions of motion were aligned. The four displayed trajectories demonstrate the broad behavior that the agent can display. The first simulation rewarded only avoiding obstacles and the agent learns a strategy of avoiding the first obstacle, then turning towards a corner of the space, and staying there. This reduces the chance of further collisions
Figure 6.3: Example data of a human subject executing the task 'pickup litter and avoid obstacles' using the parameterization of the state space for three component behaviors. Top: Representation of the distance and angle to the closest litter object during navigation along the walkway. Middle: Representation of the distance and angle to the closest obstacle. Bottom: Representation of the signed distance and angle between the subject's walking direction and the midline of the walkway.

with obstacles. In the second simulation only staying on the walkway was rewarded and the agent walks straight to the target at the end of the walkway. In the third simulation the agent was only rewarded for picking up litter objects, and indeed it approaches the targets and tries to pick up as many as possible for the duration of a single trial. Finally, the fourth simulation shows a trajectory resembles fairly closely the actual trajectory taken by the human subject as it executed the task of staying on the walkway, avoiding obstacles, and picking up litter. The reward weights for the agent were set to 0.1 units for the walkway following, 0.12 for the litter...
pickup, and 0.78 units for obstacle avoidance.

6.6.3 Estimated reward weights

The reward weights for each of the ten subjects in the two task conditions of picking up litter while following the walkway and in avoiding the obstacles while following the walkway were obtained using the inverse reinforcement learning algorithm proposed. This allows in effect to only estimate a single factor for the entire value function instead of requiring to estimate the full reward function with as many unknowns as states for each individual learner. The corresponding rewards are shown in figure 6.5. These results are encouraging in that they are in accordance with several aspects of the observed behavior of the navigating subjects. First, the positive reward weights for the litter picking up behavior are larger for the condition in which subjects were instructed to approach the litter objects (mean reward 0.65) and smaller for the condition in which they were instructed to avoid obstacles (mean reward 0.17). This is also consistent with the observation that the trajectories in these conditions differed in the number of litter objects that were targeted. These results are also consistent with the data describing the navigation behavior relative to the obstacles. While the reward weight was estimated to be smaller in magnitude (mean reward -0.28) for the litter picking up condition, it was significantly larger in the avoidance condition (mean reward -0.78).

6.7 Conclusion

If every-day visuomotor activities are learned and the algorithms that best describe this type of learning are reinforcement learning algorithms, it is necessary to extract the parameters of such models from the observed behavioral data of human subjects executing natural tasks. The hope associated with such a methodology is to be able to quantitatively characterize human behavior in a fundamental way. By knowing the respective value or cost of an action it is
possible to characterize and hopefully explain, which tradeoffs are responsible for orchestrating the sequences of behaviors that humans execute in daily goal directed behavior.

The present chapter constitutes a first attempt at extracting the value functions underlying human behavior in navigating along a walkway while avoiding obstacles and approaching targets. The proposed methodology uses a simplified version of the Bayesian inverse reinforcement learning algorithm by (Ramachandran & Amir, 2007) by using additional constraints. These constraints are derived from the specific knowledge of the problem in question and allows reducing the reestimation problem from estimating as many reward values as individual states to the estimation of a single factor for each of the component learners. Further testing of this methodology needs to establish how well this method works in terms of accuracy and precision of the obtained reward estimates. The first results of the relative reward weights associated with the learning modules responsible for obstacle avoidance and target approach for the two task conditions considered are encouraging, because they correspond well to the properties of the trajectories taken by the human subjects.
Figure 6.4: Comparison of the actual human trajectory (blue) for subject AF executing the combination task of following the walkway, avoiding obstacles, and picking up litter with four model trajectories obtained with different reward weights for the component tasks (black). The reward values are reported for the walkway following, the litter pickup, and the obstacle avoidance. From top to bottom: (0, 1, 0), (0, 0, 1), (1, 0, 0), (0.10, 0.12, 0.78)
Figure 6.5: Estimated normalized reward weights for the litter picking up behavior (purple) and the obstacle avoidance behavior (blue) for the 10 subjects in the task conditions ‘pickup litter’ and ‘avoid obstacles’.
7 Conclusion

This thesis investigates human visuomotor behavior in natural tasks in order to better understand how exactly tasks exert their control on human visuomotor activity. The tasks under analysis are those of avoiding, approaching, and following in walking, some of the most commonly executed and early learned action sequences in humans. Behavioral, computational, and theoretical methods are employed in order to address this question on different levels of description. The core hypotheses of this thesis are that human visuomotor activity is goal directed and has to be learned.

7.1 Contributions of the thesis

7.1.1 Task dependence of human visuomotor behavior

Chapter two provides empirical evidence from psychophysical experiments further demonstrating that human gaze during natural behavior is highly task dependent. The data demonstrated significant differences in behavior between tasks and high similarity across subjects within task conditions. Previous research has demonstrated that human fixations are directed to task relevant locations when these are needed, and that the statistics of eye movements change with the ongoing task.
The current results are not only a further demonstration of the importance of the task in visuomotor behavior, but they are also a challenge to current normative models of gaze allocation. The currently predominant hypothesis is that low level feature properties such contrast in feature dimensions such as luminosity, orientation, and color somehow capture attention by virtue of their visual properties alone. The presented data can find no support for such a claim. Indeed, saliency models are bottom-up models that do not incorporate any notion of the goal of a fixation within a action sequence and accordingly fail to model the observed data. The model by (Sprague et al., 2006) incorporates explicitly the goal of behavior in human vision and gaze selection but makes the prediction that there is no difference in time spent on relevant parts of the scene depending on the concrete computation that has to be performed but that instead the differences only arise from different reward weights and different uncertainties associated with the component tasks. The presented data clearly speaks for additional differences that are inherent in the specific computations such as avoiding an obstacle or approaching a target.

7.1.2 Task sensitive feature statistics

Chapter three further quantifies some statistical properties of features at gaze during the three walking tasks. The motivation for this analysis is at least twofold. First, current saliency models conceptualize vision as a passive process and postulate that gaze is attracted to regions in the image dependent solely on some image properties. In the past it has been reported repeatedly, that contrast is elevated at fixation location compared to non-fixated image regions. Here, similar statistics were obtained for the navigational tasks and results are reported that contradict previous conclusions based on free-view and search tasks, in that in one particular task condition of approaching targets, the contrast at fixation location was reduced compare to contrast at randomly selected parts of the scene. This demonstrates that it is not the case that particular visual features intrinsically always attract gaze, but that the statistics of features are task dependent.
An important consequence of this analysis is that it clearly shows that models of representational learning have to take into account, how the visual system actively engages in visuomotor behavior. The input to the visual system is clearly dependent on the ongoing task and the statistics of this input change significantly on timescales of minutes. This suggests that models such as those that learn basis functions that are adapted to representing the visual input in some optimal way or adaptation processes that adjust the output statistics of representational units should take into account how the statistics of the visual input actually change.

### 7.1.3 Specificity of features at gaze

Quantitative analysis of the statistics of features at the point of gaze reveal the strong influence of the task on the input to the visual system, but these data are obtained in a virtual environment with scenes that are rendered by graphics packages. It is therefore important to validate the methodology of using such visual input by demonstrating that well known properties of such images are comparable to those of natural images. In order to exclude biases in the analysis due to the fact that the visual environment was rendered artificially, summary statistics such as the powerspectrum and edge filter response distributions for an image ensembles obtained in the virtual environment are presented and demonstrate that they match well those obtained from images in natural environments.

Further analysis was carried out by applying dimensionality reduction techniques to these filter responses. These methods show that the visual stimuli encountered in the virtual environment can be well approximated with a low dimensional representation. This methodology can be thought of extracting features of intermediate complexity for the description of the visual scenes. The statistics of the selection of these features dependent on the ongoing task shows that only a small fraction of all the patterns present in the environment are targeted during navigation. Again, this demonstrates that these features are significantly different for the task conditions. These results therefore further suggest that vision operates highly selectively depending on the task and that features of intermediate complexity are not only relevant in tasks
such as object recognition, but also in visuomotor tasks.

7.1.4 Learning concurrent tasks

The behavioral results further demonstrate, that task based models of visuomotor behavior need to be constructed. One considerable issue is how to map such task based behavior onto known learning and action selection substrate in the human brain. The success of reinforcement learning as a general framework for the formulation of learning algorithms that can reproduce human and animal learning as well as neuronal signals during such learning is here extended to learning of visuomotor tasks. Chapter five presents a credit assignment algorithm for the learning of multiple tasks, when these are executed concurrently, and only a single global reward signal is available, which reflects the sum of all individual reward contributions. This algorithm exploits the fact that different task combinations may be executed at different times and considers the special case, in which the individual statespaces of the component tasks are independent of each other, but the actionspace may be shared.

The credit assignment problem is an important problem for multi-agent systems, but even more so for a biological system as a behaving animal in a complex environment has to solve multiple concurrent problems. The work is related to other approaches that start out with compositional solutions to individual problems and then devise methods in order to combine a large number of such elemental solutions, but it is able to do so from a single global reward signal. This algorithm is applied to a formulation of the walkway navigation task and it is demonstrated that it is indeed feasible to learn to coordinate the visual behaviors in this setting. Therefore, this gives further evidence that visuomotor behavior can be learned by providing an algorithm that can solve such tasks.
7.1.5 Consumable rewards

The above credit assignment algorithm that addresses the question of how individual goals can be learned from a single global reward signal is furthermore implemented using a method that does not require intricate communication schemes as has been previously proposed. The construction of consumable rewards in which individual component learners consume the global reward at a rate that is proportional to the current estimate provides a solution that minimizes the infrastructure of communication between individual modules but allows to take advantage of the accumulated knowledge from previous task executions.

A further extension that takes the relative uncertainties in the current reward estimates of the individual learners into account is shown to converge quickly, as it reduces the number of samples that have to be obtained from interacting with the environment. This Bayesian reward weighting, which is similar to a kind of cue-weighting, is particularly advantageous, if the different component tasks do not have all the same probability of being executed. While tasks that have been learned well have higher confidence in their reward estimates, other tasks that are less well learned can take advantage of the estimates from the well practiced learners.

7.1.6 Inverse reinforcement learning using constraints

In order to validate the proposed reinforcement learning based models, it is important to re-connect the parameters of the models with the experimental data. The data describing the human behavior in the navigation tasks is modeled using the multiple-module based algorithm and the individual trajectories of human subjects in the different task conditions are fit well by the model. The correct parameters are extracted using a recent inverse reinforcement learning method, which has been augmented by additional knowledge about the task and the functions describing it. It is possible to extract those parameters from the human behavior that can then be used to reproduce the observed trajectories well only by utilizing these additional constraints.

This methodology opens up the possibility of extracting the rewards implicitly associated
with human actions as expressed in natural behavior. This is quite intricate in the case of natural tasks, because participants are not restricted to a small number of actions but instead execute long sequences of actions which require high dimensional statespaces describing the state of the environment. Using the proposed modular decomposition in the reinforcement algorithm, estimates of these rewards for the different task combinations are obtained and translated into a common currency as reinforcer, which is based on empirical data on the metabolic energy consumption in walking. These results show that human trajectories while navigating through the environment are well described by the value functions found by the reinforcement learning algorithm. Furthermore, the extracted parameters of the model quantify the differences in behavior between tasks and the similarities across individual subjects. This can be seen as further evidence that the spatiotemporal organization of visuomotor behavior is optimized to maximize obtained reinforcers, even in natural every-day behavior. The equivalence in the mathematical formulation with reward based learning suggest further investigations about which secondary reward mechanisms can mediate the reinforcing signals.
References


Baddeley, R. J., & Tatler, B. W. 2006. High frequency edges (but not contrast) predict where we fixate: a Bayesian system identification analysis. Vision Research, 46, 2824–2833.


Sprague, Nathan, & Ballard, Dana. 2003 (August). Multiple-Goal Reinforcement Learning with Modular Sarsa(0). In: *International Joint Conference on Artificial Intelligence*.


