The Dissertation Committee for Karl S. Muller
certifies that this is the approved version of the following dissertation:

Modelling visually guided natural locomotion

Committee:

Mary M. Hayhoe, Supervisor

Lawrence K. Cormack

Alexander C. Huk

Alexander Huth

Wilson S. Geisler

Modelling visually guided natural locomotion

by

Karl S. Muller

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2021

Dedicated to all my friends and family, past present and future

Acknowledgments

I would like to first thank my advisor Mary Hayhoe. Your continuous guidance regarding scientific, career, and personal matters was incredibly valuable and is highly appreciated. I think it is easy to take for granted the immense freedom and flexibility I was afforded during my thesis work, and so I also wanted to say thank you for letting me basically not live in the city where I was going to school. This freedom also extended to my scientific pursuits, where I felt free to explore what I deemed interesting, which was immensely rewarding.

To my committee members, thank you for the helpful and interesting feedback and suggestions at different stages of my research projects. I would also like to draw attention to classes that I have taken from each of you at different stages of my time here, all of which were foundational in shaping how I approach scientific problems, and even how I view the world.

Thank you Jonathan Matthis, for providing mentorship when I was an undergraduate and graduate student. Working with you on this dataset which you pioneered was what drove my initial interest in this kind of work, and I would not be here without that experience. I would also like to thank Dana Ballard for feedback throughout different projects, and for keeping everyone on our toes at lab meetings.

Thank you to all my past and present lab mates and colleagues, Dan Panfili, Kate Bonnen, Jake Whritner, Ruohan Zhang, Lijia Liu, Sariel Li, Suna Guo, Dawei

Liang, Matt Tong, and Oran Zohar. Also thank you to all my other colleagues in INS and CPS. My experiences and interactions with each of you have taught me so much, and made my time here a great experience.

Thank you to Larry Abraham, my first mentor as an undergraduate. Working with you was my first peek into academic research and science, and the great experience undoubtedly pushed me towards further pursuit. Also thank you to Michelle Dunn for your kindness and generosity in helping me get involved with Larry's research to begin with.

Finally, thank you to my family and friends for your support throughout my time here. It is definitely a privilege to be supported and encouraged while I find what I want to do in life. I could not have done it without you.

Thank you all so much.

Karl S. Muller

The University of Texas at Austin
December 2021

Modelling visually guided natural locomotion

Publication No. _____

Karl S. Muller, Ph.D.
The University of Texas at Austin, 2021

Supervisor: Mary M. Hayhoe

Vision is an active process where an organism must seek out and acquire the information necessary to support different behavioral goals. This makes understanding these behaviors important for understanding how visual processes unfold in the brain, which is adapted to perform the necessary computations for these behaviors. Bipedal locomotion is one such behavior, which is of particular importance in evolutionary history. In this work I examine locomotion over complex terrain using a mobile eye tracker and motion capture system. This allows for an integrated record of eye and body movements, as well as approximation of the retinal input image. Computer vision methods were applied in order to extract visual motion and to reconstruct environment geometry. This allowed an unprecedented opportunity to examine the visuo-motor decision processes controlling locomotion

in natural terrain. Our results reveal the statistical regularities in motion signals that depend on gaze angle and terrain, and have implications for how the visual system might process this information. Gaze angle shapes the spatial distribution of both speeds and directions of visual motion, which has implications for how the visual system might account for this relationship. Terrain differences also manifest in motion signals as deviations from flat ground motion, the magnitude of which is correlated with proximity of gaze allocation to the walker. We also find that gaze is partly predictable on the basis of body orientation and image features. Finally we find that foot placement reflects the avoidance of height changes, with the degree of this influence being modulated by subject leg length. Walkers appear to factor this information into their decision making across multiple spatial scales. Thus foot placement reflects a complex interplay between energetic costs and the need for stable footholds, all taking place as walkers maintain their forward momentum. The conclusions drawn from this new dataset, as well as the novelty of the dataset itself are important contributions towards a deeper understanding of how vision is used to guide locomotion in the natural world.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Vision is an active process that involves acquiring task relevant visual information as different behaviors unfold. This means that in order to understand visual processing in the brain, we need to examine the problems that visual systems need to solve in natural contexts. This is because natural contexts represent both the conditions in which the visual systems underwent evolutionary selective pressures, and the conditions that they are adapted to over their lifetimes. By better understanding these natural behaviors we can build models of visual processing that incorporate information and constraints that shaped the visual processing systems themselves.

One behavior of particular importance to human evolutionary history is locomotion. Locomotion is a complex sensorimotor decision-making process that involves incorporating incoming visual information into an ongoing motor plan, especially when the terrain being navigated is uneven. Humans learn to seamlessly perform this behavior, suggesting that the brain has converged on efficient solutions to this complicated problem. While much is already known about the mechanisms of locomotion, much less is known about the visual control mechanisms.

In this work I explored complex terrain navigation, focusing on components of the visual input and behavioral output. For visual input, motion and depth were considered, and for behavioral output, I looked at foothold location selection. All of the analyses in this thesis centered around data acquired using an apparatus that combines motion capture and eye tracking data of humans as they navigate complex rocky terrain. The resulting data is synchronized streams of body positions and gaze directions over the course of multiple traversals of rocky terrain.

In natural behavior, the actions of the observer create a complex time-varying input on the retina. To understand the stimuli the visual system must deal with, it is necessary to understand this input. Therefore this work began with approximating and measuring the properties of retinal motion, something that has been hard to measure in the past. Chapter 2 explores the characteristics of visual motion inputs during natural locomotion, and how these are affected by the ongoing behavior. This involved developing a novel method that combines optical flow estimation with eye tracking in order to combine the effects of eye rotation and translation through the environment in order to approximate retinal motion. The next question concerned the extent to which gaze allocation can be predicted on the basis of image features and body configuration (Chapter 3). In this work we used convolutional neural networks and linear regression to predict gaze direction from scene images and body position. This project revealed that it would be difficult to understand gaze allocation without understanding foothold placement, since this is the primary task and likely a major driver of gaze allocation. In Chapter 4, I discuss a novel method for combining motion capture, eye tracking, and 3D environmental structure data obtained via photogrammetry. This addition to the data was important for computing more accurate foothold and gaze locations, and also allowed analysis of environmental structure in relation to foothold placement. In Chapter 5, I discuss results from analysis of this new foothold location data set. Here the focus was on

the relationship between environmental structure and foothold selection. We developed a method for measuring behaviorally relevant properties of the environment, and applied CNNs and a boosting classifier method to measure the relationship between environment structure and foothold placement. Chapter 6 provides a more general discussion of the findings and the contributions of this work. The next section provides background on eye tracking research, investigations into visual motion processing systems, and locomotion research, and discusses some unanswered questions that this thesis makes some progress towards addressing.

1.2 Background

1.2.1 Mobile eyetracking

Background

The human visual system shares a feature with many other primate visual systems, namely, foveation. Foveation means that only the central few degrees of visual angle have anatomical structure that supports high resolution. Because of this feature, animals must move their eyes and direct the fovea towards whatever point of interest in the external world is relevant to ongoing behavior. This means that eye movements are intricately linked to behavior and are therefore informative about the underlying neural processes. Eye tracking has been used since the early 1900s, although work by Yarbus provided a major advance. In the 1967 book Eye Movements and Vision [1], Yarbus demonstrates how tasks heavily influence the pattern of eye movements, since different information about a scene or the environment (in this example, a painting where subjects must deduce different things about it) is relevant depending on a particular task.

While this study was done on subjects with restrained heads who were looking at a painting, a particularly important development was eye trackers mounted

on the head. These "mobile" eye trackers confer the benefit of allowing subjects much greater freedom of movement, and therefore the ability to engage in natural visually guided behaviors. While this does come at the expense of the control and precision one can expect from a head fixed subject, the advantages are important when it comes to natural behaviors. Subjects can simply perform tasks as they would normally, with the added oddity of an eye tracker attached to their heads dissipating rather quickly (particularly for newer, lightweight, and very noninvasive trackers).

As of 2021, there are many commercially available mobile eye trackers, and they are improving rapidly. For work done in this thesis, we used the "Pupil Labs Core", recently renamed in light of newer models. Trackers like the Tobii Pro 2, and SMI Eye Tracking Glasses provide similar data. Most, if not all of these glasses work using a similar processing pipeline. They use infrared illumination of the eye, combined with infrared cameras in order to capture images of the eye. The black and white infrared image is then processed using various image processing techniques in order to extract the location of the pupil center. Meanwhile, an outward facing camera simultaneously records the subject's first person perspective. A calibration procedure where the subject directs their gaze and fixates on locations that are identifiable in this outward facing camera recording is done. Periods of the recording where subjects are looking at particular locations are noted, and a mapping between particular pupil positions in the infrared camera images, and positions in the outward facing camera images is estimated. The particular method of doing this varies however once this mapping is established through the calibration procedure, it can be applied to the remainder of the recording, yielding gaze positions as subjects carry out whatever behavior of interest.

Distinction from head fixed

The ability to move the eye and head simultaneously is important, since si-

multaneous eye and head movements are often the preferred method for making even low amplitude eye movements. This style of combined movement is not possible during head fixed paradigms, and the extent to which the inability to do this produces differences in behavior is not quite clear [2]. However, using mobile eye trackers one can record eye movements while also allowing subjects to use natural coordination patterns.

Besides simultaneous head and eye movements, mobile eye trackers also enable the study of dynamic behaviors and actions that involve the entire body. For example moving across the room to pick something up, or moving to another room to look for an object. These everyday behaviors involve a combination of eye movements and body movements working together. Understanding this process as it unfolds naturally, where the subject is able to freely move their whole body, requires allowing this unrestrained natural version of the behavior to unfold. In natural contexts eye movements are always related to the ongoing behavior, and mobile eye trackers enable the study of behaviors as they happen, whereas head fixed paradigms, while conferring the benefit of tight control over the retinal input, will always be missing this crucial component. The ability of the subject to move around has allowed a better understanding of the tight linkage, in time, between gaze and behavior [3], [4], [5]. This paradigm allows identification of the particular information that is useful for a particular action, the time at which the information is needed, and allows exploration of the costs and benefits of action choices.

The study of eye movements in the context of sensorimotor decision making

Considering sensorimotor control as a series of decisions being made based on visual information is a useful framework for understanding natural behavior [6], [7], [8] (Also see [9] for review). Mobile eye tracking enables the study of natural behaviors as they unfold. This is important because it allows examination of the

interaction between visual perception and decision making (which often manifests as specific movements). Eye tracking can establish what the visual input is from moment to moment, and can thus be used to better understand how this information is used to make decisions. It can also help in identifying and understanding behavioral goals during behaviors as they happen. For example, we know that while steering during driving we direct our gaze towards tangent points on the inside of curves before turning [10], or that cricket players make anticipatory saccades to where they predict the ball will bounce in order to hit it [11]. Measuring eye movements can also help us determine explicit versus intrinsic rewards during behaviors. It was been shown that saccades are made to targets with explicit and implicit rewards associated with them [12], [13], [9]. Measuring gaze allocation can be helpful in understanding what sources of information are important and are associated with reward (in many cases, successful completion of a task) for different complex behaviors.

Mobile eye tracking in this thesis

In this dissertation, a mobile eye tracker (Pupil Labs Core) [14] was used to understand gaze allocation and eye movements during a natural locomotion task. Human subjects wore the eye tracker and were instructed to walk back and forth across complex terrain. Gaze allocation is important in this case because it gives some insight into the timing of when particular visual information is relevant to the brain at particular moments as subjects traverse the terrain. Locomotion over uneven surfaces is a complex process involving different brain regions, and knowing the visual input at precise moments during this process is an important step towards understanding the neural basis of this process. Using the eye tracker's scene camera and gaze position estimate, we can approximate the retinal image, which is the input to the visual system. Having a quantification of this input is important for understanding and modeling whatever processed occur between this input and the

resulting motor output. In the case of visual motion signals, the eye tracking data is important in another way. Since the movement of the eye affects the resulting motion pattern from the retina's reference frame, we need both the measurements of eye position as well as eye movements at given moments in time in order to approximate the retinal motion input at that time.

1.2.2 Locomotion and optic flow

Background

Locomotion involves what can be thought of as the successful execution of several behavioral modules. In the context of flat ground, selecting appropriate places for footholds is not an issue, however it is still necessary to control steering. Steering requires cortical modulation of the central pattern generation signal that allows top down control of the direction of locomotion, in order to transport an organism to a particular location. Because of this relationship to particular locations that an organism is trying to reach, steering requires some kind of visual signal to modulate direction. Thus vision is typically a necessary component even in the case of steady state flat ground locomotion.

Patla (1998) and Warren (1998) provide reviews on use of visual information during locomotion [15], [16]. There are multiple sources of visual information one can take advantage of. One emphasized by Warren and that has become a focal point in visually guided locomotion research is optic flow. Optic flow is the pattern of visual motion that results from an observing moving through a static environment. It has been typically framed as a guiding signal for steering during locomotion, since the focus of expansion or FOE (the critical point of the optic flow signal, where motion is emanating from) coincides with the direction that the observer is currently moving. Because of this, it is commonly thought that the focus of expansion is used for steering towards a goal [17], [18], [19], [20], [21], [22]. However,

this is the case only for motion in a constant direction. Using retinal motion signals calculated from eye tracking during locomotion, it has recently been demonstrated that instability of the focus of expansion as momentary heading varies through the gait cycle makes use of the focus of expansion an unlikely steering strategy [23]. Instead, it is most likely that the optic flow patterns are used to monitor movement of the body with respect to the local surroundings and thereby monitor posture and foot placement. Thus the measurement of flow patterns in a behavioral context has allowed a different interpretation of how it is used. The context-dependent measurement of flow patterns may also have implications for understanding the role of visual motion sensitive regions in the primate brain like MST, that have been implicated in the control of locomotion. Regions like MST exhibit sensitivity to the global motion patterns that comprise optic flow. In the next section I will provide a review of some studies into optic flow and locomotion, as well as the interpretation of the role of MST in the processing of optic flow. However first I will briefly describe optic flow and some of its features.

There are a few important features of optic flow that are worth highlighting. This optic flow signal contains information both about the observer's environment, as well as the observers motion through the environment. When an observer moves relative to a fixed environment, the way that the environment moves relative to the observer is entirely determined by how the observer is moving and what the structure of the environment is. Thus the structure of the environment is implicit in the optic flow signal, and can be extracted through various means. When an observer moves through an environment, the exact direction of the observer's movement in the observer's visual field will contain a point (the FOE) from which all motion appears to emanate. This geometric features means that the FOE corresponds exactly to the direction of motion in the case of an observer that is not rotating and whose gaze corresponds to the direction of heading. The weight of this purely

expansive component simply scales with the speed at which the observer is moving in a particular direction. However, even in the case of a rotating observer, for example eye movements, the fundamental structure of an FOE with other points in the scene appearing to emanate outwards from the FOE is retained, although an additional pure translation component is added to the visual motion. Geometrically this results in a shifted signal, where the FOE is shifted along with all other structures in the motion vector field. There is an additional third type of transformation feature in optic flow, where the eye is rotating about its viewing axis, This results in rotation in the optic flow signal about the FOE which introduces rotation into the whole image about the point of gaze. Thus optic flow can be decomposed into a expansive, rotational, and translation component, each with their associated behavioral drivers.

Optic flow for the control of locomotion and the role of MST

There is a large body of literature examining the role of optic flow during locomotion. For example Pailhous et al (1990) found modulation of stride length and stride frequency induced by optic flow that was not concordant with the actual walking pattern [24]. Additionally, effects on walking velocity have been found by Konezak (1994) [25] and Zijlstra (1995) [26]. Modulation of walking velocity is achievable through systematic manipulation of optic flow signals. This effect is well documented and reproducible in different presentation paradigms like VR [27], [28], and across different populations like young and old [29], and even in recovering stroke patients [30]. Prokop et al (1997) demonstrated that even when instructed to maintain a particular walking velocity, subjects can be easily manipulated into modulation of their gait via presentation of an optic flow signal consistent with different walking speeds [31]. This effect and others such as phase-locking gait through control of expansion and contraction flow patterns [32], or modulation of ankle activation [33], demonstrate an intricate relationship between optic flow and locomotion. This relationship between optic flow stimuli and locomotion suggests

some kind of reliance on this signal for the control of locomotion. In the primate visual cortex, many regions have shown varying degrees of motion sensitivity. However, MST has shown a particularly strong selectivity for global motion patterns consistent with optic flow. Based on the relationship between optic flow and locomotion, and the sensitivity of MST to optic flow stimuli, it has been implicated in extracting information relevant to locomotor control [17], [18], [19], [20], [21], [22].

Studies of MST and its response properties are thus often framed in the context of control of locomotion by estimating heading direction. Since the current movement direction is crucial for planning and regulating ongoing motion, heading direction is a variable of interest. In Duffy and Wurtz (1995), monkeys were presented with motion stimuli with radial and rotational components, meant to mimic optic flow [34]. It was found that in most MST neurons, the response differed when the center of these stimuli was varied. Others have found similar results, such as Lappe et al (1996) and Gu et al (2006) [17], [35] . These results suggest that MST can represent the current direction of motion, which means it can support computations that require such an estimate.

I am not aware of studies that explicitly link MST activity to the phenomena of gait modulation described earlier, given the difficulty of neurophysiological recordings in mobile primates. However if MST neurons are truly responsible for heading estimation, which is important for controlling locomotion and steering, then such studies are warranted. The present work did not involve neural recording. However some insights can be provided by measurement of the kinds of visual stimuli that provide useful information for the visual system, as well as quantification of the kinds of behavioral outputs that need to be generated (control of foot placement or steering towards particular locations while navigating complex terrain). A better quantitative understanding of how these different systems work together would be helpful for extending the applications of this gait modulation phenomena. Although

such applications are not directly related to this thesis, there have already been some advances in therapeutic applications that are worth a brief discussion.

Applications of the link between optic flow and locomotion in stroke rehabilitation

Earlier the phenomena of modulation of gait due to varying optic flow signals in stroke patients was mentioned. This represents an interesting application of this phenomena to facilitate stroke patient recovery. The idea is to promote recovery of normal gait by modulating the stroke patients walking behavior by systematic manipulation of optic flow signals. The aforementioned study Lamontagne et al (2007) found that despite a decrease in the degree of modulation of walking based on varying optic flow signals, stroke patients still exhibit increased walking speeds when presented with slow optic flow signals. Subjects were instructed to maintain a constant walking speed, and when presented with an optic flow signal corresponding to slower walking speeds, sped up their gait in order to compensate for the perceived slow down. This kind of method could be used in gait rehabilitation to promote post stroke recovery. A study by Kang et al (2012) found significant improvements in stroke patients who used optic flow speed modulation based rehab compared to regular treadmill training and therapy [36].

Despite a lack of a quantitative model of how the visual system parses and interprets optic flow signals, and how this processed visual information is eventually used by the motor system to modulate gait, this effect of optic flow modulation on gait has been used to treat stroke patients with impaired gait. A more detailed understanding of how optic flow is used to control and modulate locomotion can only help in developing even more targeted ways to exploit and employ this phenomena for treatment purposes.

Optic flow and locomotion in this thesis

In this thesis, optic flow signals experienced by subjects during a natural

locomotion task are measured. One of the aims is to provide a quantitative measure of the input to the visual system while subjects do this task, as well as a measure of result of motor output (foothold locations chosen, walking trajectory chosen).

A fundamental limitation of this research is that we are unable to manipulate the visual input, since we are simply approximating it as it happens using the mobile eye tracker. However this also provides an important advantage, which is the measurement of this input as is during natural behavior, so it is not necessary to make assumptions about the nature of a visual input. Optic flow has been simulated in a variety of ways each with its own set of assumptions about what the signal looks like naturally. However, it has been demonstrated by Matthis et al, (2021) that the optic flow patterns on the retina vary through the gait cycle in a way that depends on gaze location relative to the direction of travel [23]. This can be seen in Figure 1.1, which is reproduced from Matthis et al, (2021).

Here retinal motion during locomotion was approximated with numerical simulations based on real walking and gaze data. The sign of the curl values in the retinal optic flow input indicate the direction of fixation relative to the direction of travel. The point of maximum divergence in the retinal optic flow signal projected onto the ground plane corresponds to the direction of travel projected onto the ground plane. This Figure illustrates the extensive modulation of the flow patterns during a single step while gaze is held stable in the world. This modulation of the flow pattern corresponds to substantial variation of instantaneous heading, and raises the question of how the flow patterns are used perceptually. The gait induced variation in heading makes it unlikely to be useful for steering towards a goal. A more plausible scenario is that walkers learn the flow patterns consistent with stability and use flow to modulate posture. The demonstrated effects of flow manipulations on gait, discussed above, are consistent with this. The importance of this study for the present work is that it demonstrates the importance of directly measuring

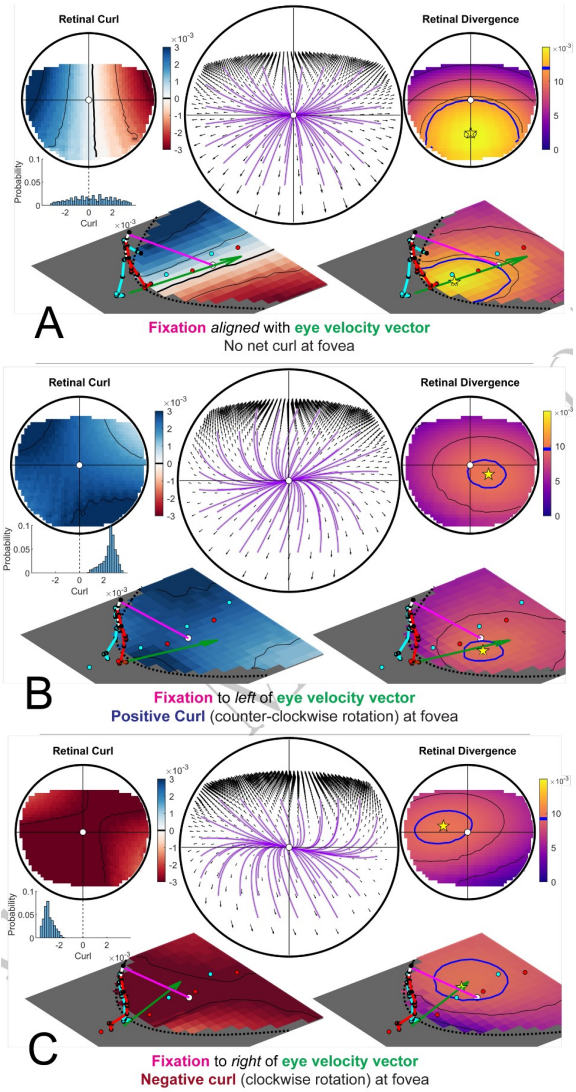


Figure 1.1: Retinal curl and divergence during natural locomotion. Figure from [23]. Each panel shows simulated retinal flow patterns along with the local curl and divergence values. (A) corresponds to fixation in the same direction as the direction of travel (eye velocity vector). (B) and (C) correspond to fixation to the left and right of the eye velocity vector respectively, resulting in positive or negative curl values. The point of maximum divergence projected onto the ground plane coincides with the eye velocity vector projected onto the ground plane in each case.

the motion patterns on the retina during natural locomotion, and shows how such measurements are critical for interpreting the perceptual role of motion information.

This has been the only attempt to measure the actual pattern of flow during locomotion. The present work adds to the Matthis et al study by computing the statistics of the motion input to the visual system during natural locomotion.

1.2.3 Locomotion on treadmills

Background

Bipedal locomotion provides clear evolutionary advantages like energy conservation, and a better line of sight. One disadvantage when compared to quadrupedal locomotion is that it is inherently unstable and somewhat ballistic at higher speeds. While locomotor tasks like moving towards a particular location involve interaction between many different brain regions, locomotor signals that generate the patterns of activation that produce a preferred gait originate in the brainstem [37]. In fact basic locomotor function remains intact in decerebrate cats [38]. Many studies have shown that decorticate and decerebrate cats, with varying levels of transection, can still exhibit different gait patterns. This illustrates the automaticity of such a behavior, requiring no executive cerebral control. This level of automaticity lends itself towards experiments that measure repeatable and reliable gait characteristics. Many studies that seek to understand human locomotion are conducted on treadmills or closely monitored stretches of flat ground. This enables precise measurement of the kinematics of movement, and even allows for electroencephalography (EEG) and electromyography (EMG) recording during the behavior.

Certain lawful behavior during steady state locomotion has been found. For example Borghese et al 1996 found invariant kinematic patterns across a range of locomotion speeds, reflected by similar gait trajectories in principal component space [39]. Despite differences in body composition between humans, certain similarities

emerge as a result of some kind of optimization principle. The idea of optimization for motor control is powerful, since it provides a framework for explaining the emergence of particular strategies for dealing with motor problems given a particular set of tools. In the case of locomotion the bipedalism results in similarities. One important well conserved pattern is the gait cycle (see Kharb et al 2011 for review) [40]. The gait cycle refers to a cyclical pattern that a walker undergoes during locomotion. A person might start standing with two feet firmly planted on the ground, then commence walking. This first step is abnormal relative to the gait cycle, but once it commences the walker cycles through a heel strike with the foot that made the first step. The foot is then flat and firmly on the ground, when the other foot leaves the ground in the toe off phase. There is then a single support phase where only the first planted foot supports the body, while the recently toed off foot swings, and is then planted in front, where another heel strike occurs. The walker thus alternates between double support and single support as one or two feet are in contact with the ground.

A lot is known about how locomotion functions from biomechanical modeling of locomotion. One of the simpler but highly effective models is the linear inverted pendulum or LIP [41]. This model uses the observation that dynamics of bipeds during the single support phase can be well approximated by an inverted pendulum. This conceptualization means that the walker takes advantage of basic dynamics of their body, meaning they do not have to add a lot of kinetic energy into the system, they instead 'swing' from foot plant to foot plant while conserving momentum. Different variations on the LIP have different advantages, but the simplicity of the LIP makes it well suited for modelling locomotion particularly for robotics applications when low dimensionality is important for real time computation.

The neural mechanics of locomotion are more difficult to observe and understand since one cannot rely purely on measured kinematics. However different

technologies have made it possible to record from different parts of the nervous system during locomotion in humans and other animal models, enabling us to understand some of the neural mechanisms. For example, electromyograms (EMG) can be placed on the surface of the skin above muscles in order to measure the electrical activity in the muscles. This measures the electrical activity in muscles that arises as a consequence of nerve stimulation. This allows quantification of muscle activity, which is important for modeling. For example Arsenault et al (1986) used EMG to compare muscle activation in the soleus, rectus femoris, biceps femoris, vastus medialis, and tibialis anterior during either treadmill vs walkway locomotion [42]. They concluded based on correlations between the corresponding muscle activations on treadmills vs walkways that treadmills despite minor differences like larger EMG amplitudes with smaller variation, treadmill and walkway locomotion is quite similar. This is an important point to make since a large portion of locomotion research uses treadmills due to the experimental control it provides.

In this thesis, the data comes from a non treadmill locomotion task, however it is important to consider studies that use treadmills since they are similar enough that insights derived from treadmill studies should be applicable to locomotion in general. However it is also worth focusing on treadmill studies that more closely approximate our task. Specifically, data in this thesis comes from walking that requires visual feedback and dealing with complex terrain. I will briefly discuss treadmill studies that deviate from normal flat ground walking, since these demonstrate some of the adaptive capacities of our locomotor systems.

Abnormal treadmill studies

Spit-belt treadmills are widely used in locomotion research for a variety of applications. A key feature of the split-belt treadmill is the ability to both present a walking task that deviates from a typical task that a subject is used to, and the ability to systematically vary the amount of deviation by varying the speeds of each

belt. The split-belt treadmill works by incorporating two separate treadmill belts into one, with each belts speed being independently controlled. For many studies this is used to simulate more varied environmental demands for example complex terrain. This simulated complexity can then be used to see whether there are differences in the capacity to deal with in stroke patients for example. In Reisman et al (2007), individuals who had suffered from a stroke were compared to controls were compared in their spit-belt walking adaptation. It was shown that the stroke patients did not exhibit impaired adaptive abilities [43]. Another study by Riesman et al (2013) showed that using a split-belt treadmill training process could be a viable treatment option post stroke, to improve step length asymmetry [44]. Step length asymmetry refers to the ability to produce alternating steps that vary in step length, which is a deviation from preferred gait and required for navigating more complex terrain.

In addition to split-belt treadmills, treadmills with uneven terrain have been used to examine the effects on locomotion. For example Voloshina et al (2013) used a treadmill with elevated foam patches of varying heights tiled on the belt [45]. They showed a net metabolic energy expenditure increase due to the terrain complexity. They suggest that the increased expenditure could be due to increased mechanical work at the knee and hip joints to accommodate the uneven terrain. Kent et al (2019) used a similar treadmill to look at the walking dynamics [46]. They found an increase in angular momentum range and variability that normally is close to zero during level walking, and has been linked to falls when present. These studies quantify differences that are introduced by terrain complexity, which highlights the importance of studying complex terrain navigation as it unfolds as is done in this thesis. Doing so in more unconstrained contexts as in this thesis will deepen our understanding of how these adaptations manifest during natural behavior.

Applications of findings

Understanding how locomotion unfolds, particularly when it comes to more

naturalistic settings like complex terrain (despite human efforts to maximize the amount of flat ground in our environment), is essential for applying what we know about locomotion practically. Whether in the domain of the design of legged bipedal robots, for treatment of motor problems resulting in locomotor deficit through therapy, or even the design of exo-skeletons or prosthetics to to address those deficits or even augment normal function, it is critical to establish quantitative models of how these systems function. Through these models one can better design and implement such applications. However a missing component is a deeper understanding of the underlying neural basis of these systems. This level of understanding is crucial in developing quantitative models.

Bipedal locomotion is a desirable characteristic in ambulatory robots, since it leverages momentum in order to conserve energy. Because of its complexity, the design of bipedal robots often draws inspiration from biological systems. For example Morimoto et al (2008) design bipedal humanoid robots using a sinusoidal pattern generating coupled oscillator [47]. This model is inspired by the simplicity of biological systems. Saputra et al (2015) show that using a model neuro-locomotor system with 'motoric' and 'sensoric' neurons is effective for the control of 3-D locomotion [48]. Since biological bipedal locomotion is a process that was reached through many years of evolution, and due to the adaptive nature of neural systems, it is likely a highly optimized process worth understanding and recreating to achieve effective bipedal locomotion in robots.

Recovery of locomotor function after ailments like stroke, spinal cord injury, or lower-limb amputation is a highly desirable clinical outcome. The extent to which it is possible to restore locomotor function is related to the ability to establish metrics that indicate normal behavior. Such metrics can be used to track the efficacy of different treatment strategies over time when applied to clinical populations. Various metrics have been proposed, for example the locomotor capabilities index

or LCI [49], which is meant to measure lower-limb amputees' walking ability with prostheses. The LCI is a questionnaire with 14 questions that individuals self report answers to. The questions range from whether someone get can get up from a chair, to whether they can step up and down a sidewalk surb. As individuals recover their ability to perform different locomotor functions, they will be able to complete more of the tasks listed in the LCI. The different tasks in question were derived based on locomotion research and a panel of experts and people with lower limb amputation.

Locomotion in this thesis

This thesis centers around data collected from human subjects performing a complex terrain locomotion task. While there is some indication that energetic costs influence locomotion, in natural terrain the walker must integrate different costs factors such as energy and stability at the same time as gathering time-varying visual information from the terrain to plan and control foot placement and direction. There have been few attempts to consider locomotion in this complex decision-theoretic context. One of the goals of this thesis is to provide empirical data about how locomotion unfolds over complex terrain in a naturalistic setting. The natural context allows us to address these more complex issues of action choices.

1.2.4 Photogrammetry

Background

Photogrammetry is a computer vision technique that attempts to reconstruct objects and the environment from sequences of images. It leverages the fact that objects and the environment are staying still while the camera is moving relative to them, which can be used to infer the structure of the environment and how the camera is moving relative to the environment. While there are subtle differences between photogrammetry, and other techniques like SLAM (Simultaneous Localization And Mapping) and SfM (Structure from Motion) [50], the main outputs of

environmental structure and camera trajectory remain the same.

Photogrammetry in this thesis

While head mounted eye tracking allowed insights into gaze control and how it reflected the underlying neural decision machinery, it relied on linking the gaze behavior both to the actions performed, and the object in the world. Thus, if a subject is fixating a cup, both the hand action and the identity of the item in the scene identify the cognitive goal. Thus a complete analysis of natural behavior requires both the monitoring of the body, together with gaze, and a metrical representation of the scene. While virtual reality can provide known scene structures, analysis has been problematical for eye movements collected in real environments. In this thesis, photogrammetry allows 3D metrical representations of the terrain subjects walk over. Together with the motion capture date for body posture. this allows a more quantitatively precise analysis of both the visual input and the visuo-motor control loop guiding foot placement. Critically, it allows relating features of the 3D structure of the terrain to the ongoing behavior.

Chapter 2

Visual motion statistics

2.1 Introduction

For a moving observer travelling through a stationary environment, visual motion is determined by a combination of the observer's movement through the environment and the structure of the environment. The resulting pattern of motion is commonly referred to as optic flow [51], [52]. This motion input carries information about the environment as the eye moves through space and rotates. This in principle extends to all visual motion signals, not just those arising from ego-motion. Even in a relatively stable state (such as sitting at a desk), there are small movements of the head and eye through space while the eye remains fixated. In these situations the motion pattern is likewise determined by observer movement through space, and the structure of the environment. Processing and using information about the environment extracted from optic flow is thus an important function of visual motion processing systems, since this can support adaptive behavior in different contexts.

Visual motion selectivity is present in the responses of V1 cells, which exhibit both direction and velocity sensitivity. V1 projects both directly and

indirectly to primate area MT [53] which appears to integrate and segment computed motion signals originating in motion sensitive V1. Additionally, direction encoding in MT is correlated with behavioral variability, suggesting that a motion perception signal upon which decision making is based could originate in MT [54]. However understanding more complex response properties of this and other motion sensitive areas ultimately requires knowledge of the natural statistics that shape the neural selectivity at both evolutionary and developmental timescales. While there has been progress in relating the properties of MT and MST to perception and behavioral goals [55], [56], [57], it is difficult to do this effectively without an explicit knowledge of the motion input and how it is shaped by behavior. Informal thinking about how stimulus and actions are linked in natural scenarios can be limiting [58]. For example, while the response to particular stimuli can be evaluated accurately, the extent to which this is informative of how the system functions normally is limited by how accurately the stimuli reflect natural inputs. This concern is valid for areas like MT, but is even more pressing in further downstream areas of the processing hierarchy where response properties can become even more complex. For example area MST has been well established as preferentially responding to global motion patterns, like those arising during self motion [59], [60], [61]. However not much more is known for certain beyond this general selectivity. Given the critical relationship between natural behavior and resulting retinal input, a different approach may be necessary. Analysis of visual response properties in the context of knowing how visual inputs are related to behavior goals may be important for deeper understanding of the visual system.

In this paper we examine eye movements during locomotion, and explore how they shape visual motion properties. Previous work has examined

retinal motion statistics in a similar fashion [62], [63]. Here, retinal motion inputs were approximated by recording head trajectories, and eye direction statistics during locomotion, and using depth images in order to approximate retinal motion signals by simulating these head and eye movements through 3D environments. In these two studies, it was found that direction of ego-motion and gaze direction heavily influences the motion signal, and that position in the visual field is highly related to speed and direction statistics. They also found that tuning properties of model neurons derived using measured input statistics match those of motion processing neurons in primates. In the present work, we simultaneously record gaze and image data while subjects walk in a variety of different natural terrains. Since terrain is a profound influence on gaze deployment, the in situ data collection strategy allows a more precise evaluation of the natural statistics than the previous studies, and we focus on interactions between gaze behavior and the resulting motion patterns. We find a stereotyped pattern of gaze behavior that emerges due to the constraints of the task, and this pattern of gaze drives much of the variation in the resulting visual motion pattern. The manner in which gaze affects the motion pattern is systematic, which has implications for neural processing. Vertical gaze angle has a large effect on the velocity pattern of motion, whereas horizontal has a large effect on the direction pattern of motion. We also examine how the relationship between gaze and motion signal interacts with different terrains, and what implications it could have for neural processing.

2.2 Experimental task and Data Acquisition

Participants: Each dataset used in this study was collected using the same apparatus, but from two separately conducted studies with similar experimen-

tal conditions. One group of participants (n=3) was recruited with informed consent in accordance with the Institutional Review Board at the University of Texas at Austin. The second participant group (n=8) was recruited with informed consent in accordance with the Institutional Review Board at The University of California Berkeley.

Equipment: Infrared eye recordings, first person scene video, and body movements of all participants were recorded using a Pupil Labs mobile eye tracker [14] combined with a Motion Shadow full body IMU based motion capture system (Motion Shadow, Seattle, WA, USA). The eye tracker has two infrared eye cameras, and a single outward facing scene camera. Each eye camera records at 120Hz at 640x480 resolution, while the outward facing scene camera records at 30Hz with 1920x1080 pixel resolution, with a 100 degree diagonal field of view. The scene camera is situated approximately 3cm above the right eye. The Shadow motion capture system is comprised of 17 3-axis accelerometer, gyroscope, and magnetometer sensors. The readings from the suit are processed by software to estimate joint positions and orientations of a full 3D skeleton. The sensors record at 100Hz, and the data is later processed using custom Matlab code (Mathworks, Natick, MA, USA). See [64] and [23] for more details.

Experimental Task: For both groups of participants (Austin dataset and Berkeley dataset) the experimental task was similar, with variation in the terrain types between the two locations:

For the Berkeley participants, the task involved walking back and forth along a hiking trail that varied in terrain features and difficulty. This trail was traversed two times in each direction by each participant. Different portions of the trail were pre examined and designated as distinct terrain types, being labeled as one of pavement, flat, medium, and bark, and rocks.

For the Austin participants, the task involved walking back and forth along a rocky dried out creek bed. Participants walked 3 times in each direction. This is the same terrain used in [64]. This terrain difficulty is most comparable to the rocks condition in the Berkeley dataset. For each of the rocky terrains, the ground was sufficiently complex that subjects needed to use visual information in order to guide foot placement (see [64] for more details).

Calibration and post-processing: At the start of each subjects recording session, they were instructed to stand on a calibration mat that was used for all subjects. The calibration mat has marked foot locations that are at a fixed distance from a calibration point that is 1.5 meters away from the center of the two foot locations. Subjects were then instructed to maintain fixation on this calibration point throughout a calibration process. The calibration process involves rotating the head while maintaining fixation, to 8 different pre-determined head orientations (the cardinal and oblique directions). This segment of each subjects recording is later used to align and calibrate the data. This is done by finding an optimal rotation that aligns the mobile eye tracker's coordinate system to that of the motion capture system, such that the distance between the projected gaze vector and the calibration point on the mat is minimized. This rotation is then applied to the position data in the recording, aligning the rest of the recording in space. Each system's timestamps are then used to align the recording temporally, as timestamps are recorded to a single laptop computer on a backpack worn by subjects throughout the recording. (see [64] for more details). The 100Hz motion capture data is upsampled using linear interpolation to match the 120Hz eye tracker data.

Fixation detection: Fixation detection is a crucial step in understanding the motion input during this behavior, and is a non-trivial problem in head-free paradigms. Many different solutions have been proposed from dis-

persion [65], to combinations of acceleration and velocity [66], to even higher order measures like jerk [67]. Recent work has also proposed using neural networks [68]. Here we use a velocity and acceleration threshold method with thresholds set such that detected fixations best match hand coded fixations. We use a threshold of $65deg/s$ velocity and $5deg/s^2$ acceleration. Frames of the recording that fall below these thresholds are labeled as fixation frames, and sequential fixation frames are grouped accordingly into fixation instances.

Fixation idealization via image feature pinning: Analysis of fixational eye movements revealed robust stabilization of gaze location in the world, while the eye counter-rotates in the orbit as the body moves forward during a step (see Figure 2.3). This stabilization is primarily accomplished by the vestibular-ocular reflex (VOR). The effectiveness of this stabilization mechanism was determined by measuring the deviation of gaze from initially fixated locations over the course of fixations, which had a median deviation of 0.83 degrees, and a mode of 0.26 degrees (see 2.8.1 for more details). Given this result we opted to idealize fixations, enforcing 0 deviation from the initial fixation location over the course of fixations. This trades off eye tracker noise with however much instability there actually is. Re-introduction of arbitrary amounts of retinal slip can then be used to gauge the impact of this manipulation.

Retinocentric coordinate system and eye movement directions: The first goal of the study was to compute the retinal motion statistics caused by movement of the body during fixations. This retinal motion is caused by expansion of the image as the walker takes a step forward, together with rotation around the yaw axis. After fixations are detected and each frame is classified as either a fast (saccadic) or slow eye movement (stabilizing) frame, directions of eye movements are computed for each. This is done by consid-

ering sequential frame pairs that are either both stabilizing frames or both saccadic frames. Then for each first frame in a pair, eye coordinate basis vectors are calculated. Here the gaze direction is the third dimension (orthogonal to the plane within which the eye movement direction will be calculated), the first dimension is the normalized cross product between the eye direction and the negative gravity in world coordinates. The second dimension is the cross product between the first and third dimensions. This convention assumes that there is no torsion about the viewing axis of the eye, and so one dimension (the first) stays within a plane perpendicular to gravity. Using these coordinates, the direction of the next eye direction (corresponding to the eye movement occurring over frame pairs) is computed in the reference frame of the first eye direction's coordinates (x, y, z). The direction is then computed as

$$atan2(y, x)$$

. A schematic depicting this coordinate system can be seen in Figure 2.1

2.3 Oculomotor patterns during locomotion

Figures 2.2 and 2.3 show excerpts of representative data from a subject during locomotion. When the terrain is complex, subjects mostly direct gaze toward the ground a few steps ahead [64]. This provides visual information to guide upcoming foot placement. As the body moves forward, the subject makes a sequence of saccades to locations further along the direction of travel. Following each saccade, gaze is held approximately stable for periods of around 200 msec so that visual information about upcoming foothold locations can be acquired while the subject moves forwards during a step (Bonnen et al 2021).

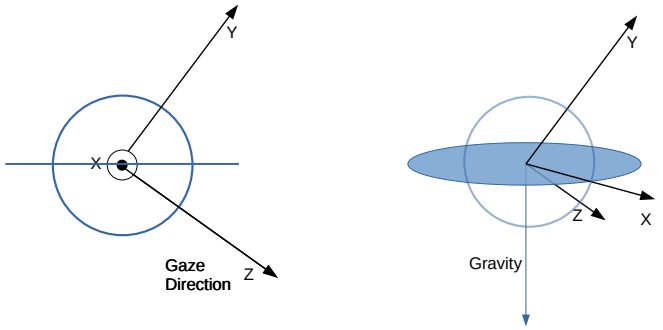


Figure 2.1: Schematic depicting eye relative coordinate system. Left and right show basis vectors used from different view points. Z corresponds to the gaze direction in world coordinates. Then X is the vector perpendicular to both Z and the gravity vector. Finally The Y coordinate is the vector perpendicular to both X and Z. The X vector thus resides within the plane perpendicular to gravity.

This results in downward slow rotations of the eye to offset the forward motion of the body. Stabilization is most likely accomplished by the vestibular ocular reflex, although other eye movement systems might also be involved [23]. There is variation in how far ahead subjects direct gaze between terrain types (see Figure 2.4), although this pattern of saccades followed by stabilizing eye movements is conserved.

This pattern of eye movements is important to highlight because in conjunction with the pattern of head movements resulting from locomotion, it determines the pattern of retinal motion. Approximating this retinal motion pattern thus requires taking the eye movements into account. It is particularly important to detect and isolate fixations since saccadic suppression interferes with vision during saccades [69], [70], although the neural mechanisms are still poorly understood. In order to take this observation into account in our analysis of motion statistics, saccades are excluded from visual motion

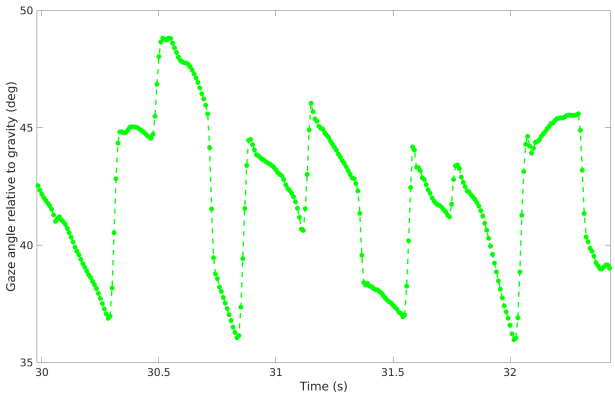


Figure 2.2: Characteristic excerpt of vertical gaze angle trace during rocky terrain navigation. As subjects navigate complex terrain, they acquire visual information through alternating saccades and stabilizing eye movements. Gaze behavior is stereotyped, with a general pattern of alternating upwards saccades (increasing gaze angle, over a short period of time) followed by downwards stabilizing eye movements (decreasing gaze angle, over a more extended period of time).

approximation related analysis. However the statistics of saccade direction and magnitude are computed separately for completeness (see Appendix 2.8.4).

Analysis of eye movement direction statistics reveals a strong bias in the downward direction for slow eye movements, and upward direction for saccadic eye movements. Data from this analysis can be seen in Figure 2.5. For the slow eye movements, the average direction was -91.5 with a standard error of 1.29 degrees (with -90 corresponding to straight downward). In contrast the saccadic eye movement average direction was 89.125 with a standard error of 1.97 degrees (90 degrees being straight up). The standard deviation of all directions (across all subjects) for each distribution was 51.167 degrees and 48.997 for stabilizing and saccadic movements respectively.

The observed pattern of eye movement direction distributions is strongly driven by the particularities of the ongoing behavior. As walkers search for

Figure 2.3: Schematic depicting typical sequence of eye movements. As subject approaches new unknown terrain, saccades to new potential foothold locations are made (left) followed by stabilizing slow eye movements while visual information is extracted (right).

and identify suitable footholds, they must direct gaze further along the path in order to anticipate arrival at future locations. However, due to saccadic suppression mechanisms including image blur, useful visual information is unlikely to be extracted from the environment during saccades. This means that walkers need to fixate points on the ground, and the oculomotor system must produce a slow downward eye movement in order to stabilize the image on the retina as the walker approaches the fixated location. These stabilized downward eye movements occur regularly throughout different terrain conditions. After sufficient visual information is extracted during the slow eye movement, gaze must then be directed upwards in order to begin search and identification of new footholds further along the path. These upwards eye movements are mostly saccadic. These two stereotyped eye movements give rise to the difference in the direction distributions of stabilizing and saccadic eye movements.

This pattern of eye movements has a strong effect on the resulting retinal motion pattern [23], making it essential to incorporate the effects into any

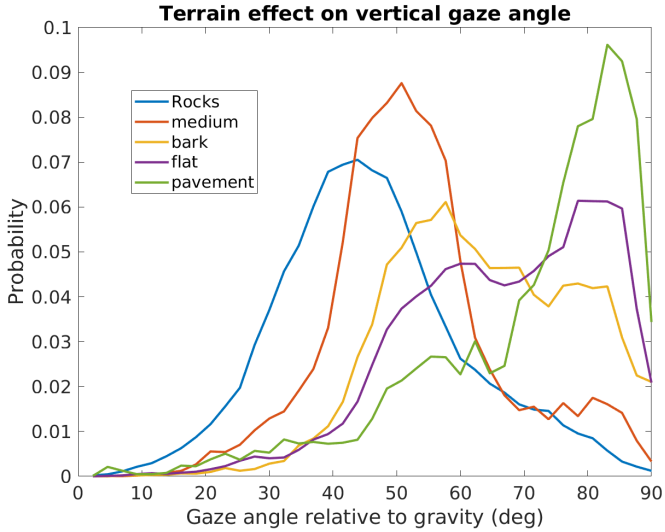


Figure 2.4: Vertical gaze angle (angle relative to gravity direction) distributions across various terrain types. Subjects walked across various terrain types which induced different gaze behaviors. Terrain variability ranges from almost entirely flat (pavement) to complex, requiring lots of visual information to navigate (rocky)

analysis of motion inputs.

2.4 Motion statistics depend on gaze angle

2.4.1 Optic flow estimation based retinal Motion approximation

In order to approximate retinal motion input to the visual system, we combine eye movement measurements from the mobile eye tracker with a computer vision optic flow estimation algorithm (DeepFlow [71]). Video from eye tracker's scene camera (subject first person perspective) is broken into sequential frames and input as frame pairs into the optic flow estimation algorithm. This provides for each sequential frame pair a motion vector for each pixel providing an estimate of where the pixel moved from the first frame to the second. Because the scene camera video is first undistorted so that a pinhole camera can be

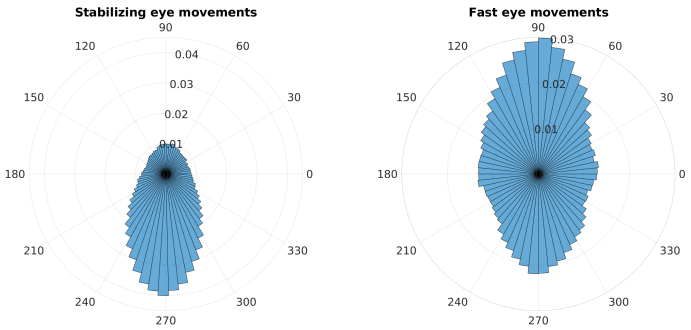


Figure 2.5: Polar histograms of eye movement directions. Eye movements were classified as either fast or slow based on previous fixation detection. Stabilizing (slow) eye movement distribution on the left, fast (saccades) on the right. 270 degrees corresponds to straight down in eye centered coordinates, 90 corresponds to straight up. Stabilizing eye movements are largely in the downward direction, where stable locations on the ground are fixated by subjects as the approach them, causing a slow downward eye movement. Some upward eye movements occur and are likely due to small saccades that are classified as fixations and perhaps gait-related head/body sway while fixating. Saccadic eye movement are largely in the upward direction, as subjects look further ahead as they navigate the terrain, and have already acquired visual information from the currently fixated location. Some saccades closer to the body are also observed presumably when foot placement requires additional information

assumed, these motion vectors can be thought of as existing in an image plane situated at a distance of one focal length away from the camera (perpendicular of the direction of the camera). Using the focal length of the camera in pixels (estimated using camera calibration [72]), the 3D direction vectors relative to the camera center of each of the pixels, as well as each of their corresponding pixel destinations computed by DeepFlow can be computed. Thus for each frame pair, for each pixel, there are a pair of 3D vectors, one corresponding to the initial location of the pixel, and one the final after the visual motion has

occurred. Theses pairs of vectors can be thought of as rotations of each pixel around the camera.

The eye is then assumed to be in the same location as the camera, which is convenient for the purposes of retinal motion approximation (the resulting error is trivial, since this would be the difference from recording from a hypothetical camera located at the eye center, which is a few centimeters from the actual camera). For each frame pair of the video, the corresponding eye in head movement (also represented as an initial point and final point in the image plane, and as a result two 3D vectors, or a rotation) can then be calculated. The rotational effect of the eye movement is then applied to all of the end points of the pixel motion vectors, incorporating the effects of the eye movement into the camera relative visual motion, resulting in eye relative motion. This eye relative motion is still in the reference frame of the camera (pixel coordinates), and is then converted into retinal coordinates by mapping eccentricity and polar angle to a 2D plane (eccentricity is known since the gaze location on the camera's image plane is known). This mapping is done through nearest neighbor sampling of the camera image (which has for each pixel a corresponding rotation, which is converted to a speed and direction in eye coordinates).

The retinal motion signal is represented as a 2D grid where grid points (x, y), which are each in the directions (X, Y) as in Figure 2.1, correspond to polar retinal coordinates (θ, ϕ), by the relationship

$$\theta = atan2(y, x)$$

$$\phi = \sqrt{x^2 + y^2}$$

. Thus eccentricity in visual angle is mapped linearly to the image plane

as a distance from the point of gaze. At each (x, y) coordinate there is a corresponding speed in $\frac{deg}{s}$ and direction $atan2(\Delta x, \Delta y)$ of movement.

Approximation of retinal motion that incorporates the eye translations and rotations through space as a consequence of the motion of the body during locomotion is especially important for this behavior, where there is a stereotyped pattern of saccades and fixations as subjects walk forwards (see Figure 2.2). We include only the data from fixation frames, as determined by the fixation detection process described in 2.2.

2.4.2 Mean motion speed and direction statistics

Subject gaze angle has a large effect on the pattern of retinal motion because of the resulting planar geometry [73]. Here we first consider the average motion signal across all data, where the effects of the overall distribution of gaze angles produce particular features in both the speed and direction patterns. We will then explore the effects of gaze angle more directly.

Approximated retinal motion described in 2.4.1 is averaged over all terrains and all subjects to compute a mean flow field. Both speed ($\frac{deg}{sec}$) and direction are averaged, and the resulting mean flow field is visualized in terms of speed and direction. Speed is colormapped to blue to yellow, with blue being the lowest velocity, and yellow being the highest. Directions at different retinal locations are represented by unit vectors indicating the direction.

The average flow field shows regular patterns in the spatial distribution of speed and direction. Results from this analysis can be seen in Figure 2.6. Speed increases as a function of eccentricity, although this increase is not radially symmetric. The lower visual field has steeper speed increases as a function of eccentricity compared to the upper. The left and right visual fields are even lower than the upper visual field. Average speeds in the lower visual

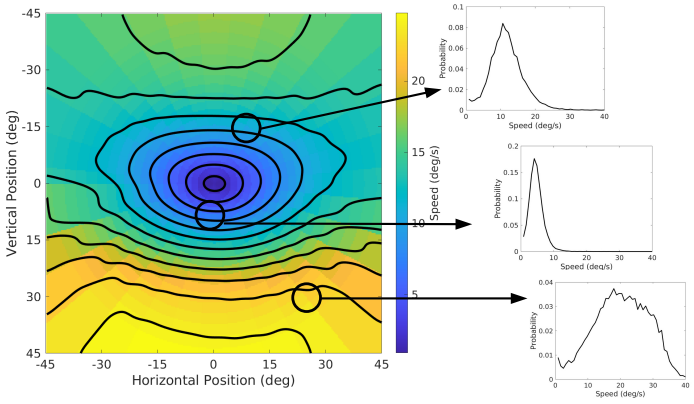


Figure 2.6: Average speed of retinal motion signal as a function of retinal position. Speed is colormapped (blue = slow, red = fast). Average is computed across all subjects, and terrain types. Speed is computed in degrees of visual angle per second. Shown on the right are three example distributions at particular retinal locations. Speed variability also increases along with the mode speed of the distributions, which can be seen in the distribution widths

field peak at 24.6 deg/s (at 45 degrees eccentricity) whereas the upper peaks at 17.4 (also 45 degrees eccentricity). Average speeds within the central 5 degrees of eccentricity do not exceed 3 degrees/second.

The average directions of flow exhibit a radially expansive pattern. This can be seen in Figure 2.7. Motion direction is away from the fovea, however the degree of pure expansive motion directions (directly away from center) is also not radially symmetric. Directions are biased towards vertical directions, with only a narrow band in the left and right visual field exhibiting leftward or rightward motion.

The main features of the mean motion signal arise due to the particular pattern of gaze behavior. The ground fixations result in a situation where from the perspective of the eye, the ground plane is translating towards the eye while also rotating (as the observer moves mostly parallel to the ground plane at eye height, and fixates a ground point resulting in rotation). The

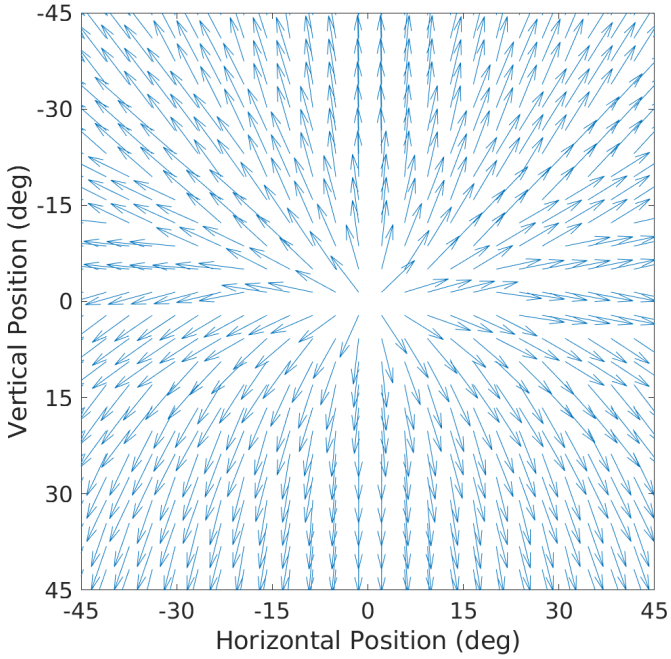


Figure 2.7: Average direction of retinal motion signal as a function of retinal position. Direction is indicated by a unit vector drawn at particular location. Here direction corresponds to direction in a 2d projection of visual space, where eccentricity from the direction of gaze in degrees is mapped linearly to distance in polar coordinates in the 2d projection plane.

component of eye velocity towards the fixation location results in the expansive pattern that can be seen in the spatial distribution of direction. The rotational component, as well as the eccentric angle that the ground plane is viewed from (not perpendicular like with a fronto-parallel plane) results in distortions from a radially symmetric expansive pattern. This can be seen slightly in the spatial distribution of directions, where directions are away from the fovea, with slight upwards and downward biases in the upper and lower visual fields respectively. The increasing speeds as a function of eccentricity, along with the radially asymmetry of this effect, are also related to the towards and rotational

components of the ground plane relative to the eye.

2.4.3 Effects of horizontal and vertical gaze angle on motion pattern

While navigating complex terrain, subjects direct gaze towards the ground, which affects the average motion pattern. However gaze angle changes substantially over the course of the behavior, depending on the demands of the terrain, which in turn drives changes in the experienced motion pattern.

The same method for computing an average flow field is applied to subsets of the data, separated by horizontal and vertical gaze angle. Vertical gaze angle is defined as the angle of gaze in world coordinates relative to gravity. Horizontal angle is defined relative to the overhead projected direction of head translation, with 0 degrees meaning gaze is in the same direction as head translation, and positive being rotated clockwise when viewed from above, negative being counter-clockwise. Frames of recordings for all subjects are binned by their respective horizontal or vertical angles (-60 to 60 in 40 degree increments for horizontal, and 0 to 90 in 30 degree increments for vertical, and average flow fields are calculated within bins. Vertical gaze angle is measured relative to gravity, and so the effect is driven by different terrain demands that cause the subject to direct gaze closer or further from the body. Horizontal gaze angle is defined relative to the direction of travel, and as a result reflects interactions between gaze direction and body motion.

Horizontal gaze angle (defined as clockwise angle of gaze direction relative to ground plane projection of eye translation direction) has a substantial effect on the spatial pattern of motion direction, with a smaller effect on the spatial pattern of motion speeds. This can be seen in Figure 2.8, panel A. The orientation of iso-contour lines changes slightly, although a pattern of increas-

ing speed with eccentricity particularly in the lower visual field is maintained. The global motion direction pattern shows a systematic dependence on horizontal gaze angle, changing from expansive motion centered at the fovea to expansive motion with additional rotation around the fovea (either clockwise or counter clockwise depending on if gaze is directed to the left or right relative to the translation direction). This makes motion directions more perpendicular to the radial direction of the retinal location of the motion, as gaze becomes more eccentric relative to the translation direction. In other words, as gaze is directed at a higher angle relative to the direction of travel, there is more rotational motion about the gaze location introduced into the input.

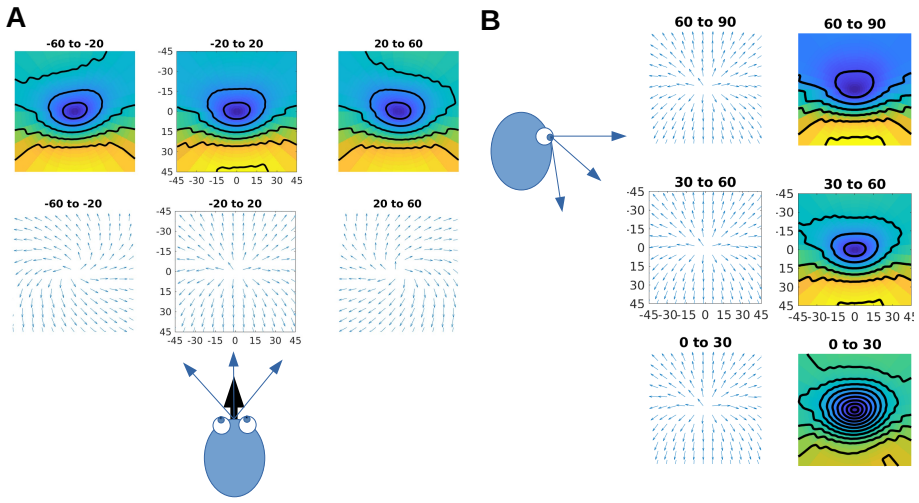


Figure 2.8: Effect of horizontal and vertical gaze angles (measured relative to the head translation direction and relative to gravity respectively) on motion speed and direction. Here average motion speed or direction is computed across the visual field, for ranges of horizontal or vertical gaze angles. Frames of the recording where the horizontal or vertical gaze angle falls within bins of width 40 degrees, with edges ranging from -60 to 60 for horizontal, and 30 degree width bins from 0 to 90 for vertical, are averaged within bins.

The relationship between angle and speed or motion direction is differ-

ent for vertical gaze angle, which can be seen in Figure 2.8, panel B. There is not a noticeable effect on the spatial pattern of motion direction. However the motion speed shows a significant change. As gaze is directed more towards the horizon, the pattern of increasing speed as a function of eccentricity becomes more radially symmetric. The opposite occurs with decreasing gaze angle, as gaze is shifted more towards straight downwards, the distribution of motion speed becomes more radially asymmetric, with much of the upper visual field showing similarly low speeds to the fovea, and the lower visual field having higher speeds.

Horizontal gaze angle had a noticeable effect on the spatial distribution of motion direction, with only a slight effect on speed, whereas vertical gaze angle has more of an effect on speed than direction. Horizontal gaze angle has a general effect of adding either clockwise or counter clockwise rotation the expansive motion pattern, resulting in an expansive and rotating flow field centered at the fovea, with the degree of rotation scaling with gaze eccentricity from straight ahead. These flow fields are reminiscent of the types of optic flow stimuli known to provoke strong responses in MST neurons, whose role in processing optic flow is well studied [17], [18], [19], [20], [21], [22]. Many MST neurons have exhibited tuning in 'spiral space' [74], of which rotation is a dimension that seems well captured by variation in horizontal gaze angle. Vertical gaze angle alters the spatial pattern of speed such that it is more radially symmetric at higher angles (more towards the horizon). At higher angles (more towards the horizon) speeds are much higher in the lower visual field than upper. The radially symmetric pattern experienced at lower vertical gaze angles is most reminiscent of optic flow fields one might generate using a dot field, however the radially asymmetric signals experienced at higher gaze angles (more within the distribution observed for more difficult terrains) has

the distinct upper vs lower field asymmetry, which is prominent in the mean flow signal (see 2.6).

2.5 Visual motion generated by saccades and imperfect stabilization

2.5.1 Eye movement speed statistics

Following categorization as either saccadic or stabilizing eye movements, the speed of the eye movements was quantified. The results are shown in Figure 2.9. Each plot shows a 2D histogram of eye movement velocities, broken into horizontal and vertical components.

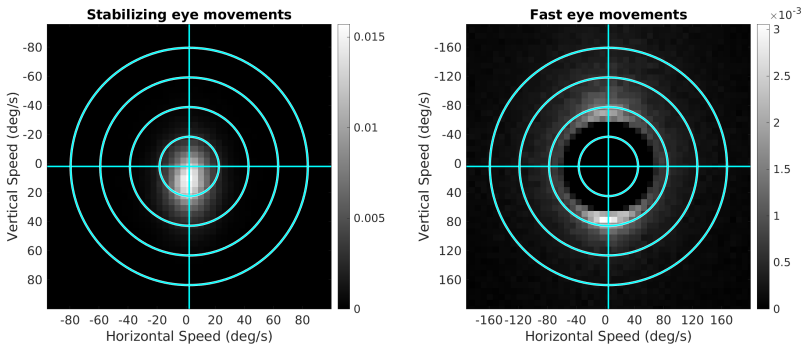


Figure 2.9: 2D histograms of vertical and horizontal components of eye movements. Note axes on right plot are double the scale of the left plot. Outermost circle on left plot fits into blank region in center of right plot. Grayscale colormap shows probability of vertical and horizontal eye components for either stabilizing (left) or fast (right) eye movements. Stabilizing eye movements are concentrated downwards, whereas fast eye movements are more dispersed, with a higher concentration in the upward direction.

For saccadic eye movements (Figure 2.9, right), the distribution of vertical speeds is bimodal, with peaks at -80 and +80 deg/s (negative corresponding to upwards), with a higher density in the upwards direction. These eye movements mostly direct gaze towards new possible foothold locations that are further along the path, with some saccades back downwards during the search. The distribution is horizontally symmetric.

For stabilizing eye movements (Figure 2.9, left), the distribution of vertical speeds is shifted in the downward direction, towards -9 deg/s. Like with saccades, the distribution of horizontal speeds is symmetric. These slow eye movements in the downward direction stabilize the image of the upcoming terrain on the retina as subjects approach as they walk forwards.

The hypothetical effect of saccadic eye movements on the visual motion signal is quite large, since it would add the speed of the saccade to an instantaneous motion input, and shift motion directions in the opposite direction of the saccade. On the other hand, the stabilizing eye movement is what confers the observed properties to the motion signal described above. While it is difficult to estimate the exact contribution of imperfections in stabilization (slight over or under compensation for the observers forward movement by the downwards slow eye movement), we provide a metric that suggests stabilization performance results in maintenance of initial fixation targets within a small region of the fovea (2.8.1).

2.6 Gaze angle independent terrain effects on motion signal

2.6.1 Between terrain vertical gaze angle

Subjects navigated a path with variable terrain features. Different segments of the path were segmented manually by observing common features and walking path complexity. Since subjects walked the same path the labelling of separate terrain types was consistent across subjects. In this analysis the distributions of vertical gaze angle were computed for different terrains across all subjects. Vertical gaze angle is the angle between the direction of gravity and the current gaze angle.

Gaze angle distributions vary across terrain types, which can be seen in Figure 2.4. In order of increasing terrain complexity (Rocks, medium, bark, flat, and pavement), the median vertical gaze angles (relative to straight downwards) were (in degrees) 43.7, 50.33, 62.54, 67.95, and 75.9 respectively. With the exception of Rocks, the distributions all had degrees of bimodality, with lower peaks at 50.77, 57.70, 60, and 62.3 (for medium, bark, flat, and pavement respectively), and higher peaks of 80.77, 78.46, 78.46, 83.08.

Walkers seem to adopt different gaze strategies as the demands of the terrain change. From most complex (rocky) to least complex (pavement), vertical gaze angle changes from closest to feet, to mostly directed towards the horizon. For the rocky terrain, the amount of uncertainty regarding stability of foothold locations is higher, since terrain is more variable compared to flat ground with elevations and crevices due to the large rocks. In contrast, the pavement is completely flat by design, and so past experience would suggest to walkers that no visual information is necessary to coordinate foot placement,

so gaze is directed towards the horizon. Interestingly, the 'flat' condition also has flat ground where stability is guaranteed, however it is not paved. This makes walkers more cautious than on paved ground, with gaze being directed downward more often, possibly because surface friction is less than on the paved road, hence more visual information might be needed about more immediate stepping locations.

2.6.2 Photogrammetric reconstruction based retinal motion approximation

The relative contributions of non flat ground plane structure and the ground plane itself to motion are of interest. Presumably, because of the deterministic nature of flow patterns for completely flat ground given a particular eye translation and rotation, walkers can learn the expected signal for completely flat ground. Thus the visual system may be able to exploit this regularity by assessing deviations from the expected signal, however the magnitude of the deviation needs to be calculated to explore this. This could in principal be done by computing a 'flat ground equivalent' for each frame of data, and the computing the difference. However any noise in the head and eye orientation measurements would result in slight misalignment of corresponding locations in visual space when comparing the two signals. To account for this, retinal motion is approximated using photogrammetric reconstruction of the terrain. Doing so allows explicit control of this alignment, since both motion signals are computed based on the same eye orientation measurement, and the terrain used to compute the motion signal is manipulated. This is done by computing depth maps and treating these as a ground truth geometry. We use Photogrammetry package Meshroom [75] in order to simultaneously reconstruct terrain and approximate camera position and orientation from the eye tracker

scene camera video. Meshroom does this by taking as input a sequence of images and using a series of image and 3D point processing methods. The output is a 3D triangle mesh representation of the environment, in the same coordinate system as the estimated camera positions and orientations.

Using Blender [76], the 3D triangle mesh representations of the terrain are combined with the spatially aligned eye position and direction data. A virtual camera is then placed at the eye location and oriented in the same direction as the eye, and a depth image is acquired using Blender's built in z-buffer method. Thus the depth image input at each frame of the recording is computed. Visual motion in eye coordinates can then be computed by tracking the movement of projections of 3D locations in the environment onto the plane described in 2.4.1 as a consequence of translation and rotation of the eye (see [77] for generalized approach).

For each instance of retinal motion input calculated from a depth image, a corresponding flat ground version is also computed. This is done by computing the motion input assuming a flat ground plane. The flat ground plane is normal to gravity, and is placed at a height equal to height of the fixated location in the original depth image.

The same convention for representation of retinal motion described in 2.4.1 is used here. See Appendix 2.8.2, Figure 2.15 for comparison of eccentricity dependent speed distributions between methods.

2.6.3 Flat ground normalized MT response distance

Here for each terrain type, retinal flow inputs are resampled such that the associated vertical and horizontal gaze angle distributions match the overall combined distributions (to avoid bias due to gaze angle effects). Then the flow fields computed from depth maps and simulated flat ground flow field for

the same instance are transformed to MT-like representations of local motion (see Appendix 2.8.3 for analysis on speed rather than model cell response). Here a similar approach to [78] is taken. A polar grid 20 x 32 total retinal locations are considered, with 5 x 8 different model MT cell responses for each location, corresponding to 5 different speeds and 8 different directions. There are 20 different eccentricities (ranging from 2.25 to 45 degrees at 2.25 degree intervals intervals) with 32 different polar directions, and motion within a pooling area similar to MT receptive field size given the eccentricity [79]. The 5 x 8 different MT cells correspond to 5 different preferred speeds (2,4,8,16,32 $\frac{deg}{s}$) with 8 different preferred directions (cardinal and oblique directions). We use the same model described in [78], where the activity of each model MT cell r_{MT} at location (x, y) with preferred speed ρ_{pref} and preferred direction θ_{pref} is given by:

$$r_{MT}(x, y; \theta_{pref}, \rho_{pref}) = d_{MT}(x, y; \theta_{pref}) s_{MT}(x, y; \rho_{pref}),$$

Where d_{MT} and s_{MT} are the direction and speed responses of the unit respectively. The direction response is given by:

$$d_{MT}(x, y; \theta_{pref}) = exp(\sigma_\theta(cos(\theta(x, y) - \theta_{pref}) - 1))$$

with $\sigma_\theta = 3$. Speed response is given by:

$$s_{MT}(x, y; \rho_{pref}) = exp\left(- \frac{log^2(\frac{\rho(x,y)+s_0}{\rho_{pref}+s_0})}{2\sigma_\rho^2} \right)$$

with $\sigma_\rho = 1.16$ and $s_0 = 0.33$. The resulting response is bound by [0,1]. See [78] for more details.

The 20 x 32 x 5 x 8 different MT responses are computed for each flow

field, both on the actual flow field and the corresponding flat ground flow field (see 2.6.2). The euclidean distance between the 40 dimensional (5 x 8) firing rate vector at each of the 20 x 32 locations is computed between the actual and corresponding flat ground flow fields. The result is a 20 x 32 distance map for each flow field. These distance maps are then averaged within terrain types.

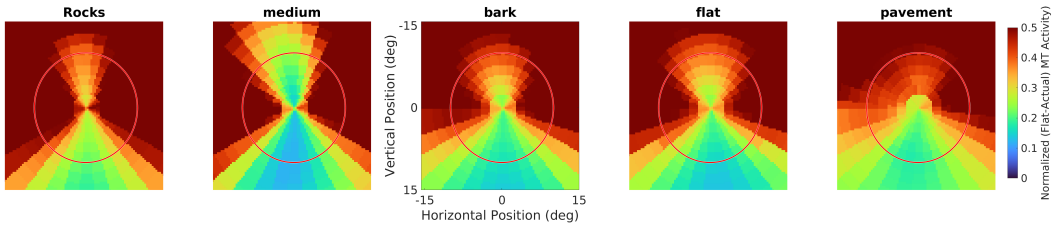


Figure 2.10: Visualization of mean normalized distance between MT like representation of motion signal at different retinal locations across terrains. Red circle with 10 degree radius indicates central region considered in next analysis. At each retinal location, the distance between the normalized firing rate vector for each simulated flat ground vs actual recorded data frame is computed. Distances are then averaged across all inputs within each terrain type. In colormap red corresponds to higher distance (more deviation from flat ground).

Average normalized MT firing rate vector distances show differences between terrains, which is shown in Figure 2.10. For the central 10 degrees of visual angle, the median distances between actual and flat ground simulated MT responses were 0.245, 0.259, 0.296, 0.346, and 0.460 for pavement, medium, bark, flat, and rocks respectively. As shown earlier in Figure 2.4, each terrain induced a different gaze angle distribution. In Figure 2.11 the relationship between flat ground deviation and median vertical gaze angle is shown. Median vertical gaze angles for the different terrain types were 75.90, 67.95, 62.54, 50.33, and 43.70.

The computed deviation from flat ground for each terrain type's average motion signal was inversely related to the median vertical gaze angle. In

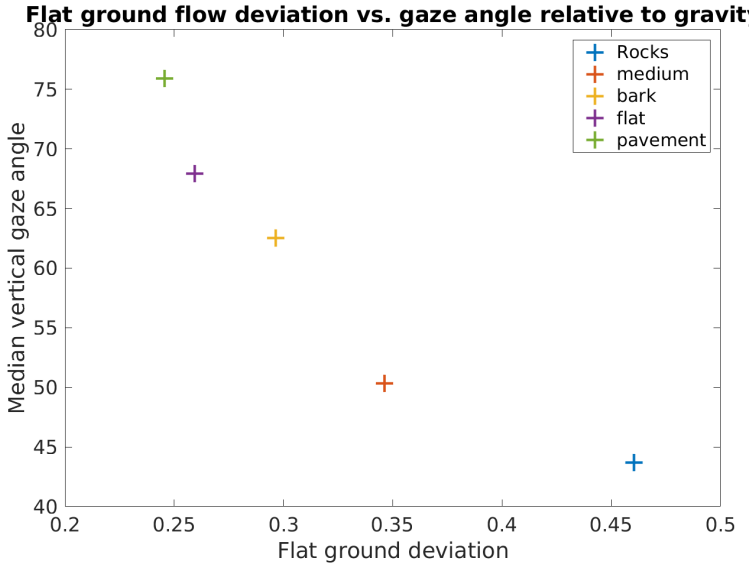


Figure 2.11: Median distance of foveal MT like representation between flat ground simulated and actual input plotted against median vertical gaze angle for each terrain. The median flat ground deviation (measured in terms of distance between normalized model MT firing rate vectors of simulated flat ground vs actual motion) is calculated for the central 10 degrees of eccentricity for each terrain.

other words, the more complex (less like flat ground) a terrain, the closer to their body the walker needed to direct gaze as they navigated the terrain. This provides some insight into how the gaze allocation and foothold finding strategy might change across terrains. It is also evidence of some kind of visual signature (in this case in the motion pattern) that indicates terrain complexity, and could be used to signal different contexts in order for the walker to adjust strategies. Humans are known to flexibly switch visuomotor control strategies depending on context [80], [81], and this signal might be related to how different contexts are recognized. This could be tested with experimental setups that explicitly manipulate terrain complexity.

2.7 General Discussion

Here we present a novel method for approximating retinal motion signals during natural behavior, as well as the simulated corresponding flat ground motion signals. This has allowed us to measure eye movement statistics, gaze angle distributions, retinal motion statistics, and VOR stability. Additionally it has allowed measurement of surface complexity in a perceptual variable (the normalized firing rate vector distance between flat ground and actual signal).

2.7.1 Mean speed and direction of retinal motion

It is worth highlighting that the eye translations and resulting stabilizing rotations are what give rise to this particular pattern of motion. The flow field is reminiscent of Gibson's formulation of optic flow [51], with the focus of expansion centered right at the point of gaze. The emergence of this expansive flow field centered at the point of gaze might suggest that subjects preferentially look at the focus of expansion. The theoretical focus of expansion resulting from a paralyzed eye translating through space is however distinct from this observed average flow pattern, which results from the previously described combination of ground plane translation and rotation relative to the eye. [23].

Average signals have implications for the interpretation of neurophysiology experiments. For example, [78] shows that heading direction tuning in MSTd neurons could simply be an emergent property of efficient coding of the statistics of the input stimulus, without the specific optimization for decoding heading direction. Statistical summaries of input signals and the associated behaviors (such as gaze angle in this case) could prove useful in understanding the responses of motion sensitive neurons, since some responses properties may arise as a consequence of efficient coding. For example, one might expect to

find perceptual biases or over representation of neural tuning for greater speeds as a function of eccentricity. However this many not be a radially symmetric effect, since average motion speeds tend to increase most rapidly towards the lower visual field, followed by the upper, and then the left and right (as seen in Figure 2.6). Additionally, one might expect to see receptive fields that reflect the band of low velocities that results from the geometry of the ground plane, which can be seen in the contour lines in average speed signal visualizations.

It is worth mentioning that these average signals reflect those experienced during this particular class of locomotion, where visual information is needed. Different tasks with different gaze behaviors and scene geometries could result in different patterns of visual motion, however the influence of gaze angle on the resulting motion signal would still be of importance even for different tasks.

2.7.2 Effects of horizontal and vertical gaze angle on motion pattern

These effects have implications for how the brain might process optic flow information, since particular eye orientations have statistical relationships to particular flow patterns, which could be learned and exploited by the visual system. Effects of eye position on response properties of visual neurons have been extensively observed in different regions [82], [83], [84]. Perhaps this eye direction tuning is related to how similar or different a particular pattern of stimulation is to the average stimulus given a particular eye direction. "Mismatch" neurons have been observed in the visual cortex of mice, where firing rate is related to the degree of mismatch between the anticipated motion signal given a particular movement of the mouse and a fixed environmental structure, and the observed motion signal (which is controlled by the experimenter)

[85]. It is possible that a similar strategy might be employed given a static environment and a known statistical relationship between particular eye and body movements and patterns of visual motion.

2.7.3 Eye movement direction statistics

Each direction distribution has a clear peak in either the upward or downward direction, with some variation about the peak direction. Figure 2.9 does reveal key differences about the two classes of eye movements. The saccade distribution is more dispersed, and more asymmetric with peaks in both the upwards and downwards directions. The higher variation could be due to search saccades as subjects look for new suitable foothold locations, and this search spans both upwards and downwards directions. In contrast the stabilizing eye movement distributions are more concentrated downwards, as subjects counter rotate their eyes to stabilize the image as the translate forwards. It is worth noting that it is possible that small saccades get classified as stabilizing eye movements, resulting in wider distributions than one might expect for the stabilizing eye movement directions and horizontal and vertical speed. However, most of the variation in the direction distribution of the slow eye movements can be explained by three different factors. The first is that stabilization during the side to side sway of the body during locomotion requires downward eye movements with sideways components to compensate for this sway. This results in small deviations from straight downward movement directions. The second is that not all fixations are directly on future destinations, which leads to a similar effect as for side to side sway but even more pronounced if gaze direction differs substantially from the current movement direction. Finally, even in cases where fixations are directly on future destinations, this fixation could be taking place during a turn to face said destination, which would also

result in variations around straight downward eye movements.

The measured eye movement statistics raise an interesting question about eye movement systems. Given its evolutionary importance locomotion and the accompanying patterns of gaze may have shaped characteristics of eye movement systems. For example [86] shows higher resolution, and over representation of upper visual field superior colliculus neurons in macaque monkeys, which contrast with previous assumptions about symmetry. Whether such asymmetries exist in systems governing slower stabilizing eye movements is still unknown, but worth investigating given our results. [87] found modulation of retinal projections to the superior colliculus based on arousal, further demonstrating the importance of examining these systems in different contexts.

2.7.4 Gaze angle distribution across terrain types

The observed vertical gaze angle distribution shows cognitive flexibility regarding context shifts for the visuomotor system. Walkers are able to adopt different strategies (although the nature of these differences beyond gaze angle is unclear) depending on visually identifiable terrain characteristics.

Another interesting feature of these distributions is the bimodality of all of them except for the 'rocks' condition. Each has one peak more towards about eye height out in front of the walker (45 degrees), and one peak more towards the horizon (90 degrees). This suggests a common strategy between the terrains of foothold-finding fixations with look-ahead fixations interspersed. The height of the lower peak decreases with terrain complexity, whereas the horizon peak increases with decreasing terrain complexity. This would indicate that the amount of look-ahead fixations compared to the amount of foothold-finding fixations increases as terrain complexity decreases, since information

about the ground is less important. Walkers are then more free to acquire information about the upcoming path.

2.7.5 Stability of VOR during locomotion with ground fixation

Because of this low slippage, the remainder of our analysis which focus on retinal motion signals will rely on a method that trades off between reliance on this low slippage assumption, and eye tracker noise. Specifically, the method for tracking initial fixation locations described in is used and the estimated gaze locations during fixations are replaced with the initial fixation location. This effectively idealizes fixations, pinning them at the initial fixation location for each detected fixation. This enforces 0 motion at the fovea, since the required eye movement perfectly counter rotates to stabilize the initially fixated location. The reported average error for Pupil Labs mobile eye tracker is 1 degree, although this may in most cases correspond to bias as opposed to variance. However when considering the worst case frame by frame scenario, this can result in large amounts of motion at the fovea. Additionally, the sampling rate of the eye tracker (30Hz) is low in relation to the timescales at which saccades begin and end. Thus we rely on this pinned fixation method for gaze location estimates over the course of fixations.

Stability of VOR needs to be further investigated with specific experimental setup although it does have interesting implications regarding the nature of most visual motion inputs (whose patterns arise from this stable fixation + head movement relative to fixated location). This constraint on motion inputs that results from fixating stable objects is important to consider when interpreting the results of experiments on visual motion processing systems.

2.7.6 Motion resulting from terrain complexity

The nature of the relationship between the flat ground response deviation for each terrain and median gaze angle is still unclear. Previous studies have observed higher variability of vertical gaze angle for irregular terrains [88], which was also observed in our results. Higher deviation terrain (more uncertainty) could be inducing a change in gaze allocation strategy, although what kind of benefit this might confer is not immediately obvious. For flat ground versus ground that has some obstacles, the need for visual information induces more ground fixations, however it is unclear why there is a gradient of effects for more variable terrain. This could be related an interaction between the spatial resolution of information needed given different terrain types, and the amount of certainty needed to maintain efficient gait. Understanding this relationship more precisely would require causal manipulation.

What is also still unclear is whether this change in strategy is optimal, and if so the sense in which it is optimal. What benefit does shifting gaze closer to the body confer in terms of visuomotor control? Is it related in any way to uncertainty? The deviation from flat ground metric does capture something like degree of uncertainty when no visual information is available, and so perhaps the level of uncertainty about the terrain is what drives differences in gaze allocation[89], [90], [91]. The precise relationship between terrain structure, uncertainty, and visuomotor strategy still needs to be further investigated.

2.8 Appendix

2.8.1 Within fixation initial target deviation

Sequential fixation frames are treated as single fixations, where an initial fixation target can be calculated. This is simply the gaze location in the camera image at the first fixation frame. Using optic flow vectors computed by Deepflow ([71]), this initial fixation location is tracked for the duration of the fixation. This is done by indexing the optic flow vector field at the initial fixation location, measuring it's displacement across the first frame pair, computing its new location in the next frame, and then measuring the flow vector at this new location. This is repeated for each frame in the fixation. The resulting trajectory is that of the initial fixation location in camera image space over the duration of the fixation. For each frame of the fixation, the actual gaze location is then compared to the current location of the initial fixation location (with the first frame excluded since this is how the initial fixation is defined). The locations are converted into 3D vectors as described in 2.4.1, and the median angular distance between gaze location and initial fixation location is computed for each fixation.

The vestibular-ocular reflex (VOR) was shown to be stable over the course of individual fixations. Data from this analysis can be seen in Figure 2.12. The median of the distribution of median initial fixation position deviations was 0.826 degrees. The 75^{th} percentile is at 2.8, while the 95^{th} percentile is at 14.97 degrees. It is likely that the long tail of the distribution results from errors in specifying the fixations rather than failure of stabilization.

Despite the frequent gaze shifts resulting in short fixations at rapidly varying locations, subjects were able to stabilize the fixated ground locations relatively well. For 75% of all fixations, the initially fixated location remained

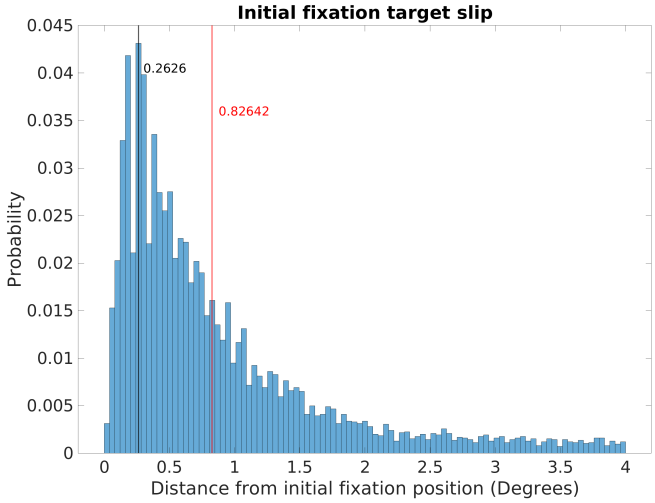


Figure 2.12: Histogram of deviation from initial fixation location over the course of fixation. Initial fixation location is computed and tracked over the course of each fixation, and compared to current fixation for duration of each fixation. Median deviation value is calculated for each fixation. The histogram captures the extent of variability of initially fixated locations relative to the fovea over the course of fixations, with most initially fixated locations never deviating more than 2 degrees of visual angle within the fixation.

within the 2.8 degrees of eccentricity, staying within the fovea. Higher acuity visual information is likely necessary for foothold selection, and the efficacy of stabilization that occurs means that it is often available. It is worth noting possible sources of noise including small saccades that are grouped into fixations resulting in higher displacement of the initial fixation location, as well as eye tracker noise. What is also not clear due to hardware limitations (image based slip calculations must be done at 30Hz, resulting in noise) is retinal slip statistics in terms of visual motion speed at fovea. A different experimental setup is necessary for this kind of measurement, however our measured initial fixation location deviation statistics and the walkers' ability to perform this task effectively suggest that slippage is modest. Treadmill studies have found

long periods (about 1.5s) of image stability (defined as less than 4 degrees per second of slip) during slow walking, with this being decreased to 213ms for faster walking [92].

Manual reintroduction of retinal slip (which our measurements suggest arise from a gain in the VOR of ¡1) simply results in added downward motion to the entire visual field, whose magnitude is equivalent to the slip. This also has the effect of slightly shifting structure in the motion pattern upwards, by however far from the fovea the eccentric location with the same speed as the slip is. When considering the average signal this shifts the zero point upwards to 4 degrees of eccentricity for 4 deg/s of retinal slip. The other structural features of the signal are conserved (radially asymmetric eccentricity speed gradient, variation with gaze angle).

2.8.2 Comparison of speed vs eccentricity relation for Deepflow computed, and Meshroom computed flow signals

To confirm the validity of using both methods for different analyses, results from the two were directly compared. Shown in Figures 2.13 and 2.15 are corresponding speed distributions as a function of retinal eccentricity. Here motion vectors within eccentricity bands centered at the indicated values are pooled and their histograms are plotted. For Figure 2.15, the nearest distribution for each of the Deepflow and Meshroom distributions is its equivalent. They are well matched, except for at higher eccentricities where they diverge slightly. This is likely a result of how the Meshroom based flow fields were calculated. The Meshroom method samples more at higher eccentricties, since during the depth image based motion vector computation, where there are missing values the depth that corresponds to a ground plane is imputed. This leads to more sampling of eccentric values than in the Deepflow based method,

where values are ignored. Whether this more closely approximates the retinal motion signal depends on whether this oversampling of more eccentric, but more flat ground locations leads to better approximation of the actual distribution of depths, or if ignoring these eccentric locations without imputing flat ground depth values does this better.

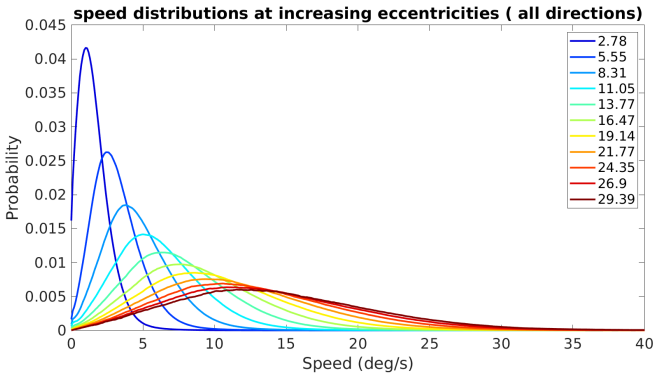


Figure 2.13: Distributions of retinal motion speeds as a function of eccentricity

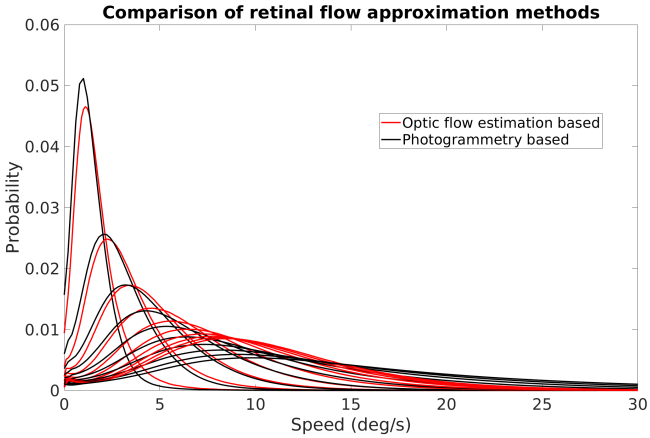


Figure 2.14: Distributions of retinal motion speeds as a function of eccentricity computed using either optic flow estimation based or photogrammetry based retinal motion approximation.

2.8.3 Comparison of average speeds across terrains

Here a similar method to that described in 2.6.3 was used to sample retinal motion values within each terrain type (where gaze distributions are matched). However the mean speed was computed instead of the flat ground deviation metric. Here it was difficult to interpret the speed differences, and when averaging speeds any differences between terrains are not likely to show. This is because speeds could be higher or lower than what might expect for flat ground when there is structured terrain. The reason for this is that one could be perched on a high rock, where relative to that perch most of the terrain ahead is actually lower than in a flat ground scenario (resulting in lower speeds). Conversely standing in a trough would result in higher than flat ground speeds. Thus the deviation from flat ground metric was more informative about between terrain differences.

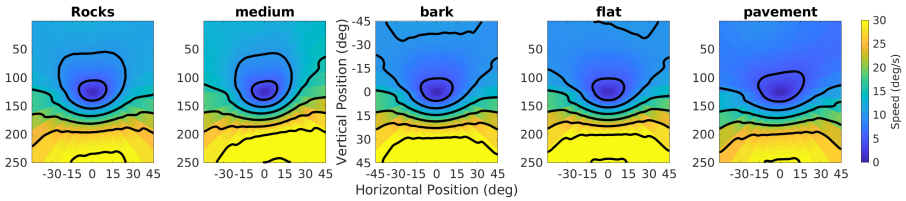


Figure 2.15: Average speed of retinal motion signal as a function of retinal position. Horizontal and vertical gazes were matched across terrains in order to control for their affects when sampling from different terrains.

2.8.4 Effects of saccades on average motion signal

Here we simply include saccade frames in our analysis of average motion speeds. In Figure 2.16 we compute motion statistics for only saccade frames. This results in high speeds throughout the visual field. In Figure 2.17 we compute average speed for all frames, both fixation and saccade. The result is a

distribution that resembles the original fixation only mean speed distribution, but with values shifted upwards as a consequence of averaging in the high speeds from saccades.

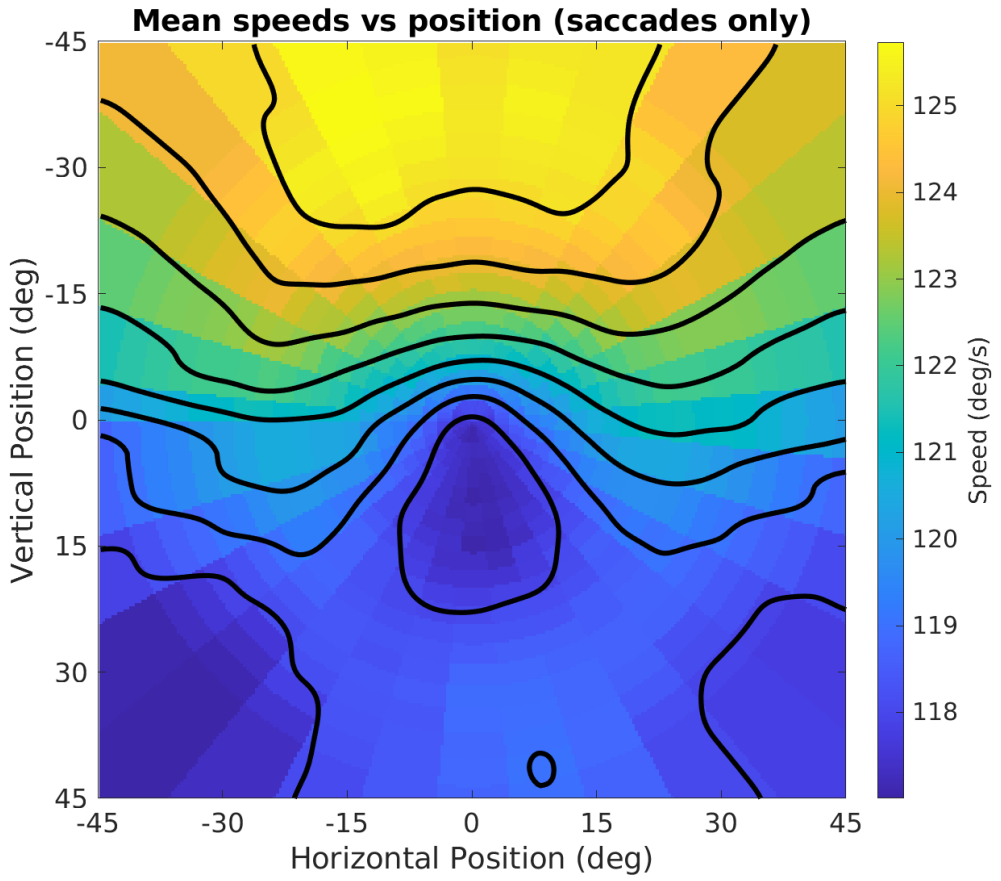


Figure 2.16: Average speed of retinal motion signal as a function of retinal position (saccades only). Retinal motion was computed only for saccade frames

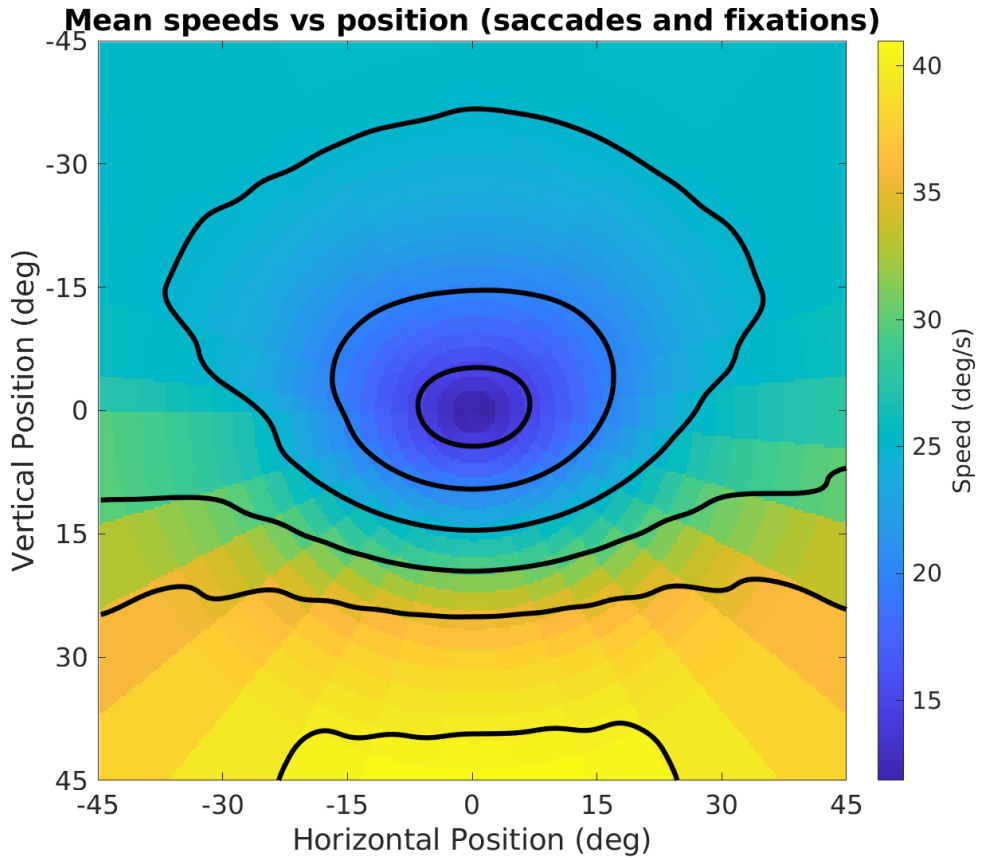


Figure 2.17: Average speed of retinal motion signal as a function of retinal position, with saccade frames included in calculations. Speed is colormapped (blue = slow, yellow = fast). Average is computed across all subjects, and terrain types. Speed is computed in degrees of visual angle per second

Chapter 3

Gaze prediction

3.1 Introduction

The next point of interest in this dataset was gaze allocation. We wanted to understand the extent to which we can predict it based on behavioral and environment variables. Gaze allocation and eye movements strongly shaped the motion input, but in addition the subjects are presumably fixating certain locations for a reason. Therefore determining what drives gaze allocation is important for understanding the underlying computations taking place as this complex behavior unfolds

First we were interested in body position measured by the motion capture suit, as well as foothold locations. However simply using the known future foothold locations would obviously suffice for predicting gaze to a degree given the clustering of subject gaze on ground around future foothold locations [64], [93]. Here we decided to test whether subjects were looking at desired future foot locations, based on their biomechanics (where they would like to step given their current walking pattern). We used a simple method for calculating future preferred foothold locations in order to test this, although found that

these locations were ineffective for predicting gaze locations.

In addition we wanted to understand the extent to which visual features in the environment were predictive of gaze locations. If vision is being used to actively search for viable foothold locations as opposed to confirming the desirability of preferred future locations (which is the case given our previous result), then there should be some kind of visual signature that is predictive of gaze allocation, one which could be framed as attracting gaze. Here we used convolutional neural networks in order to build a model capable of predicting gaze locations based on input images. The network did so above chance, although the exact interpretation of this result is uncertain.

3.2 Body position explains some variance in gaze direction

3.2.1 Methods

This analysis first required computing predictor variables for a linear regression scheme. Future preferred foothold locations were computed in order to be used as a model of gaze. These preferred foothold locations were computed by extrapolating subject gait by computing an average right and left step. These average right and left steps were computed by first resampling body positions over time for each left foot to right foot step, and each right foot to left foot step. These resampled sequences of body positions during right steps and left steps were then averaged across all right or left steps, yielding an average right step and average left step. Then, at each recorded step (as they appear in the original recording), the current "body orientation" was computed, by calculating the change in position of the center of mass of the

subject between the last step and the current step location. The average left and right steps were then aligned such that the "body orientations" of the average steps matched the current step. Then the average left and right steps were cumulatively added onto the original step location, yielding a sequence of preferred future foothold locations. This was repeated for the next 5 preferred foothold locations, yielding the next 5 preferred step locations relative to each step in the recording.

The other predictor variables for this analysis were simply the center of mass relative body positions of the subject measured by the motion capture suit at each frame of the recording. These were computed by taking each body position and subtracting the center of mass position at each frame.

For the first analysis, the computed next 5 preferred step locations were used to predict gaze locations. First the locations themselves were tested as predictions for gaze location, where for each frame of the recording the closest foothold location to the current gaze location was treated as the prediction. The locations were transformed to spherical coordinates, with a vertical angle (elevation relative to gravity) and horizontal angle (about the gravity axis), and R^2 was computed for each angle.

For the next analysis, linear weights were fitted to each of the 5 preferred step locations in order to predict the gaze locations. This can be seen schematized in Figure 3.1, where the next analysis is also described. This next analysis used subject center of mass (COM) relative body positions as predictors with linear weights.

R^2 was used as a performance metric for each of the three cases.

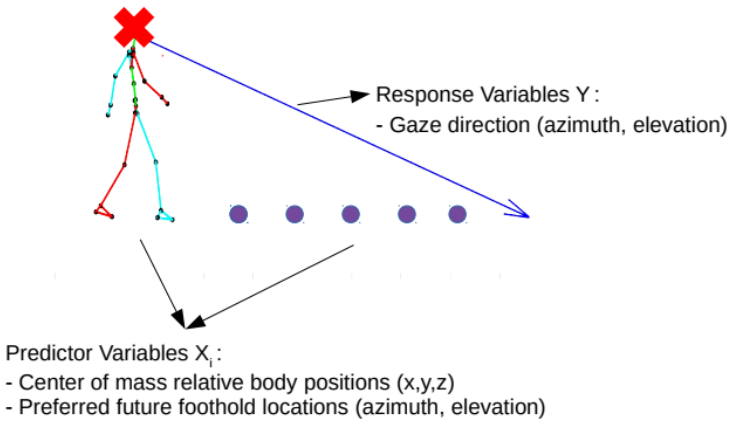


Figure 3.1: Schematic of linear regression problem. Subject body positions and future foothold locations are used as predictor variables. The head position is excluded due to central tendency of gaze, where head orientation is highly correlated with gaze direction. The response variable is gaze direction, represented in spherical coordinates.

3.2.2 Results

A consistent pattern emerged for both horizontal gaze angle (Figure 3.2) and vertical gaze angle (Figure 3.3). "Raw" which refers to the best performing future preferred foothold location at each frame had negative R^2, meaning it is outperformed by using the mean gaze angles rather than the foot locations. The poor performance of preferred future foothold locations is further exemplified by low R^2 when fitting linear weights to the foothold locations.

In contrast, the "B" model, or COM relative Body positions used as inputs in linear regression performed well for both horizontal ($R^2 = 0.6$) and vertical ($R^2 = 0.48$).

3.2.3 Discussion

The low performance of both future preferred foothold locations as gaze location predictions, and preferred foothold locations as inputs in a linear regres-

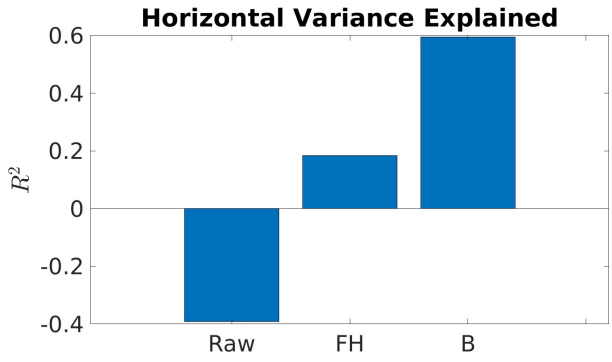


Figure 3.2: Variance in horizontal gaze angle explainable by different models. 'Raw' corresponds to using the best performing calculated future preferred foothold location as the prediction for gaze location. 'FH' refers to using the foothold locations, with linear weights fitted to them. 'B' refers to using COM relative body positions as inputs to a linear regression.

sion model suggests that gaze is actually not being deployed to check future foothold locations for viability. This conclusion however does rely on the assumption that the preferred foothold locations that we computed accurately portray where subjects would step given completely flat ground. This is a somewhat reasonable assumption, however the possibility that we are not correctly capturing where subjects would prefer to step given their current gait can not be ruled out.

The higher performance of the COM relative body position model is somewhat mysterious. Higher horizontal performance may be owed to the tendency of subjects to direct gaze prior to making a turn, as observed in Bernadin et. al. (2012) [94]. However the source of the performance on the vertical gaze angle prediction is less clear.

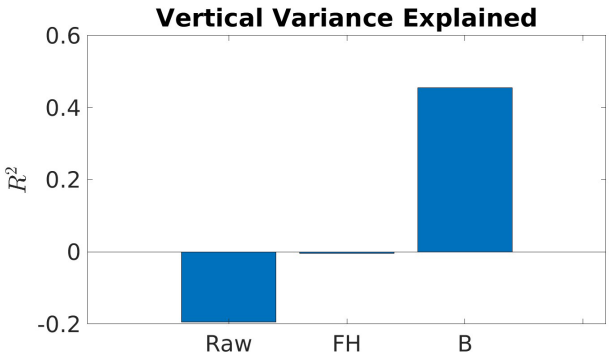


Figure 3.3: Variance in vertical gaze angle explainable by different models. 'Raw' corresponds to using the best performing calculated future preferred foothold location as the prediction for gaze location. 'FH' refers to using the foothold locations, with linear weights fitted to them. 'B' refers to using COM relative body positions as inputs to a linear regression.

3.3 CNN suggests image features are predictive of gaze locations

3.3.1 Methods

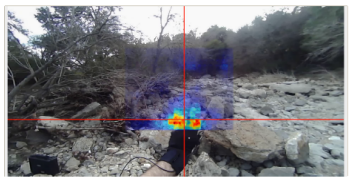
For this analysis, a convolutional neural network (CNN) was used to predict gaze locations in the eye tracker's outward facing camera image space. The input was the subject first person perspective image itself, and the target output was a gaussian distribution centered on the recorded gaze location in image space.

The images were first pre-processed using a random cropping procedure. The image was randomly cropped to a smaller image with half the width and height (from 1080x1920 pixels to 540x960), and the crops were selected randomly such that they contained the actual gaze location. The gaze location within the smaller cropped images was uniformly distributed. This was done in

order to avoid the tendency of CNNs to overfit by predicting central locations for gaze data. This arises due to center bias of gaze, which results from the eyes being preferentially situated centrally in their orbit. This central bias means that the gaze locations are strongly biased towards the center of the image. By randomly cropping an image that contains the gaze location, and using this smaller cropped image as the input, the CNN will not overfit to a particular location of gaze since gaze is uniformly distributed.

The CNN architecture was a convolutional deconvolutional architecture, with 3 convolutional layers, followed by 3 transposed convolutional layers. This arhcitecture allows generation of a probabilty map of gaze locations. The target probability map is created by centering a gaussian distribution with standard deviation equal to the assumed error of the eye tracker (approximately 1 degree of visual angle). The CNN is then trained using KL-divergence between the target distribution and its output as the loss function to produce the gaze probability map when using the cropped image as input.

As a baseline for comparison, we used the classic Itti-Koch saliency model for comparison. The performance metric used was area under the receiver operating characteristic curve (AUC). This provides a measurement of the accuracy of the predicted gaze locations generated by the model. It treats gaze prediction as a binary classification problem, where each pixel must be labelled as either a "gazed upon" pixel or not, and AUC measures the resulting performance on the problem when framed in this manner. AUC can be thought of as the average proportion of pixels that can be reliably labelled as not gaze pixels while still accurately labelling the gaze pixel. AUC was computed for both the Itti-Koch saliency model as a baseline, and our model.



Predicted versus actual gaze distribution

Performance

Prediction Method	AUC
Random Guess	0.502
Saliency-Based Model (Itti 1998)	0.479
Neural Network (Ours)	0.747

AUC: 0.5 – random guess; 1 - perfect

Figure 3.4: Single frame example of predicted gaze location (high probabilty shown in red, low probability shown in blue). Predicted region is a smaller random crop of the original image. The trained CNN performs well above chance with an AUC of 0.747, compared to baseline of 0.5

3.3.2 Results

Our model performed well above chance, with an AUC of 0.747, compared to the baseline of random guessing (average 0.5), as well as the Itti-Koch saliency model (AUC=0.479). A single frame example and table can be seen in Figure 3.5.

3.3.3 Discussion

The above chance performance can most likely be attributed to the ability of the CNN to rate the terrain regions of the image as higher probability than the surrounding scene. This is reasonable since the subject is likely restricting their gaze to the terrain portion of the scene, since that is where the task relevant information is (see Figure **??** for example). However this is not the level of visual feature we were interested in extracting, and so a deeper understanding of how gaze is affected by visual features will likely require a different approach.

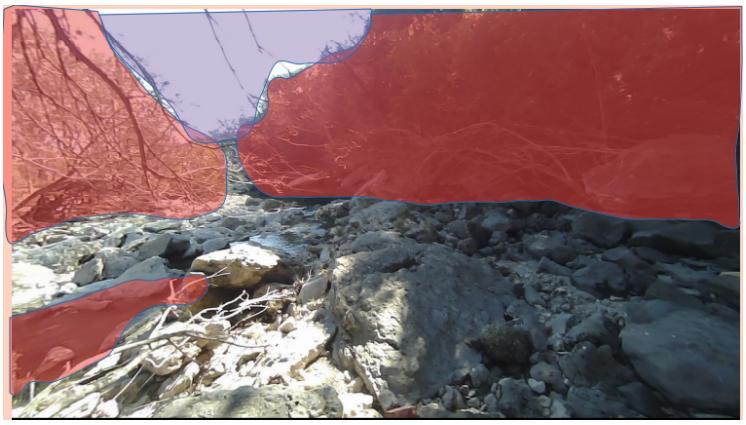


Figure 3.5: Single frame example of scene where CNN could achieve above chance performance with a relatively trivial strategy of identifying terrain vs surrounding foliage/above horizon regions of the image. Doing so would allow consistent labelling of not gazed pixels since subjects rarely direct gaze anywhere besides the terrain.

3.4 General Discussion

This investigation was mostly inconclusive regarding how walkers are directing gaze and using visual information in order to perform this task, however it did provide a few relevant insights. The body is predictive of gaze direction, and people tend to mostly direct gaze at the task relevant portions of the scene.

From here it seemed that a change in approach would be best, where we focused on elements of the task that we knew were of importance. Specifically we decided to focus on foothold locations. A big issue with focusing on gaze allocation is that it becomes difficult to attribute intention to specific eye movements and fixation locations during this behavior. However we assume there must be some relationship between foothold locations and gaze, as well as the environment and foothold locations. And so we shifted our focus to foothold locations for the remainder of this investigation. However, first I will discuss our novel method for getting the necessary data in the next chapter.

Chapter 4

Reconstruction of terrain and head trajectory

As described in earlier sections, photogrammetry is a computer vision method that allows simultaneous estimation of camera poses and environmental structure from sequences of camera images. Photogrammetry has been an important tool for further augmenting the dataset, allowing new analyses for the retinal motion statistics portion of the thesis, as well enabling all of the analyses in the next chapter. In this chapter I will provide a more detailed description of how photogrammetry was used and incorporated into the dataset.

4.1 Detailed overview of Meshroom pipeline

In this section I will give a brief overview of how Meshroom works, since it is a collection of many different computer vision and 3D data processing techniques from a variety of sources.

Natural feature extraction: This step involves extracting image features from each image in the input sequence. The features are selected such that

their descriptors are minimally variant to viewpoint changes, which is crucial in this application since they need to be reliably detected across different viewpoints in order for the reconstruction to be possible. Here the scale invariant feature transform (SIFT) [95] is used in order to extract such features from images. The invariant property of these descriptors is crucial for the next steps.

Image matching: Here after feature descriptors are computed for each image, these descriptors are stored into a 'vocabulary tree' [96]. This allows efficient lookup of an image's descriptors, which is used to determine the amount of shared features between images. This is used to determine matching images from the set of all images. Each image can have multiple other image matches since the input images are from first person video traveling through an environment and hence have shared features across sequential frames of the video. Matched images are then used in the next step.

Features matching: For each pair of matched images, corresponding image feature locations are found for all feature descriptors that appear in both images. This finds the same location in world space in the image space of each matched image. Computing each of the locations allows one to compute the displacement of the feature descriptors across the matched images, which contains information about how the camera moved through the environment, as well as information about the structure of the environment. This information (camera and feature movement) is extracted in the next step.

Structure from motion: With image pairs selected as well as corresponding feature locations computed, scene geometry and camera position and orientation for each image can be estimated. Meshroom uses an iterative process for reconstruction, starting with a single image pair's features, and iteratively adding information from new views that match either of the initial image pairs.

After one pass of adding in new views, the reconstruction and views are refined using PnP [97] and RANSAC [98]. Then a Bundle Adjustment [99] is performed, removing observations with high reprojection errors. The resulting camera pose estimates were a crucial component in aligning Meshroom's output to our existing data.

Depth maps estimation: Structure from motion yields a sparse reconstruction of the environment, represented by 3D points in space. The depth map step tries to use this sparse reconstruction in order to estimate a depth for each pixel of the input images, using the estimated camera pose for each image and the 3D points in view of that camera pose. This is done by computing multiple depth candidates per pixel, using multiple nearby camera views, then consolidating and filtering. An additional filtering step is applied to enforce consistent depth estimates between camera views.

Meshing: After each camera view has an estimated depth map, all depth maps are fused into a global octree. Then 3D delaunay tetrahedralization [100] is performed, followed by a voting procedure [101], [102]. The Graph Cut Max-Flow [103] is then used, followed by Laplacian filtering to remove artifacts. This process yields a 3D triangle mesh that can be further simplified to reduce excessive vertices. These triangle mesh outputs were essential for our analyses in Chapter 5.

Texturing: The final step is texturing of the computed 3D triangle mesh. This is done by iterating over each triangle and using the visibility information of each vertex to retrieve texture information from the original camera images. The pixel values are averaged, averaging more views in low frequencies than higher frequencies [104]. This yields a fully textured 3D triangle mesh, where each vertex has a corresponding color. The texture aspect of the meshes was not strictly important for our analysis, but makes visualization and hence

intuition building easier when once can more easily recognize terrain features.

4.2 Terrain reconstruction

Output formats: Each step of Meshroom's pipeline generates intermediate processed data, which contain different useful pieces of information. For the terrain reconstruction, we leveraged the 3D triangle mesh output, which contained information about the structure of the environment. The 3D triangle mesh representation was particularly useful for the depth map estimation method used in our analysis. The triangle representation was also resampled as colored 3D point clouds using CloudCompare [105]. Here rather than vertices and their corresponding triangle edges, the 3D triangle mesh was resampled to 3D points. This means that there are arbitrary numbers of points representing the boundaries and surface of each triangle rather than 3 vertices and the connecting edges. This resampling of the mesh was crucial for certain parts of the analysis.

Some details on the data structures: The 3D triangle mesh representation as the name suggests, is composed of 3D triangles. The vertices of the triangles are stored as a list of points (N x 3), with each point having a unique index to identify it. The triangles are represented by a list of coordinates (M x 3), where each row has 3 indices indicating which of the points compose that triangle. So there are a total of M triangles, represented by different connections of N points. Each of the N points can be a member of multiple triangles, depending on the structure of the terrain wherever the point is located.

Use of triangle representation: The triangle representation was used for two important steps in our analysis, as well as for visualization. The first was computing local terrain slant. The terrain reconstruction is aligned such that

the vertical coordinate corresponds to gravity. As a result the triangle plane normal vectors' angles relative to straight upwards indicate the surface slants of the triangles. This slant value was used in analysis that involved filtering parts of the terrain that are walk-on-able by human subjects. There are many methods for approximating slant values of meshes but in this case, computing a value for each triangle, and then averaging over a 3D volume the size of a human foot allowed a satisfactory approximation of local slant.

The second use was for computing depth maps (described in detail in the next section). Having the triangle mesh representation allowed for easy use of Blender's z-buffer method in order to calculate depth maps. These depth maps were then used to compute retinal motion inputs using known translations and rotations of the eye (see [77] for generalized form).

The textured triangle mesh representation made visualization much more intuitive. Using MATLAB's trisurf function, code was developed to visualize subject body and eye movements relative to the fully textured terrain. The video outputs of this visualization process were important for deriving insights about the ongoing behavior by watching it unfold, both in real time and frame by frame when necessary.

Use of point cloud representation: The point cloud representation was important for foothold localization and scaling of the motion capture data. Specifically, after alignment (described in the next section), the foothold locations on the terrain point cloud were found using a nearest neighbor approach, allowing for calculation of the best scale factor to align the different data. This approach would be difficult with a triangle representation since one needs to search over the entirety of each triangle rather that just the vertices when determining intersections of the walker's foot and the terrain.

4.3 Motion capture alignment

Head pinning and Meshroom camera trajectory: One of the important outputs from Meshroom that is used is the camera trajectory. Meshroom provides for each frame of the original input image sequence, an estimate of the camera location and orientation in the coordinate frame of the reconstructed terrain. This camera trajectory is used to align the coordinate systems of the motion capture and eye tracking data to that of Meshroom. This is done by first pinning the head marker's location to Meshroom's camera position estimate at each frame.

Rotation optimization: After translation, the rotation between the two coordinate systems must be computed. A single rotation that best aligns the world camera orientation in the motion capture data's coordinate system to the orientation of the camera in Meshroom's coordinate system is computed. This computation is done using MATLAB's fminsearch function, where the euclidean distance between the two orientation matrices at each frame (Meshroom's camera, and the camera basis vectors in the motion capture space) is used as an objective function, and three euler angles are used as the input. The optimization seeks to find the three euler angles that minimize this error when applied as a rotation.

Scaling: Once the appropriate translation and rotation has been applied, the motion capture data must be scaled to match the arbitrary scale of the Meshroom reconstruction. This is done by taking advantage of the subjects foot plant locations. Since there is always at least one foot in contact with the ground since the subject is not running, the foot locations can be used to compute the correct scale factor. This is done by sweeping over a range of scale factors to apply to the motion capture data, and for each footplant

frame previously determined by analyzing the motion capture data, and for each scale factor, computing the nearest neighbor of the planted foot to the 3D points in point cloud representation of the terrain. The median distance to each of these points is then calculated for each scale factor, and the scale factor resulting in the smallest median distance is used and applied to the entire recording.

4.4 Drift correction

IMU related drift: The absolute position data from the motion capture suit is not measured but rather inferred by double integrating acceleration signals from the accelerometers. The relative positions of the joints are inferred based on assumptions about a skeletal structure and measurements of joint orientations, however the estimate of absolute position of the entire skeleton in world coordinates relies on double integration. This means that the absolute position estimate accumulates noise over time, causing the estimate of the skeleton's location in space to drift over the course of the recording. This is not a significant issue over short time periods, however it becomes a concern when computing the locations of future footholds relative to the current body location, which would accumulate noise with increasing distance.

Image based reference frame drift correction: The position estimates from Meshroom's reconstruction process can be used in order to correct for this drift. Since the coordinate system of these position estimates is the same as the reconstructed terrain, the terrain serves as a stable reference point. While there is a small amount of drift associated with image based reconstruction (since the terrain is iteratively added to, leading to small accumulated position errors), it is more stable than the position estimates from the motion capture

suit.

4.5 Between subject alignment

Between terrain keypoint identification and selection: In order to visualize between subject variation of path selection, all subject data was aligned to the same coordinate system. Because each recording was used to reconstruct its own terrain, the different recordings had their own coordinate systems. To align the coordinate systems of different subjects, CloudCompare [105] was used in order to inspect each terrain reconstruction, and manually extract the location of 5 reliably detectable locations on the terrain, that were visible across all reconstructions of the terrain (recognizable rocks, markings, etc.). This process was repeated for all reconstructions.

Similarity transform application: The first subject's recording was used as the reference coordinate system, and for each other subject the similarity transform (simultaneous translation, rotation, and scaling) that aligned each of the 5 corresponding key points was computed, and applied to the entire non reference terrain. This aligns the non reference and reference terrains in a coarse manner.

ICP: After the similarity transform was applied, two iterations of the Iterative Closest Point (ICP) [106] algorithm were applied. This step also yields a similarity transform, which constitutes a small adjustment compared to the initially calculated transform from the 5 corresponding keypoints, adjusting the orientation, location, and scale of the non reference terrain only slightly. However error (measured by distance to the closest point in the reference mesh) still decreases. The coarse and fine transforms are then combined, resulting in a single transform that aligns each of the non reference terrains to the refer-

ence terrain. This same transform is then applied to aligned motion capture data, putting the step locations and head trajectories of all subjects into a single coordinate frame. We computed our own error metric rather than using overall MSE as is typical, since we are more concerned with reconstruction and alignment accuracy at foothold locations themselves. So our metric (described further in Section 4.7) was MSE but for foothold locations specifically. This error was quite low, with over 75% of errors being less than 0.2 foot lengths, and all errors being less than 1 foot length.

4.6 Geometry based motion estimation

Blender depth map estimation: 3D Terrain data also allows approximation of retinal motion inputs based on terrain geometry rather than image based motion estimates. This is the process used to compare flat ground simulated and actual motion in previous sections. In order to do this, Blender [76] was used with the terrain data and aligned gaze data. Using a Python script, depth images were captured using Blender's z-buffer method along with a virtual camera located at each frame's recorded head position and pointed in the gaze direction. The resulting depth image is then stored, and this process is repeated over each frame of the recording.

Use depth map to compute motion based on geometry: Once the sequence of eye relative depth images are captured, they can be used to approximate motion based on the recorded eye movement from the current frame to the next frame, and the current frame's depth data. The eye's translation and rotation through space, as well as the depth values at each retinal location can be used to calculate the visual motion of each point in space relative to the eye. This method allows for direct comparison of actual motion from a

particular terrain to a simulated flat ground instance. This is important since it isolates the effects of the terrain.

4.7 Error measurement

Manual foothold clicking results: The most effective way to validate our photogrammetry derived data, including the resulting improved foothold localization is still not clear. The chosen method was to assume the reconstructions camera orientation estimate was reliable, and to manually annotate the subjects' foot locations in the scene camera video when visible. These manually annotated locations were then projected out from the image plane using the estimated camera orientation, and the closest location on the terrain mesh was computed for this projection. These 'ground truth' foot locations were then compared to those computed using the previously described alignment of the motion capture data. The resulting error distributions can be seen in Figure 5.17, B. An example frame where a foothold location is manually annotated, and the corresponding location on the mesh is computed and compared is shown in Figure 4.1.

Between subject mesh same point comparisons: Another important question is that of inter subject reliability for terrain reconstruction. For this measurement the different terrains were first aligned using the coarse to fine alignment process described earlier in this chapter. Then each subjects foothold locations were estimated using the scaling procedure, however each terrain reconstruction instance (from other recordings) was iterated over for a separate foot location calculation. In other words for a given recording, a set of foothold location estimates using each of the other recordings' terrain reconstruction was computed. Each subjects aligned terrain data was used to

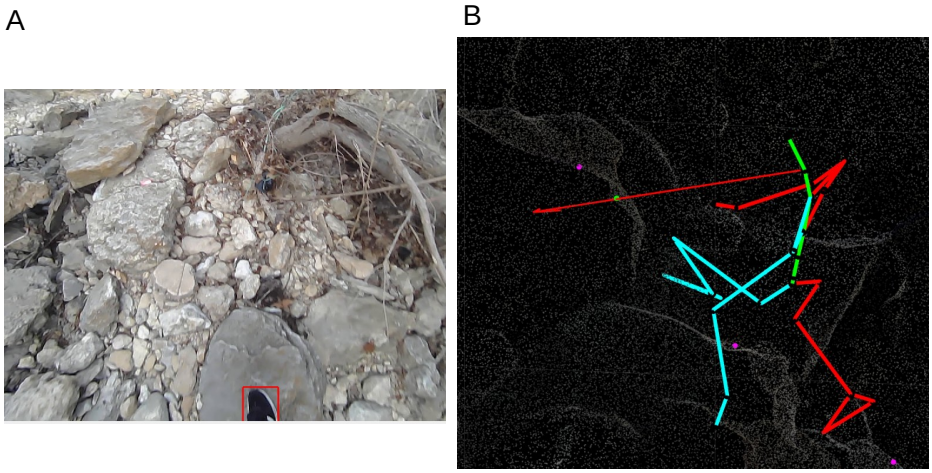


Figure 4.1: Example of manual foothold location annotation and comparison. In A, frames of recordings where subject foothold locations are in view (foot is in the frame when the subject plants it on the ground) are identified and the foothold location is manually annotated. Then in B using the camera pose estimate from Meshroom, the 3D vector corresponding to this foothold location is computed and the closest intersecting point on the mesh is determined (shown as a green dot). Then the corresponding foothold location from the aligned motion capture data is used to compute error for this example. This is then repeated for each instance where the foothold location was in the image frame.

estimate each subjects foothold locations, exhausting all combinations. The result was a distribution of estimated foothold locations for each actual subject foothold location, computed using different terrains. This distribution of estimated locations was used to compute relative error, which can also be seen in Figure 5.17, A.

4.8 Potential new approach for retinal input approximation

It is worth mentioning a work in progress from Panfili et al. [107], which involves a real time retinal simulation for virtual environments. The goal is to leverage ray tracing in order to simulate and record retinal motion inputs resulting from eye translations and rotations in virtual environments. This could eventually be used in conjunction with reconstructed natural environments in order to record retinal motion efficiently, and in real time. The real time application is particularly important for real time manipulations in experimental settings, contingent on properties of the motion input. Traditional computer vision methods for approximating retinal optic flow would be limited by processing time for this kind of application.

Chapter 5

Foothold selection

5.1 Introduction

Natural visually guided behaviors can be characterized as a sequence of complex sensorimotor decisions [9], [108]. In this context, even simple behaviors such as locomotion require consideration of sensory uncertainty, together with the momentary rewards and costs of direction choices and foothold selection. Locomotion on flat ground requires very little visual feedback, and can be accomplished with minimal cognitive control [109], but locomotion over complex terrain requires the coordination between brainstem-mediated central pattern generators, and motor cortex-mediated modifications of leg and foot trajectories [37]. These modulatory signals in turn depend on evaluation of visual information about viable foothold locations and desirable paths.

Understanding how visual information is incorporated into locomotor decisions presents a challenge, since it is difficult to create experiments that fully capture the complexity of walking behavior as it unfolds in natural settings. Much of our current understanding of locomotion comes from work characterizing steady state walking on treadmills, where it has been shown

that human beings converge towards energetic optima. Human subjects adopt a preferred gait that constitutes an energetic minimum given their own biomechanics [110], [111], [112]. The parameters over which this optimization principle holds include walking speed, step frequency, step distance, and step width [113], [114], [115]. While energetic cost minimization has been well established for steady state walking on flat terrain, the same framework can also describe some of the effects of deviations from flat terrain on gait. For example, it has been shown that on sloped surfaces, an objective function that includes stability can help explain the reduced speeds observed relative to what would be optimal in terms of cost of transport [116].

There has been little work on how these optimization principles might play out in natural locomotion. Existing models of locomotion focus primarily on the optimization of the preferred gait cycle with respect to the walker's neural and biomechanical factors. However, locomotion over rough terrain must be optimized for both the biomechanics of the walker and the structure of the environment being traversed. In these environments, a walker must make a trade-off between the efficiency of the preferred gait cycle and the need to place the feet in stable locations to support continuous locomotion. Little is known about how vision is used to identify viable footholds, and how walkers use this information to alter the preferred gait cycle appropriately for the upcoming path. Previous studies tracking the eyes while walking outdoors have found alterations of gaze with the demands of the terrain [117], [118], [119], [119], but foot placement was not measured, so it was not possible to analyze the relation between gaze and foot placement. Recent work by Matthis et al (2018) that integrated gaze and body measurements in natural walking showed that walkers modulate gait speed in order to gather visual information necessary for selection of stable footholds as the terrain became more irregular. In addi-

tion, increasing time was spent looking at the ground close to the walker with increasing terrain complexity, and subjects spent most of the time looking 2 to 3 steps ahead in moderate and rough terrain. Subjects looked slightly further ahead in rough terrain, perhaps to allow for more path planning as terrain complexity increased. While in principle it appeared that subjects optimized both energetic costs and stability by regulating gait speed, understanding the visuo-motor control loop was limited by the lack of a quantitative representation of the terrain itself. Thus, while gaze and gait were tightly linked, it is not known what visual features subjects look for in the upcoming terrain in order to choose footholds and guide body direction towards the goal. The aim of the present study, therefore, was to incorporate a representation of the terrain, linked to gaze and gait data to shed light on how subjects use visual information about the structure of the environment to choose paths in natural rugged terrain.

In Matthis et al's 2018 study, it was necessary to assume a flat ground plane and both gaze location and foot placement were projected onto this plane. In the present study, however, we took advantage of a recently developed photogrammetry algorithm that uses a sequence of camera views to reconstruct the 3 dimensional terrain structure, along with a representation of the 6 DOF camera path. Because the camera was mounted on the subject's head, we were also able to align the reference frame of the terrain with that of the walker. This allowed accurate estimates of gaze and footholds on the ground surface, and also allowed us to relate the choice of footholds to the terrain structure. The departure from reliance on the flat ground assumption and the ability to relate geometric features of the terrain to walker behavior is a key component of this work. We first demonstrated that there were in fact regularities in the paths chosen by subjects when walking over the same

terrain on a different occasion, and also that there were similarities between subjects in the chosen paths. Thus paths were not completely random, and must reflect some optimization principles. Our next step was to measure the smoothness of short segments of the chosen paths relative to neighboring regions. This was motivated by the importance of energetic costs demonstrated in previous research, and by the fact that stepping up and down on large rocks is energetically costly. We found that subjects choose paths where the average height change in a short segment is less than neighboring possible paths. We also found evidence that average height change is evaluated over a set of several future steps, indicating planning of step sequences. Deviating from a straight path also incurs energetic cost, and this cost increases when subjects avoid straight paths with big height changes. We found that subjects deviate more from straight paths as the average height change of those paths increases. Finally, we trained a neural network to recognize viable paths using depth images and chosen path segments, from the viewpoint of the walker. The network was able to learn to predict paths that subjects would take. Thus we found that subjects plan multi-step paths on the basis of visual information about height irregularities, reflecting the role of energetic costs even in rough terrain. Although only a portion of the variance was explained by this factor, there is sufficient regularity to reveal stable factors underlying foothold choices.

5.2 Methods

Participants: The data used in this study was collected by the authors in two separate studies, performed in similar conditions and using the same apparatus. The first group of participants (n=3) were recruited with informed consent in accordance with the Institutional Review Board at the University

of Texas at Austin. The second group of participants (n=8) were recruited with informed consent in accordance with the Institutional Review Board at The University of California Berkeley.

Equipment: Eye and body movements of both groups of participants were recorded using a Pupil Labs mobile eye tracker and the Motion Shadow full body motion capture system. The eye tracker has two eye facing cameras, and one world facing camera. The eye cameras recorded from each eye at 120Hz with 640x480 pixel resolution. The outward facing camera was mounted 3cm above the right eye, and recorded at 30Hz at 1920x1080 pixel resolution, with a 100 degree diagonal field of view. The motion capture suit featured 17 sensors (with 3-axis accelerometer, gyroscope, and magnetometers) whose readings were combined with software to estimate full body joint positions, as described in the Detailed Methods section and in Matthis et al (2021). The raw data was recorded at 100Hz, and was later processed with custom Matlab code (Mathworks, Natick, MA, USA).

Experimental Task: The task instructions were similar for the two groups, with only the terrain type varying slightly: In the Berkeley data set, participants were instructed to walk back and forth along a hiking trail that varied in terrain difficulty. This walk back and forth was then repeated. Terrain stretches were pre-designated as pavement, flat, medium, and rough, although only the rough terrain data was used in this study in order to best combine with the Austin data set. The rough terrain consisted of large rock obstacles with significant height deviations from purely flat terrain. In the Austin data set, participants were instructed to walk back and forth three times along a stretch of a dried out rocky creek bed, which consisted mostly of large rocks. This was the same terrain used in the Rough Terrain condition in Matthis et al. (2018). Since both terrains were rugged, it was necessary for

subjects to use visual information in order to localize and guide foot placement (see Matthis et al, 2018).

Calibration and post-processing: At the beginning of each recording, participants were instructed to stand on a calibration mat 1.5 meters from a calibration point marked on the mat in front of them. This distance was chosen based on the most frequent gaze distance in front of the body during natural walking in these terrains. They were instructed to fixate the calibration point while rotating their head along each of the 4 cardinal directions, and 4 more in the diagonal directions. This portion of the recording is then used to find the single optimal rotation between the eye tracker's coordinate system and the motion capture systems recording system such that the eye direction vector's intersection with the mat is closest to the calibration point. This rotation is then applied to each frame of the eye data. The resulting data streams are now aligned in both space and time. (See also Matthis et al, 2018.) Following the data collection, recordings from the eye tracker and motion capture system were aligned in space and time. Temporal alignment used the timestamps recorded from each device on the recording computer worn on a backpack by the subject. The motion capture systems data stream was upsampled (using linear interpolation) to 120Hz to match the frame rate of the eye tracker. The eye ball centers relative to the head center (measured by the motion capture system) are then approximated, and the eye direction vectors are centered at each respective eye.

Photogrammetry with Meshroom: In order to estimate both the environmental structure and the relative camera position from the head mounted video, we used Meshroom [75], which is a software package that combines multiple image processing and computer vision algorithms in order to reconstruct the environment from an image sequence. This allows a quantitative descrip-

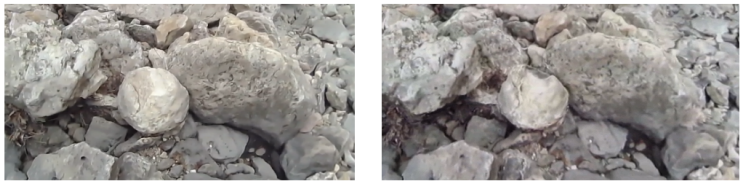


Figure 5.1: Rendered image of textured mesh from Meshroom (right) along side original RGB video frame (left) . Meshroom provides as output estimated camera positions and orientations for each video frame, relative to an estimated environmental structure represented as a textured 3D triangle mesh

tion of the 3D structure of the environment, as well as estimates of head position relative to the environment. First, features that are minimally variant with respect to viewpoint are extracted from each image. Images are then grouped and matched on the basis of these features, followed by matching of the features themselves between images. Feature matches from previous step are then used to infer rigid scene structure (3D points) and image pose (position and orientation) for each of the image pairs. We then aligned the head orientation and position measured by the IMU motion capture system with that of the camera orientation and position estimated by Meshroom (See Figure 5.2 for depiction). In previous work (Matthis et al 2018), estimates of future foot locations relative to current body location were subject to noise resulting from drift in the IMU signal. By pinning the IMU estimated head position to the Meshroom estimates, we were able to eliminate this drift by fixing the environment representation relative to the body.

In order to evaluate the accuracy of the 3D reconstruction we took advantage of the terrain meshes calculated from different traversals of the same terrain by an individual subject, and also by the different subjects. Thus for the Austin data set we had 12 traversals (out and back 3 times by 2 subjects.) Easily identifiable features in the environment (e.g. permanent marks

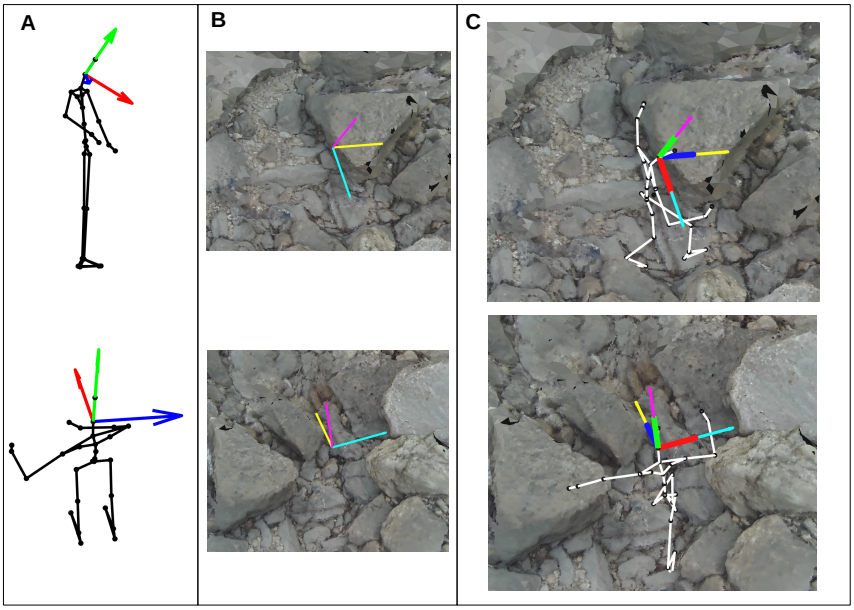


Figure 5.2: Alignment of motion capture data to Meshrooom coordinates. Motion capture coordinate system (A) is aligned with meshroom coordinate system (B) via a single rotation and translation that minimizes error between the mocaps camera axes and Meshroom's camera axes (C). The motion capture skeleton is then scaled such that the foot locations on known footfall frames distance to the closest point on the mesh is minimized at each footfall frame. This scale factor is then applied to the motion capture data at every frame.

on rocks) were used in order to align coordinate systems from each traversal. A set of corresponding points can be used in order to compute a similarity transform between points. Then the iterative closest point (ICP) method is used to align the corresponding point clouds at a finer scale by iteratively rotating and translating the point cloud such that each point moves closer to its nearest neighbor in the target point cloud. The resulting coordinate transformation is then applied to all recordings such that they are all in the same coordinate frame. There is high agreement between terrain reconstructions, with small

errors in foothold localization (see https://youtu.be/llulrzhIAVg for example subject traversal). More details available in Appendix. A visualization of aligned motion capture, eye tracking and terrain data is shown in the video at https://youtu.be/TzrA_iEtj1s. The heatmap overlayed on the terrain image shows gaze density, and future foothold locations are shown in magenta.

5.3 Between subject path similarity

The first issue we needed to address was whether there was a meaningful relationship between the environment and chosen paths. The level of agreement between subjects can shed some light on this question, since regularities in chosen paths indicate influence of the environment on path choice. Figure 5.4 shows the paths for the Austin and Berkeley data sets. In Figure 5.4, the two different subjects are shown in different colors for A and B and paths in both directions are included. One subject's outwards facing eye tracker camera was angled further from the ground, resulting in localization issues for Meshroom, and was thus not included in this analysis. For the Berkeley data, a similar issue arose, where for one subject the lighting for the day resulted in very low contrast in the video, resulting in noisy estimates from Meshroom. The remaining 7 subjects are shown in an excerpt showing one traversal of this terrain segment in C. For both terrains it can be seen that there is considerable regularity in the chosen paths, especially at particular points in the path, with divergence in other regions. The points of convergence suggest that the are indeed some terrain features that drive path selection. This is most notable for C, where the central region features the same path for all subjects. Inspection of the first person video data reveals a well trodden path with large boulders and fallen trees flanking it, resulting in convergence of paths for the different

subjects. There is more variability before and after this section of the path. This is also notable to a certain degree in A and B, where certain regions of the terrain have multiple paths that go through them, although there is much more divergence for this terrain compared to C. An example of convergent and divergent paths from subject perspective can be seen in Figure 5.3

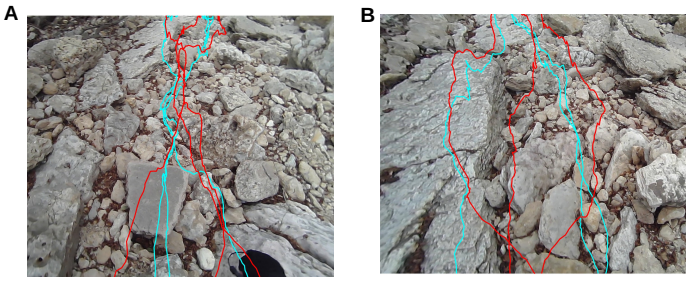


Figure 5.3: Examples of convergence (A) and divergence (B) of paths from subject perspective. Each color corresponds to repeated traversals by one subject.

5.4 Gaze allocation relative to chosen footholds

A second issue we need to address before we investigate the role of specific terrain features is the nature of the relationship between gaze and foot placement. This was the concern of our earlier investigation (Matthis et al 2018), which showed that fixations were clustered in the region 2-3 steps ahead of the walker's current foot plant. We took up this issue again with the improved estimates of gaze location and footplants using the 3D meshes from Photogrammetry. A detailed analysis of this question is beyond the scope of the current paper and is dealt with separately in Panfili et al (2021). However, for the purposes of the current work, we reproduce one of the Figures from that paper, showing the distribution of fixation locations relative to footholds

N+1, N+2, and N+3, and N+4 for one subject, in rough terrain (see Figure 5.5). It can be seen that the fixations are centered on the foothold locations, with a standard deviation of 38.8, 27.3, 26.6, and 29.1 cm for steps N+1, N+2, N+3, and N+4 respectively along the direction of travel. The standard deviations in the orthogonal direction were about 31.7, 25.4, 27.7, and 29.1 cm. Thus subjects look close to the locations where the feet are placed, with the future foothold often falling in the parafoveal retina. Another notable characteristic of gaze is that it seems to be used to anticipate undesirable foothold locations. Figure 5.6 shows fixations further along the path assuming the subject maintains a straight walking path. However there are instances where the subject ends up not taking the straighter path, which seems to be triggered by some kind of visual information acquired during the forward fixations. These results suggest a critical role for gaze in the control of foot placement and path planning, further motivating analysis of structural features of the terrain.

5.5 Role of height changes

Previous work has demonstrated human tendency to minimize energetic cost during locomotion. It is plausible that subjects avoid stepping up and down over large rocks in order to reduce energy expenditure in this naturalistic complex terrain. This would also result in more stable locomotion, since it deviates less from flat ground which is most stable. On top of this, excessively large rocks would be avoided altogether. We therefore sought to evaluate the flatness of chosen path segments relative to comparable path segments that were not chosen. Doing so required analyzing the terrain in a way that properly accounted for the behavior goal of the task as well as the limitations of the subjects. To do so we first excluded locations on the terrain where the average

local surface slant exceeded 33 degrees (used as the maximum walk-on-able slope based on results from [120]). This provides a collection of locations that we assume capture all viable step locations. We then calculated from all recorded subject steps the maximum and minimum step lengths (in terms of leg length), maximum absolute step height (step height change either downwards or upwards), and the maximum angular deviation from a step directly towards the subject's final step location. Distributions and schematics depicting these quantities can be seen in Figure 5.7. These three constraints define possible steps between the possible step locations (as described above).

We then use the combination of viable step locations and possible steps in order to simulate possible paths across the terrain. A depiction of this process can be seen in Figure 5.8. At each subject step location, a possible path can be sampled by simulating a random walk down the viable step locations and connecting steps between these locations. Repeating this process from a single starting step location allows multiple possible paths from a given location to be sampled, and used as comparison to the actual chosen path. For this analysis we sample a sequence of 5 steps, or 6 step locations including the starting step location. These sequences will be referred to here as paths. For each step location, there is thus an associated actual path, which is the recorded subsequent 5 steps relative to that step location, as well as a distribution of *possible* paths relative to that step location. The actual path can be compared to possible paths in order to determine the basis upon which it was selected when other paths are possible given observed stepping behavior. The data from Figure 5.7 shows the extent to which subjects will deviate from their preferred gait (peaks in the distributions) in order to better handle the complexity of the terrain. How this capacity for modified steps interacts with the terrain is of interest. Here we examined a simple statistic, which was the

average step slope of paths. For each possible path as well as the actual chosen path, we computed the average height change of all steps along the path (see Figure 5.9 for schematic of mean slope calculation process). Computing the average height change of a path is an attempt at capturing the difficulty or something proportional to energetic cost associated with that path. This results in two distributions of mean step slopes, those from the randomly sampled possible paths, and those from the chosen paths. These distributions are shown in Figure 5.10.

The chosen path average slope distribution has a significant bias to lower average slopes when compared to the randomly sampled path distribution. The median value for the chosen path average slopes was 9.31 degrees, whereas the randomly sampled paths had 14.91 degrees. There is however substantial overlap between the distributions. 20.75% of the randomly sampled path average slopes lie below the median of the chosen path average slopes, and 20.89% of the chosen path average slopes lie above the median of the randomly sampled average slopes. This analysis did reveal some separation between the two distributions, although it is possible that a different statistic might show even more separation, if said statistic better captures whatever heuristic subjects may be using in selecting paths (it is worth noting that max step slope rather than mean shows a similar separation between chosen and random paths). For example one might examine path statistics that measure something over multiple steps rather than the average over individual steps, however this is beyond the scope of this paper. The current result does however suggest a preference for lower average step slopes, although the amount of overlap despite this separation indicates that this is not a strong preference, and subjects are flexible in their decision making.

Simulated possible paths were also used to compute another quantity of

interest, tortuosity of paths. Tortuosity, captures the path's curviness. More tortuous paths are more energetically costly, as they deviate more from a subject's preferred step width. If possible a subject would prefer straighter paths, however this is not always possible for complex terrain where it is at times impossible or extremely costly to continue on a straight path. Here we attempted to examine this trade off between a desire to maintain a straight path with the desire to avoid large height changes. Randomly sampled paths with tortuosity less than the median tortuosity of all paths are classified as straight paths. The average step slope of these straight paths is calculated. If subjects prefer paths with less height change, assuming they would also prefer straighter paths, one would expect a trade off between the straight path step slope, and the chosen path tortuosity. The average step slope of straighter paths captures the expected step slope if the subject were to go straight, which is presumably the preferable option for flatter terrain. Comparing this value to the tortuosity of chosen paths allows measurement of the trade off between height change avoidance, and straight path preference. A schematic depicting a straight path vs a chosen path, as well as accompanying results can be seen in Figure 5.11

Chosen path tortuosity was positively correlated with straight path average step slope for each subject. The highest correlation coefficient was 0.84, and the lowest was 0.51. In this case higher correlation indicates a higher increase in tortuosity as a function of straight path average step slope. A subject with a higher correlation value would be more likely to take a tortuous path when straighter options are expected to have larger height changes. The positive correlation between straight path step slope and chosen path tortuosity across subjects suggests that there is a tradeoff being made between the two. However, more specific understanding of this tradeoff would require a con-

trolled experiment where the two quantities are manipulated. Another useful approach may be to represent each of the quantities in units of energetic cost with a biomechanical model. This may provide better insight into the nature of the tradeoff.

Subject leg length may be related to preference for height change avoidance. For example a taller subject may be less affected by large height changes and hence is less likely to opt for a more tortuous path. Leg length was strongly negatively correlated with the correlation coefficient for straight path slope and path curviness of the particular subject ($r = -0.8662$, see Figure 5.12 for reference). Subject leg lengths ranged from 810mm to 1035mm, with corresponding correlation values of 0.8399 and 0.5089, meaning the longer leg lengths trended towards smaller correlation values. In other words, the longer the subjects legs, the less relation between the straight path height variability and the tortuosity of the chosen path.

5.6 Mean step slope, step slope over areas, depth features

For the next analyses we shifted focus to examining specific locations on the terrain and areas surrounding those locations, as opposed to 6 step paths. The 6 step path analysis attempted to capture any path level heuristics a subject might use to select paths. However given subject's tendency to fixate very close to almost every step they eventually take, we next focused our analysis on more local information around each step that could be informing subject decision making. Using the viable step locations and linking steps computed previous, the average step slope for possible incoming and outgoing steps to each viable step location was computed. This assigned to each viable step

location something like the expected step slope when stepping on that location. Each viable step location, including the locations that were actually stepped on by subjects now has an associated average step slope. This value was then used to classify a step location as chosen vs not (see Figure 5.13 for schematic). MATLAB's built in RUSBoost classifier was used (see [121]) due to the large class imbalance, as most locations were not chosen. The classifier was trained and cross validated by leaving one traversal of the terrain out and training on the remaining traversals to avoid over fitting. The average performance for each subject was then obtained by averaging across trials for a given subject. The RUSBoost classifier provides in addition to a predicted class, a class probability score, which is used to determine the class but provides additional information regarding probability. To gauge performance, the class probability score at each chosen foothold location is compared to other non chosen foothold locations within a surrounding region. This is done to attempt to address situations in which many foothold locations might be classified as not foothold locations by the classifier, but the chosen foothold location may have relatively higher class probabilities compared to other local possible locations despite not being classified as chosen. This comparison is done by computing the quantile of the class probability score of the chosen foothold locations compared to those of viable foothold locations in the neighborhood. A visualization of the computed average step slope quantity across terrain plotted along with subject paths, as well as the resulting local quantile class probability scores is shown in Figure 5.14.

All subjects had above chance (0.5) mean quantiles of probabilities generated by the classifier within a 4 step by 1 step comparison region. The lowest mean quantile was Subject 2 with a mean quantile of 0.54, with a standard error of 0.01 (measured across trials). The highest mean quantile was Subject

3 with a mean quantile of 0.65 with a standard error of 0.018. Other subjects ranged within these extremes. Interestingly, only a weak correlation between subject leg length and this quantile score was observed ($r = -0.1341$), which indicates that longer legged individuals are less predictable on the basis of average step slope, but only to a small degree. Generally average step slope to and from locations explains a small portion of variance in foothold selection. This suggests that there are other terrain related factors at play, which are explored to some extent in other analysis (where average step slope is aggregated over larger areas, or measured along paths) but these also rely on average step slope. Other factors still need to be explored. Another limitation is that the analysis and a few others implies full awareness of the terrain by subjects, which is not the case because they are limited by their perceptual abilities. This is somewhat addressed in another analysis that uses retinocentric depth images as inputs.

The next analysis was similar, but instead used average slopes aggregated over larger areas as values associated with particular step locations. While results from [64] suggest subjects mostly fixated step N+2, there is still substantial fixation of N+1 and N+3, indicating that information over a range of steps might be important. This analysis still attempts to capture step location specific information, but over a larger area than the incoming and outgoing step locations to that step. Additionally these values are aggregated over many possible steps within a region, as opposed to along paths as in previous analyses. These aggregated values were computed by pooling previously computed average step slope values over cone areas of varying angular width and length. Here we attempted to survey a range of directions and distances relative to each step that subjects might be considering height change information over. The cones were oriented such that the opening was oriented towards the final

step location, and the width and length of the cones was systematically varied such that aggregated average step slopes for a range of different cone areas was computed for each step location. This resulted in many features at each step location, which were then used again with RUSBoost to classify stepped locations versus other locations. The trained classifier was then used with MATLAB's 'predictorImportance' function in order to determine which of the cones were most important when classifying the step location. This yielded two peaks in importance for cones that were 1 step length long with 60 degree width, and 3 step lengths long with 30 degree width. These two features alone were then used to train and test the classifier, resulting in similar performance to a classifier using all features. Each feature was also tested alone, and a similar process for gauging performance (local quantile comparison) was used. Results can be seen in Figure 5.15.

Here for all three conditions (1 step ahead 60 degree cone, 3 steps ahead 30 degree cone, and combined) all subjects had above chance probability quantiles, with the exception of 3 steps ahead 30 for Subject 8, and 1 step 60 deg for Subjects 2 and 8. For the 1 step ahead 60 degree cone analysis, Subject 8 had an average quantile of 0.44 with a standard error of 0.037. Subject 1 (the highest) had 0.62 with standard error of 0.015. For the 3 step 30 degree condition, Subject 8 had 0.48 with 0.036 standard error. Subject 5 (highest) had 0.63 with a standard error of 0.032. All Subjects except subject 8 had larger combined scores than any single feature alone, with the highest (Subject 3) being 0.69+-0.017. Both cone area features predicted subject foothold selection above chance. The performance of the two together indicates some kind of interaction between the two, since both features considered simultaneously outperforms what one might expect from the combination of both. Here again leg length only had a modest relationship to performance for the three dif-

ferent categories ($r = 0.21, -0.051, -0.12$ for each category respectively). As stated earlier, the substantial amount of unexplained variance suggests that other terrain based features are at play.

It is worth noting that previous results (Matthis 2018) suggest subjects direct gaze at 2-3 steps ahead, while the features in this analysis incorporate information up to 3 steps ahead of a location. If subjects were employing a strategy suggested by this model where information up to 3 steps ahead of a given step is computed and considered, this would suggest that subjects could be incorporating information up to 5-6 steps ahead of them. Since subjects seem to direct gaze toward step locations that are being considered 2-3 steps in advance, the additional 3 step information seemingly incorporated suggests a 5-6 step window of advanced planning relative to the planted foot.

One issue not addressed by either of the previous analyses is that the subject perspective and perception is not being accounted for. The other analyses implicitly assume that a subject would have full information about the environment and the step slopes associated with each location, however in reality subjects must make eye movements and acquire this information visually. To better model this process we combined the environment mesh data with aligned foothold location, eye position, and eye direction data allowing approximation of depth image inputs to the visual system with foothold locations in the depth image space. These retinocentric depth images are then used as inputs to a CNN, where the target output is a distribution of foothold locations in the depth image coordinates. Ground truth foothold location distributions are computed by centering Gaussian distributions at computed foothold locations. Subject perspective depth maps approximate the visual information subjects have when deciding on future foothold locations. If a CNN can predict these locations above chance using depth information, this would indicate

that depth features can be used to explain some variation in foothold selection. A visualization of high and low performance, as well as the results can be seen in Figure 5.16. Median AUC values for all subjects were significantly above chance. The maximum median AUC of 0.79 indicates that the 0.79 is the median proportion of pixels in the circular image that can be reliably labeled as not a foot location while correctly labeling each foot location. Because at each frame, up to 5 of the next upcoming footstep locations are present in the image, the CNN is most likely learning local terrain structure features that are predictive of good footholds at multiple distances. The lowest performance was for Subject 3 with a Median AUC of 0.68, which is still well above chance (0.5). Interestingly, here leg length shows a modest correlation with median AUC ($r = 0.46$), which suggests that longer legged individuals foot selection is more predictable on the basis of local structure features, although the precise nature of this relationship is still unknown. The results from the depth image CNN analysis show that subject perspective depth features are predictive of foothold locations. These depth image features may or may not overlap with the step slope features shown to be predictive in the previous analysis, although this analysis does better approximate how subjects might use such information. Despite this advantage there is still the limitation of assuming full access to the full resolution depth image at all eccentricities, since depth perception falls off with eccentricity [122]. Incorporating this limitation, or others like it, may change the results.

5.7 General Discussion

The results of this investigation provide evidence that height change avoidance can explain some but not all of step location selection during a natural complex

terrain navigation task. Subject chosen paths have lower average step slopes than randomly sampled paths, which indicates a preference for lower height change. However the overlap of these distributions suggests it is a preference not a hard constraint. This preference for lower height changes also seems apparent at different spatial scales, with step location prediction with a classifier performing best when incorporating height change information over both 1 step and 3 step lengths. Walkers seem to trade off between paths with lots of height change and paths with lots of turns. The height change of straighter simulated paths correlated with the tortuosity of subject chosen paths. The variation of the strength of this correlation across subjects suggests that this is a subjective parameter, however the pattern was consistent across subjects. Finally, depth image features seem to be involved in foothold localization, however the extent to which these are picking up on the same information as height change vs other features like local surface slant is not clear.

The chosen vs random path analysis allowed capturing behaviorally relevant statistics from the terrain rather than something like height variation of all terrain points within some region, since it conveys information about how the specific terrain structure and geometry might relate to a subject's choice of step locations as they traverse over the terrain. While this does add ecological validity to the measures, there are assumptions being made about what would constitute possible steps or step sequences. Here possible steps are being considered fully specified by step slope, direction relative to goal, and step length, which geometrically is the case. However there are other factors that might determine 'possibility' of steps such as their relationship to previous steps. For example a two step sequence of steps that when considered alone each fall within the range of possible steps, but would be highly unlikely or not happen in the data (for example two steps of very large height change

may not occur sequentially, even though large height changes for single steps are observed). A deeper analysis of higher order relationships between steps is needed in order to better constrain possible steps for the purpose of step sequence simulation.

The lower median chosen path average step slope when compared to random sampled possible paths indicates a preference on the part of walkers, although not a strong one. The overlap of the distributions suggests that there are multiple possible paths that also have relatively low average step slopes, and that there isn't a specific value for average step slope that prevents a walker from taking a path. In that case there would be more separability between the distributions. There is however a point where a sharp decrease in the distribution of chosen path average step slopes occurs, at around 10 degrees. However there are still a substantial amount of chosen paths that have higher values. This suggests that walkers have a preference for lower average step slope paths but are flexible and other factors may make taking those higher average step slope paths necessary, although that is not captured by looking at this single statistic.

Analysis of individual step location average step slopes shows that there is some relationship between step slope and subject foot placement choice. The remaining unexplained variance suggests that other important factors are involved, that subjects only have a preference for low step slope, or that there is a wide availability of low step slope locations. Qualitatively it seems like a mixture of the first two based on the aligned walk visualization. Subjects tend to cross over high step slope locations, while visiting the low step slope locations at a rate above chance but still a small degree. This suggests that there are other factors driving subjects over large step slope locations when these other factors are more favorable, or that step slope is only a light con-

straint. There could also be instances where specific scenarios or values for other factors result in an increased weighting of the average step slope. This idea is based on the observed convergence of multiple walking trajectories to the same locations of low step slope, but only at certain regions of the terrain.

There are some limitations to this approach. The first is that it implicitly assumes full awareness of the potential average step slopes of a large region surrounding each location as subjects navigate across the terrain. While this is not unreasonable given the constrained region within which probability scores are compared, one might expect different results when somehow accounting for what locations subjects are able to perceive and reach given their current body state and vantage point. The best way to implement these kinds of constraints is not clear, however more variance could be explainable if subject perception and action limitations were built into the predictions. For example despite being possible (as observed in our data) particularly wide steps may be less desirable, and so weighing the probability estimates to in some way reflect this could improve prediction performance. Additionally, as previously mentioned, there are likely other factors influencing subject foot placement besides expected step slope of a location. Another example could be a particular sequence of foothold locations and configurations that works particularly well for a given body state. Identifying and incorporating these factors into the model would likely result in a higher score and as a result a better understanding of how subjects are making decisions, but is beyond the scope of this investigation.

Analysis of multiple step slope features at individual step locations show that both features (1 step ahead, 60 degrees wide, and 3 steps ahead 30 degrees wide) both capture future potential step slope given a certain step location. However each capture this at different scales, where 1 step only adds an addi-

tional step to the hypothetical planning horizon, whereas 3 steps adds 3. This indicates that subjects are taking future steps in addition to the 2-3 steps ahead [64] into account when planning their next step. This could indicate that subjects are able to acquire whatever information is necessary to determine future expected step slope peripherally from their fixations directed close to footholds.

Although each of these features that captures future step slope at different scales achieves similar performance to average step slope for a given location (although in the case of Subjects 2 and 8 for the 1 step, 60 degree cone and Subject 8 for the 3 step 30 deg cone only marginally above chance), interestingly when combined the two features explain more variance than either alone, or than the average step slope. This suggests that subjects take both factors into account, and that they capture different information useful to the subject for making decisions. It is not clear why this performance gain for combination of features occurs in the case of Subject 2, since the 3 step 30 degree feature alone did not explain foot locations above chance. Perhaps it is the result of interactions between the features.

Similarly to the result of the mean slope based classification analysis, there is still substantial variance unexplained by using even both of these features combined. This suggests again that either subjects take other factors into account, or that they are flexible when it comes to these two features and they are merely capturing a preference, or a combination of both.

The straight path slope versus chosen path tortuosity analysis revealed that consistent across subjects, the tortuosity of subjects chosen paths was highly correlated with the increase of mean slope of available straight paths. This suggests that subjects make some kind of trade off between going straight (which would minimize the overall distance travelled) and avoiding large height

changes (which is more costly that flat ground walking) when deciding on paths. The different slopes and correlation values between subjects suggest some kind of individual preference for what this trade off should be.

While this correlation is strongly suggestive of a tradeoff it relies on the assumption that a subject would have taken the straighter paths had they not had as high of average step slope values. One can only definitively make this kind of claim with explicit experimental manipulation where subjects are forced to take one of two paths. This could be carried out with a controlled experiment where the experimenter has control over parameters like the tortuosity of a path, and the average step slope of a path, and can somehow through the design of the experiment force a choice between the two paths. Through manipulation of the parameters one could determine precisely given a forced choice between the two, where the cutoffs for tortuosity and height change are in path choice. However our observations in this unconstrained task provide a solid ecological argument for carrying out such manipulations, since this tradeoff is apparent in the data.

Another limitation which opens an interesting avenue for exploration is the lack of a shared unit of cost between tortuosity and step slope. Presumably each incurs extra cost to the walker as they increase, however a better understanding of this tradeoff requires understanding these extra costs in the same units. One approach could be to use a biomechanical model of a walker in order to determine the relative energetic cost of different paths. This would also have an added benefit of developing a better understanding of between subject variation, since the different correlations and slopes might be explainable by a biomechanical model with adjustable parameters based on the different subjects physical characteristics. The different correlation values between subjects do suggest a subjective value when determining the tradeoff, and the leg

length results suggest that a lot of the between subject variation is due to biomechanical factors. Subjects with longer legs are better able to deal with large height changes and hence may opt for more height varying paths rather than more circuitous ones.

The above chance mean and median AUC values obtained when measuring the CNN foothold location predictions indicate that subject perspective depth image features explain some of the variance in foothold location selection. Looking at the depth information from the subject's point of view restricts the analysis to information that the subject was known to have had access to. This is important since the previous analyses did not take this into account, and consisted purely of environmental variables. Despite this advantage, here we do make an implicit assumption which is that there is no degradation with acuity for subject's depth perception. The CNN is receiving a full resolution depth image even at higher eccentricities whereas it is known that depth perception has a falloff [122]. More accurately modelling the limitations of the visual system might affect our results.

Additionally, the ability of the CNN to predict foothold locations using depth images compliments the other results well since depth would be the manner in which the subject could most directly infer height change information in the environment. This means that the predictive power of height change information of the environment and the depth information from the subject's point of view are from the same source, which is the structure of the environment. This means that this kind of analysis could get at subject perception specific phenomena such as occlusion. However, doing so would require framing each prediction based analysis in terms of the same performance metric in order to determine how much of subject choice is due to limitations in their perception of the environment.

It is important to note that depth features may or may not overlap with the step slope quantities focused on in other analyses. Step slope would be reflected in subject perspective depth images, however so would other geometry related features like local slant. It is however difficult to determine what depth features are being used to predict foothold locations.

One advantage of subject perspective depth maps is that they may implicitly bias the selection of depth features even at non foothold locations. The reason is that humans are known to direct gaze towards points of interest and importance to the task at hand [12], [13] (also see [9]). This means that there might be instances where the subject is generally looking towards a desirable foot location but due to other factors the fixated location is not where the subject places their foot. One might still gain insights by focusing analysis on gazed locations even when they are not chosen as footholds, rather than random locations on the terrain.

The above chance performance of the retinocentric depth image CNN is encouraging since it paves the way towards a biologically plausible model of this behavior, with subject perspective depth images being a closer approximation to the visual information accessible by subjects as they completed this task.

5.8 Appendix

5.8.1 Pre-processing

Motion capture data

For more detailed description of pre-processing of motion capture and eye tracking data, see [64] and [23].

Photogrammetry with Meshroom

Meshroom [75] is a software package that combines multiple image processing and computer vision algorithms in order to estimate camera position and environmental structure from a series of images. First, features that are minimally variant with respect to viewpoint are extracted from each image. Images are then grouped and matched on the basis of these features, followed by matching of the features themselves between images. Feature matches from previous step are used to infer rigid scene structure (3D points) and image pose (position and orientation) for each of the image pairs. An initial two-view reconstruction is created, which is then iterated on which each new image. Depth values for each pixel in the original images are computed using the inferred point cloud. Depth maps are then merged into a global octree where depth values are merged in to cells. 3D Delaunay tetrahedralization is then performed, followed by graph cut-max flow and laplacian filtering. Finally the resulting mesh is then textured, where each vertex' visibility is factored in and matching pixel values are averaged for each triangle.

Here we take the outward facing world camera from the pupil labs eye tracker and input its video into Meshroom. Pupil labs world camera video is first processed into indvidual frames using ffmpeg [123]. The individual frames are undistorted using a camera intrinsic matrix estimated by checkboard calibration [72]. This allows a pinhole assumption for the images (citation), which facilitates reconstruction. The estimated focal length in pixels is supplied as an additional parameters to Meshroom. Meshroom then takes the images and runs them through the above decsribed pipeline, resulting in a 6D camera trajectory (3D position and 3D orientation), with one 6D vector for each frame of the original video (See Figure 5.1 for rendered image of textured Mesh output).

Motion capture to mesh data alignment

Meshroom provides pose (3D position, and 3D orientation) estimates corresponding to each of the inputted video frames from the eye tracker's world facing camera. This position and orientation (6D) is in the same coordinate system as Meshroom's estimated rigid scene structure (3D point cloud or 3D triangle mesh). The next step of our analysis involves alignment of the pupil-Shadow motion capture and eyetracking data with Meshroom's coordinate system. The 3D orientation of the world camera in the previously described eye tracking and motion capture data is available from the procedure described in [64]. This 6D camera pose is then aligned to the 6D camera pose of the Meshroom estimated camera in Meshroom's coordinate system. A single 3 euler angle rotation that minimizes L2 error at each frame is estimated using fminsearch in Matlab. This transformation that best aligns the two camera poses is then applied to the entire skeleton and gaze data. The skeleton is as a result pinned both in location and relative orientation to the 6D pose of the Meshroom camera estimate (see Figure 5.2 for visualization of alignment). After the head location and orientation alignment is computed, the motion capture data is scaled such that the distance between the motion capture system's estimated foot position during footfall frames and the closest point on the mesh is minimized (ensuring maximum contact between the motion capture foot position estimates and the mesh). This maximum contact scale factor is the applied to all of the motion capture data for that traversal.

Cross subject alignment

Cross subject alignment involved the use of open source package CloudCompare (CloudCompare) in order to manually extract corresponding keypoints

between meshes to be aligned, perform coarse alignment via similarity transform, and perform fine alignment using the iterative closest point algorithm. For unique terrain segment that subjects traversed multiple times a single traversal and its corresponding Meshroom terrain reconstruction output was selected as the reference terrain. 5 reliably detectable features were chosen as key points, and these 5 features were located for the terrain outputs for each of the other traversals across the same terrain. Using the set of 5 corresponding keypoints, a best fitting similarity transform (translation, scale, and rotation) was computed and applied to the 'moving' terrain such that it would be best aligned to the 'fixed' terrain. This aligns the 5 keypoints for each of the terrains, which also aligns the rest of the terrain coarsely. Fine alignment is then performed using the iterative closest point algorithm. This locates for each point in 'moving' point cloud the closest point in the 'fixed' point cloud and estimates a similarity transform that minimizes this distance further, with multiple iterations. This fine alignment ensures even better correspondence between the two terrains.

This process is repeated for each terrain until all terrain data has been transformed into the same coordinate system as the chosen reference terrain. The transforms are then stored and applied to the aligned motion capture and eye tracking data. This allows analysis of all chosen paths in the same coordinate system. This alignment was not used in main analysis, but was used for visualization of all subject trajectories in the same reference frame (see Figure 5.14, and in addition was useful for computing a cross mesh error metric.

Possible step and path simulation

In order to facilitate analysis of the data in regards to path planning and foot placement, all possible foot locations and steps between foot locations are predetermined using various constraints. The first is a constraint on possible step locations. Maximum walk-on-able slope was previously measured in [120]. Here we use the maximum value for the walk-on-able slope since our participants would not have to maintain gait over the slope for multiple steps, whereas the max walk-on-able slope was computed under those conditions in the study. Viable foothold locations are computed using mean surface slant angle in a foot length area. The 3D triangle mesh representation of the terrain allows calculation of a surface normal vector for each triangle. A mean local surface slant is then calculated for each point in the point cloud representation using an average of all triangle calculated surface slants within a radius of one foot length. After viable foothold locations are selected via mean triangle surface slant angle filtering (where all surface slant angles below the walk-on-able slope cutoff are deemed viable), viable steps *between* viable foothold locations were determined based on 3 constraints (See Figure 5.7). In the observed data, each step subjects took was used to compute a step slope (arctangent of height over distance ratio, or slope of the step), a goal angle deviation (deviation of step direction from the goal direction in the plane perpendicular to gravity), and a step distance deviation (deviation of the step length from median step length). The step slope is computed by taking the change in vertical coordinate of sequential foot locations, and dividing by the magnitude in two dimensions of the line connecting the two locations in the forwards and lateral coordinates. In other words the step vector is projected onto the ground plane, with the vertical component ignored, and the magnitude is calculated,

and the height change is divided by that magnitude. Goal angle deviation is computed by taking the direction of the step in this same vertical projected ground plane, and taking the angle between this direction vector and the vector pointing from the initial foot location in the two step sequence to the final step location for that traversal (the goal direction). Step distance is calculated by taking the euclidean distance of the line connecting each set of two foot locations for each step in 3 dimensions. The maximum observed values for each of these was computed, and all possible steps between selected viable foothold locations (pairs of viable foothold locations) that were within the maximum values for each of these (when a hypothetical step between the locations is considered) was deemed a possible step. This allows analysis of the terrain data with respect to possible steps and step locations, as well simulation of hypothetical paths given some initial step location.

Retinocentric depth image extraction

The aligned motion capture, eye tracking, and photogrammetric data was used to calculate subject perspective depth images as they traversed the terrain. Using Blender [76], a virtual camera was translated to be centered at the estimated camera location for each frame of a traversal, and rotated to be oriented in same direction as the subjects gaze based on the aligned eye tracking data. The virtual camera is then used to capture a depth image of the 3D triangle mesh representation in Blender using it's "Z-buffer" method. The virtual camera is a perspective pinhole camera, facilitating calculation of foothold locations in the camera's image plane. This is done by taking the intersection between lines connecting future foothold locations, and the current camera position, with the camera's image plane. The depth image is then transformed such that the pixel coordinates correspond to retinal coor-

dinates (theta,rho), with distance from the center of the image in pixels being convertible to eccentricity by scaling this distance by 1/2 of the width of the image and multiplying by 22.5 degrees. The polar angle of a given location in the image would correspond to the same polar angle in retinal coordinates (theta). The retinocentric depth images are then shifted such that the depth value of the center pixel (fixation point) is zero by subtracting the depth at the fixation point from the rest of the image. The depth images as a result represent depth relative to fixation point of other points in the image, with the fixation point always being 0. These subject perspective depth images allow considering information from the subject's perspective when trying to predict foothold locations, whereas other analyses implicitly assume full awareness of the environment when choosing foothold locations.

5.8.2 Detailed Analysis

Possible path vs chosen path analysis

This analysis leverages the pre-computed possible step locations and possible steps connecting them from (5.8.1). For each traversal of the terrain, each chosen step is iterated over, and the next 5 steps that the subject took relative to that step location are considered. This 6 step sequence is treated as a 'path' in this analysis. For each path, a subset of the possible step locations is selected using the 'maxflow' function in MATLAB, which can output a subset of nodes that have non-zero flow values in a directed graph given two selected nodes. This subset represents step locations that can be visited from the starting step location and still have available paths to the end location (6^{th} step in path). Other possible paths connecting the two end points of the actual path are then sampled from this subset of possible step locations and connecting steps.

For each of the simulated paths as well as the chosen path, the average step slope for steps within the path is computed and assigned to the path. We then compare the average step slope for chosen paths compared to randomly sampled paths. This analysis allows comparison of average step slopes for chosen paths compared to ones encountered if paths were randomly chosen (which with the exception of constraints on possible steps, is purely terrain driven). The choice of number of steps to include in a 'path' is arbitrary, and does not necessarily get at how a subject might be choosing paths. It does capture expected average height changes for randomly sampled paths over the chosen amount of steps, which is useful for comparison.

Mean incoming and outgoing step slopes

In this analysis each viable foothold location determined previously (see Section 5.8.1) is iterated over and the average HoD of all connected outgoing and incoming steps is computed. In other words, the average slope of steps going to or from that step location is computed for each step location. This is then used as the mean HoD for that location. Each foothold location also has an associated class, which is either a stepped on location or not a stepped on location. Here we use RUSBoost [121] a boosted classification method for unbalanced samples in order to classify actual foothold locations vs non foothold locations. Results for each traversal are obtained by using the other traversals as training examples, and testing on the withheld traversal. Classification confusion matrices are calculated for each traversal, however for evaluation of performance a different method is used. The class score obtained from the RUSBoost classifier is assigned to each foothold location. Then each actual foothold location's class score is compared to other class scores for other foothold locations within a rectangular region, with short length of 1 step in

the walking direction, and 4 steps in the perpendicular direction (see Figure 5.14, (A)). The quantile of the chosen foot locations class score is then computed relative to the rest of the scores of viable foothold locations within the region. This method of performance evaluation is used to compare the relative probability of chosen location compared to other local non chosen locations.

Mean incoming and outgoing step slopes over cone areas

In this analysis two features were computed for each viable foothold location. Each feature is the average step slope as computed in 5.8.2, averaged over multiple step locations within different areas. For the first feature, step locations within 60 degrees of the goal direction and within 1 step length are considered. This captures the likely encountered step slope given that subjects would prefer to go straight towards the goal from a given location. The other feature is the average step slope for all step locations within 30 degrees of the goal direction, but within up to 3 step lengths. This feature captures future step slope from a particular location up to 3 steps out, which contains more information about the future likely step slope from a particular location than the other feature. As in 5.8.2, RUSBoost is used to classify actual foothold locations vs non foothold locations based on these features. 3 different classification models are trained, one for each feature, and one that uses both features simultaneously. As in 5.8.2 the class score from the classification models are computed and the scores of actual foothold locations are compared to those of other viable, not chosen foothold locations within the same 4 step by 1 step region. The quantile is again used as the performance metric.

These two feature were determined through preliminary analysis using multiple combinations of step lengths and and angular distances. A grid search method was used ranging across 1 to 10 steps ahead, and between 0 and 60

degrees of angular deviation from the goal direction, each spaced at 10 intervals, yielding a total of 100 different features. A RUSBoost trained classifier that used all 100 features was trained, followed by feature selection ("predictorImportance" in Matlab, an algorithm to determine the predictive measure of association for each feature). The two chosen features were the most predictive, with performance being comparable between all features being used and only the two.

Straight path slope vs. curved path probability

This analysis relies on the paths discussed in the 5.8.2. In this analysis we also compute for each path a tortuosity metric. This is computed by taking the actual cumulative distance of the path (here computed by summing the length of each line connecting step locations), and dividing by the straight line distance of the path, or a line connecting the start and end foot locations. Again, at each step we consider the chosen path (6 step sequence), and possible paths are simulated along the subset graph calculated using maxflow. For each traversal, the distribution of tortuosities for chosen paths is calculated and the median is used to determine a cutoff for 'straight paths'. The mean step slope for the randomly sampled paths that have tortuosities below the median actual observed tortuosities are computed. These are treated as the average step slope the subject would encounter if they tried to take a straighter path for that segment of terrain. Thus for each path there is an associated tortuosity, as well as the mean step slope of possible straight paths. We then use these values in our analysis.

Assuming subjects would prefer straighter shorter paths, but also would prefer to avoid significant height changes, this analysis would capture a trade off between the two since steering to avoid large height changes would result

in increased tortuosity.

Retinocentric CNN

The retinocentric depth images with foothold locations known in the same image space are then further processed for use in a convolutional neural network (CNN). The convolutional neural network used in this work has a convolutional - deconvolutional architecture with three convolutional layers followed by three transposed convolutional layers, followed by KL divergence loss computed with a target foothold location distribution (See below table for parameters used and descriptions of each layer).

Layer	Output Shape	# Params
Conv2D	(100,100,4)	1604
BatchNorm	(100,100,4)	16
MaxPooling2D	(50,50,4)	0
Conv2D	(50,50,8)	3208
BatchNorm	(50,50,8)	32
MaxPooling2D	(25,25,8)	0
Conv2D	(25,25,16)	12816
BatchNorm	(25,25,16)	64
Conv2DTranspose	(25,25,16)	25616

The ground truth foothold location distributions are computed by taking the known coordinates of foothold locations in the depth image and smoothing with a gaussian kernel with sigma = 5 pixels, which corresponds roughly to 1 degree of visual angle, although the conversion between pixels and degrees is not constant throughout the visual field.. This is to capture any noise in our estimation of foothold location to allow more robustness in the CNN learned features. Depth images were 45 degree of visual angle across, meaning they extend to 22.5 degrees of eccentricity.

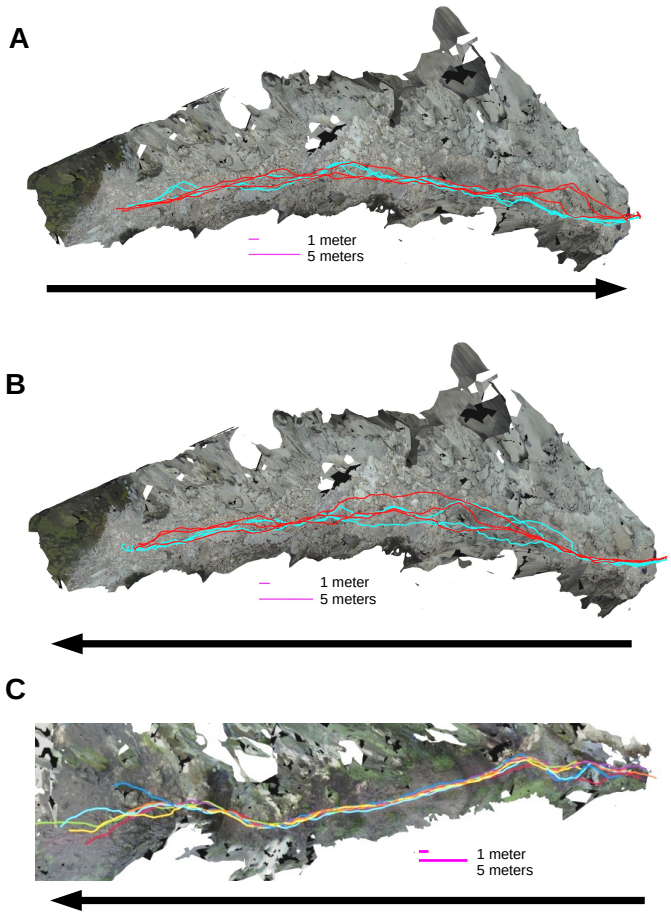


Figure 5.4: Overhead view of Austin and Berkeley data. Subjects walking from left to right (A and C) or right to left (B). Different colors correspond to different subjects, each traversing in each direction 3 times for Austin data (A and B), and once for Berkeley data (C).

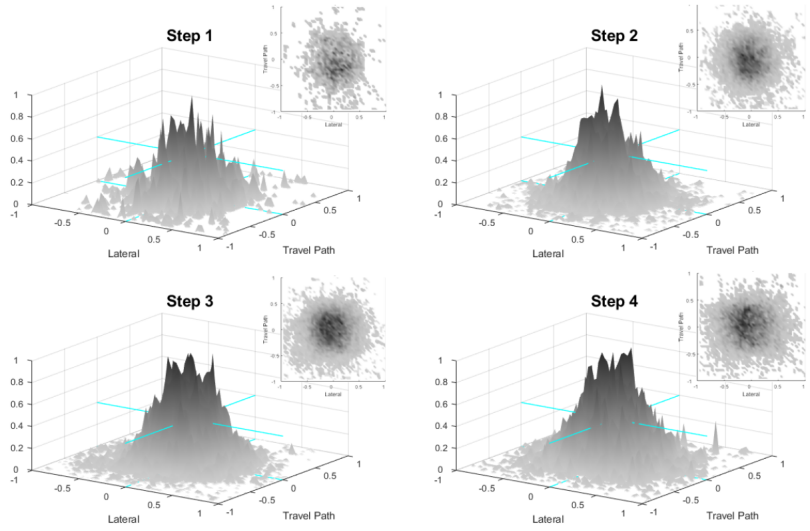


Figure 5.5: Distribution of ground fixations relative to future foothold locations. Each graph shows distribution of gaze on ground relative to upcoming foothold locations N+1, 2, 3 and 4 respectively. Axes are in meters along the direction of travel, and the orthogonal direction.

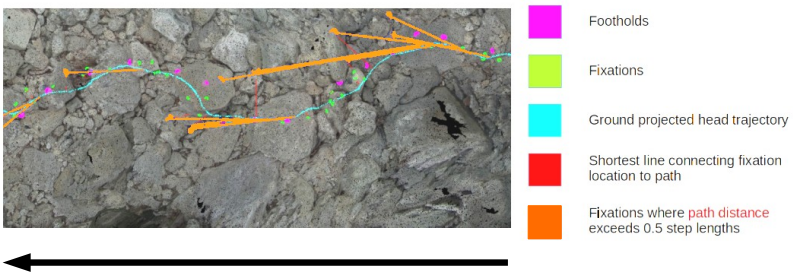


Figure 5.6: Gaze is used to select paths. Here we show a representative excerpt of data where gaze is directed further along the path, in this case at locations that are not travelled to. Gaze is apparently used to determine the viability of paths ahead of time, since fixations further ahead in straight directions often precede turns that deviate from the fixated locations.

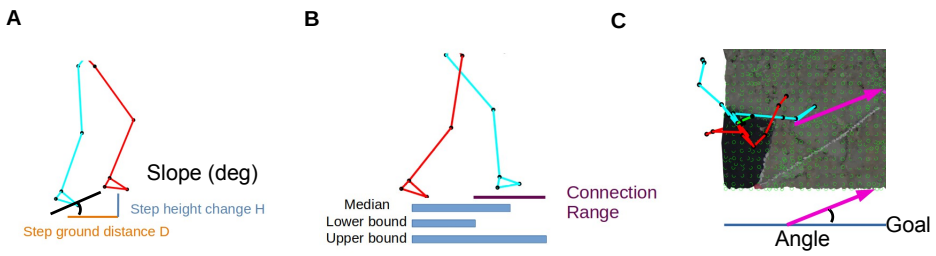


Figure 5.7: Step parameter distributions. In flat ground scenarios (such as a treadmill) walkers converge towards energetically optimal gaits. This means regular steep lengths and directions, without any height changes in their steps due to the flat ground. Because of the terrain complexity, walkers need to modulate their gait by adjusting step slopes (A), lengths (B), and directions (C) to accommodate the complexity of the terrain. The ranges of step slopes, lengths and directions can then be used to compute possible steps that did not occur.

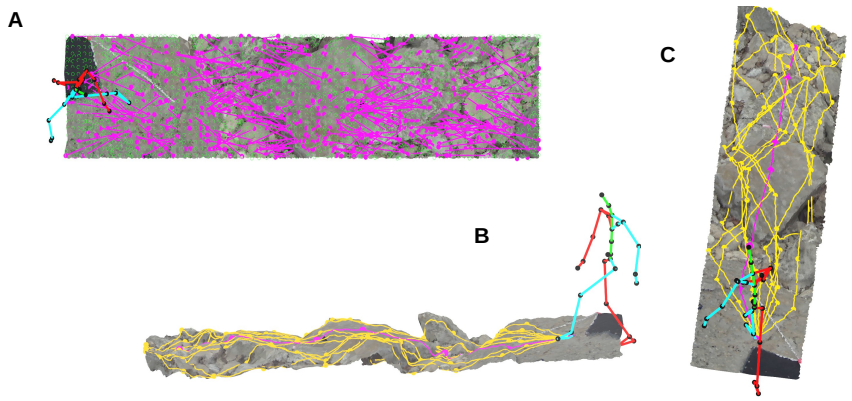


Figure 5.8: Possible step location and step schematic. Possible step locations are determined using a local slant cutoff. The slant values (the angle between a vector pointing straight upwards and the normal vector of a particular triangle face) of each triangle face on the mesh are computed. Then they are averaged across foot length sized radii, providing a local slant value at each mesh location. Each location with a local slant value below a threshold determined by previous empirical results on the maximum walk-on-able slope were used was labeled as a viable step location. Then these locations are downsampled with a voxel-wise mean filter, with the voxel width being half a foot length. Each of these possible foothold locations are then checked against all other possible foothold locations, and all pairs that fall within the ranges described in [3] are connected with a step. This precomputed graph of foothold location nodes and possible step edges (A) is used to sample possible paths from a given location. B and C show from two different perspectives the a subject's chosen path (magenta) as well as 10 randomly sampled paths (orange)

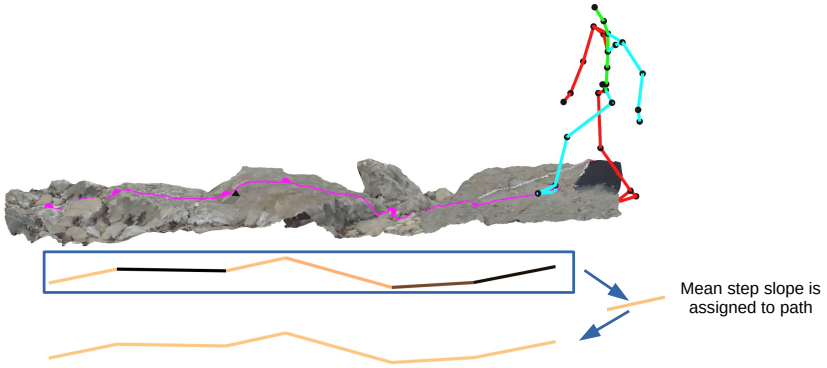


Figure 5.9: Schematic depicting how mean step slope is calculated and assigned to each path. For each path (6 foothold location sequence, either actual observed or simulated based on viable locations and steps), the step slope is calculated for each step in the path, and the mean step slope is calculated. This mean slope is then assigned to the path for later analysis. Mean step slope is calculated for all paths (both actual path, and the randomly sampled possible paths), and this statistic is used to compare chosen vs random paths in later analysis.

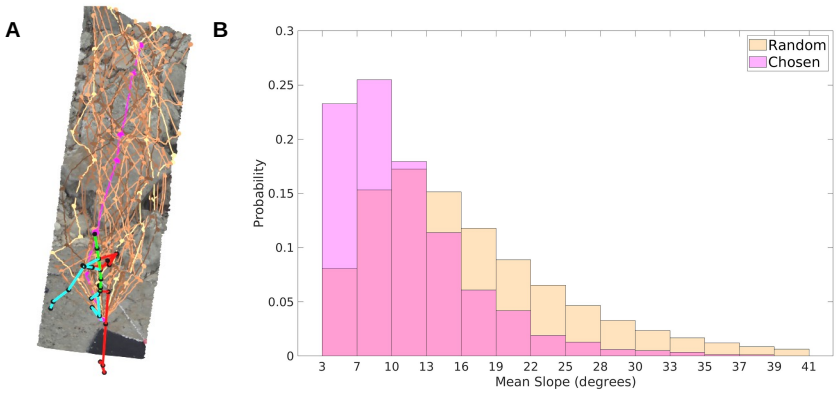


Figure 5.10: Chosen vs random path mean slope. Using the previously described method we can randomly sample available paths in order to compare them to the chosen path. (A) shows a subject's chosen path (magenta) along with a subset of randomly sampled paths with their mean step slopes colormapped onto the path. (B) Shows histograms of the mean step slope for paths that were chosen and randomly sampled paths. The chosen path distribution is shifted to the left with far less rightwards skew

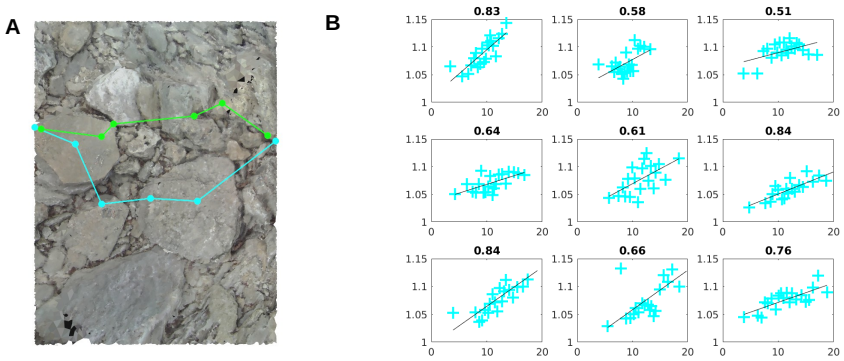


Figure 5.11: Turn probability vs straight path slope. Here 6 step sequences are analyzed. For each sequence the distance of the straight line connecting the first and last step is computed, as well as the distance of the actual path. These are used to compute tortuosity of the chosen path. In addition, 10,000 paths are simulated along the max flow graph, which includes on locations that are reachable from the start location and end location. The straightest paths (paths with tortuosity less than the median tortuosity of chosen paths) are used to compute an average straight path step slope. This average straight path step slope is then compared to the tortuosity of chosen paths

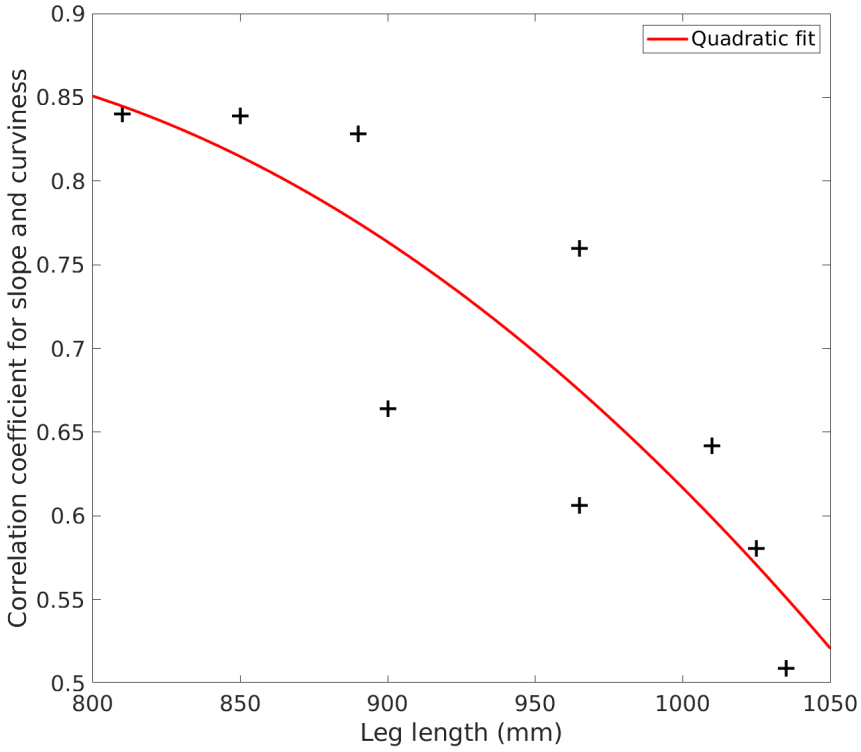


Figure 5.12: Relationship between leg length and correlation value between straight path step slope and path tortuosity. Subject length length (in millimeters) is plotted on the horizontal axis, against the correlation coefficients for each of the plots in Figure 5.11 plotted on the vertical.

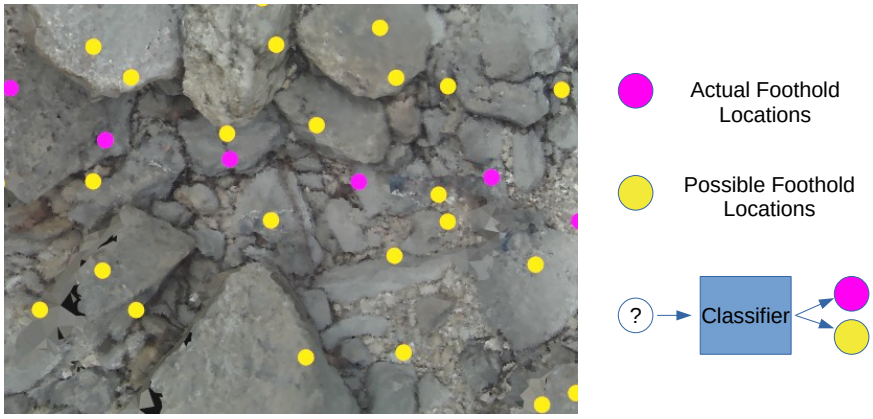


Figure 5.13: Viable foothold location classification scheme. Viable foothold locations have been previous selected based on local slant cutoff. Actual chosen foothold locations and viable but not chosen foothold locations are labeled as separate classes, and different input features corresponding to each location are used by a classifier to distinguish between the two classes.

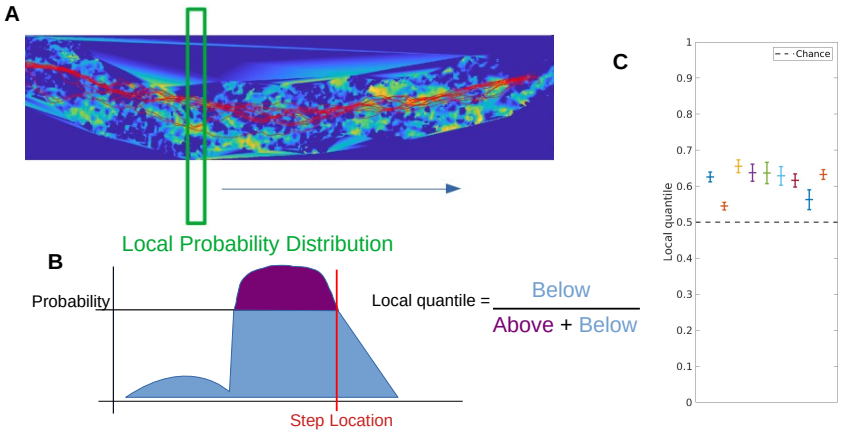


Figure 5.14: Mean incoming and outgoing slope prediction. Using the pre computed steps as described in [3], the average step slope of all steps connected to each step location (either incoming, or outgoing steps) is computed. Each step location then has a corresponding average step slope value. This value is used to classify each step location as either an actual chosen step location or not, using RUSBoost. The probability of each location being of the chosen step class is then obtained for each step location. At each chosen step location, this probability is compared to all neighboring possible step locations within 4 steps in the direction perpendicular to the walking direction, and within half a step in the walking direction. This quantile of the chosen step location probability compared to the other probabilities is measured.

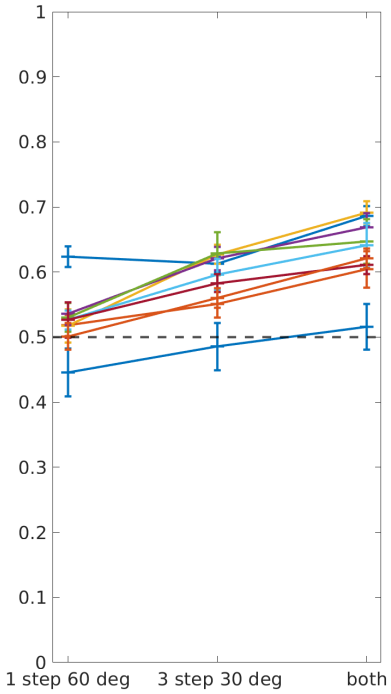


Figure 5.15: Mean incoming and outgoing slope over areas. Similar to the analysis in [7], here the average step slope values for possible step locations were averaged across space and these averaged were assigned to step locations rather than just the value at a given step location. The area over which the values were averaged was varied systematically with differing lengths and widths of a cone that extended in the walking direction. Feature selection was used to determine strongly predictive cones, with 1 step length, 30 degrees wide, and 3 step lengths, 60 degrees wide cones being strongly predictive. Probabilities generated by RUSBoost using either feature, or both features, were used to compute quantiles as in [7].

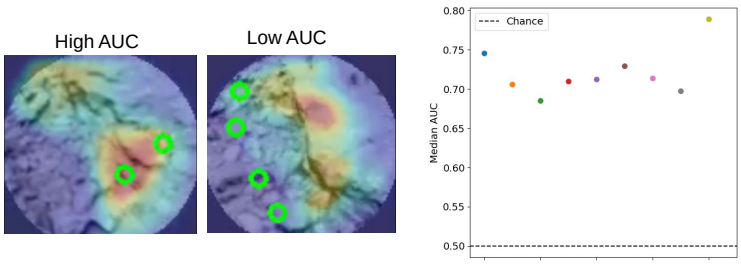


Figure 5.16: CNN based foothold location prediction. A CNN was trained to predict foothold locations in subject point of view depth images. Depth images are acquired using Blender, where a virtual camera follows the same trajectory and orientation as the subject's eye. Foothold locations in this eye centric reference frame are then calculated by intersecting eye center relative foothold location vectors with the eye's image plane. The CNN is a convolution deconvolutional architecture where the output is a probability map of foothold locations. The CNN is trained with outputs generated by placing gaussians with standard deviation [sigma] at the calculated foothold locations, and the corresponding depth image is used as in input. Performance is evaluated by computing the mean and median percentiles of the foothold locations in the output probability map.

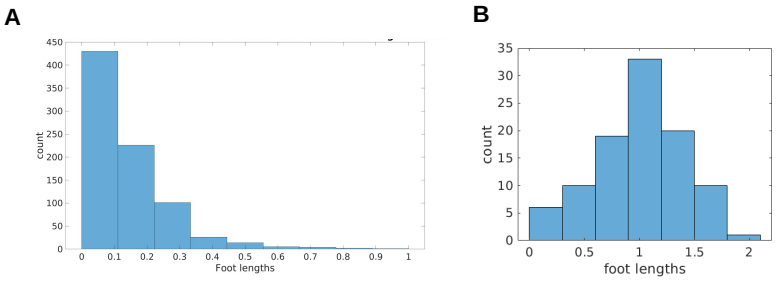


Figure 5.17: Foothold localization error. A. Distribution of between mesh errors of foothold location estimates (for the same subject traversal data. Foothold locations are estimated according to the same process described, but the terrain data used is interchanged, and the resulting different corresponding foothold locations are compared. B. Distribution of foothold estimate errors when compared to 'ground truth' foothold locations, obtained by manual annotation in the image frame, followed by projection of manually marked locations out onto the mesh depending on Meshroom's estimated camera pose

Chapter 6

Discussion

This dissertation focused on a complex terrain navigation dataset, and analyzed aspects of both the visual input and behavioral output of walkers. In this section I will discuss some of the noteworthy aspects of the datasets created in the process of this thesis. I will then highlight the key findings resulting from analysis of the data, and discuss their implications. Finally I will discuss future avenues of exploration that expand on this work, either directly or by incorporating aspects of this work into existing approaches.

6.1 Noteworthy characteristics of the new datasets beyond our results

6.1.1 Retinal motion approximation

Due to the adaptive nature of visual systems, knowing the statistics of naturally occurring visual signals can be helpful in guiding visual systems research [58]. Due to the high dependence of retinal motion patterns on behavioral variables like eye and head movements, as well as influences of the structure

of terrains, retinal motion inputs are difficult to measure. Doing so in the context of the behaviors that generate these motion signals is important for two reasons. The first is to accurately capture the statistics of the signals, since without use of accurate statistics of eye and body movements this is not possible. The second is that measurement of the natural behaviors that generate motion signals may be essential for understanding how the motion signal is used. This can be understood by examining task relevant features of the environment, and how these relate to the motion signals. The idea is that not all aspects of the input are necessarily important for guiding behaviors, and as a result an efficient visual system would not dedicate resources towards processing those aspects.

The value in this novel dataset is that it features both retinal motion patterns generated during natural locomotion, as well as information about the ongoing behavior that can be used to understand how the motion patterns might be used. This should be of use for understanding visual motion processing, beyond the scope of analysis done in this thesis.

6.1.2 3D terrain structure representation

The addition of reconstructed terrain to the dataset is of particular importance because it allows quantification of environment variables that may be important to the task. In order to better understand complex processing that takes place between environment variables and the eventual behavioral output, we need detailed measurements of both ends of the process. Reconstruction of the terrain in the same coordinate system as the recording of the ongoing behavior gives us measurements of terrain geometry like surface areas of different terrain features, their slants, and their heights. We also have quite accurate localization of foothold locations relative to these terrain features.

This novel dataset could provide a framework for testing quantitative models of terrain navigation, since it provides the necessary input and output cases that occurred with human subjects.

6.2 Key findings

Retinal motion stats:

1. Retinal motion during locomotion is characterized by an expansive pattern centered at the point of gaze, with rotations about the point of gaze and radially asymmetric speed gradients as a function of eccentricity. These effects have been quantified in the context of natural locomotion.

2. This pattern of motion is largely shaped by the pattern of head and eye movements through space, where the eye fixates points on the ground as the observer moves forwards.

3. Gaze location in the scene is relatively stable during this task, with initial fixation locations deviating only slightly over the course of fixations.

4. Gaze stabilization is accomplished by slow rotations of the eye in the head as the body moves forward. These slow counter-rotations are concentrated in the downward direction with a smaller angular spread compared to saccadic eye movements which are concentrated upwards, that is in the direction of travel. These patterns result from lateral movement of the head during the gait cycle in the case of stabilizing eye movements, and the somewhat wider lateral dispersion of gaze around footholds. This lateral dispersion is caused partly by changes in the direction of travel in irregular terrain.

5. Gaze direction relative to translation direction (horizontal gaze angle), and relative to gravity (vertical gaze angle), both have large effects on the pattern of motion directions, and speeds respectively. This effect has been

quantified in the context of natural movements.

6. Terrains of increasing complexity (as measured by normalized neural distance between model MT responses to terrain compared to simulated flat ground) induce lower vertical gaze angles. Thus this metric might be useful as a psychophysically relevant descriptor of terrain complexity and may predict terrain-dependant policies of walkers.

Gaze prediction:

1. Both vertical and horizontal gaze angle are partly predictable based on center of mass relative body position, with horizontal being more predictable.

2. The high predictability of the horizontal gaze angle is likely attributable to high correlation of body orientation and gaze direction, where the two are often aligned.

3. A convolutional neural network trained to predict gaze locations on the basis of visual features in a subject first-person perspective video can do so above chance, although with unexplained variance. This suggests that visual features are important for gaze allocation, but other factors are likely important.

Foothold finding:

1. Subjects have a preference for paths that have smaller average height change, although this is only a preference and walkers demonstrate considerable flexibility in path choices that include some steep slopes.

2. Foot location selection can be partially explained by likely height change when traveling to and from a given step location.

3. Subjects seem to trade off between paths with height changes, and curvy paths, with each subject having a subject-specific correlation value between the height variation of available straight paths, and the curvature of

their chosen path.

4. This correlation value seems to be negatively correlated with leg length, meaning longer legged subjects are less affected by variation in the height changes for the straight path.

5. A convolutional neural network trained to localize footholds based on retinocentric depth image inputs can do so substantially above chance, which suggests that depth image features are being used to select foothold locations, although unexplained variance suggests that other factors may also be important.

6.3 Discussion of findings

Each of the key findings has been discussed in previous chapters, so here I will provide some final thoughts regarding each. Particularly I will reiterate some points made in previous chapters, as well as explore some interesting future avenues of research.

6.3.1 Visual motion statistics

The average retinal motion pattern, which is expansive relative to the point of gaze, with rotations about the point of gaze, is reminiscent of Gibson's original conception of optic flow [51]. It is worth reiterating that the cause of this is not fixation of the focus of expansion as it appears due to head translation, but rather arises due to the combination of translation relative to the ground plane and fixation of stable locations on the ground. The full implications of this distinction between foveating the focus of expansion as opposed to this pattern of motion arising because of gaze behavior itself are not clear. However it is important because it changes the interpretation of how visual processing

systems may use the signal.

VOR is well understood in the context of lab experiments as well as treadmill studies. Additionally some work has been done measuring head movement statistics in real world contexts [124]. However the extent to which VOR is active and how effective it is during natural locomotion is not entirely clear. We were able to demonstrate that VOR makes the foveal image relatively stable, deviating minimally over the course of a fixation from the initially fixated location. This is important because it means that the area of interest remains foveated and high resolution as the subject acquires visual information. More precise measurements of VOR stability would be helpful since many of our analyses hinge on the assumption that the manual fixation adjustment better approximates the actual underlying eye movement than a noisy measurement from the eye tracker. The best way to do this is not clear however an experimental design and apparatus specifically geared towards measuring VOR during this behavior is probably warranted. Our measurement is limited by the frame rate of the camera (30Hz) as well as eye tracker noise. Different approaches that can address these issues, perhaps with a scene camera with a higher frame rate so that image slip can be measured at a faster time scale, would be useful in getting a better handle on this question.

Since locomotion constitutes a substantial component of human behavior, the difference in direction distributions between saccadic versus stabilizing eye movements has interesting implications for eye movement circuitry. In the same way that perceptual systems are adaptive for the natural environment, motor systems are adaptive. One might expect to see over-representation of neurons associated with saccadic or slow eye movements for particular directions, given that the demands of the task require different eye movement directions for the two. While classical psychophysics paradigms make probing

the visual system for differences in sensitivty that can confirm adaptation to natural signals possible, the way to probe the oculomotor system for such differences is not clear. Perhaps one could devise an experiment where subjects must make repeated saccades to targets at varying eccentric locations, and the accuracy of the saccades could be analyzed as a function of saccade direction. The same kind of process could be repeated for ground targets, where subjects must move in different directions relative to the targets while fixating them.

The large variation of the motion signal as a function of different gaze angles suggests that different gaze directions could provide context for an anticipated motion signal. For example, when looking at the ground forwards and to the left relative of the direction of translation, a clockwise spiral pattern with the same characteristics assuming a flat ground plane appears. This same pattern will always be present when this particular situation presents itself (see also Matthis et al, 2021). One way to explore the hypothesis of learned patterns of motion as a function of gaze angle relative to translation direction is with a virtual reality experiment with online eye and head tracking. For a given subject, their preferred gait could be recorded and then modelled in order to extrapolate their movements. They would walk repeatedly from one side of the room to the other, while fixating a point on the ground. The virtual room would be entirely black, except for the ground fixation point. As the subject approaches the fixation point, a motion pattern either consistent with or inconsistent with their current head trajectory and resulting stabilizing eye movement (which would need to be extrapolated based on their gait) is presented at low contrast, such that detection becomes difficult. One could probe the detection threshold of such a stimuli, and see whether differences between the consistent and inconsistent motion pattern exist. Differences in this threshold would suggest a different representation for the consistent pattern.

As discussed in Chapter 2, the patterns of flow speed and direction should be reflected in the patterns of speed and direction preferences in MST and possibly MT. Characteristics of the mean speed pattern like radially asymmetric eccentricity speed gradients, as well as banded regions of lower velocity due to ground plane geometry, could result in certain receptive field characteristics that exploit these properties in order to maximize information gain from the motion signal. Additionally, the dependence of these patterns and their spatial layout on gaze angle has implications for how the visual system might incorporate gaze angle information into its processing. There is evidence of eye in orbit modulation of firing rates in various visual cortical regions [82], [83], [84], which could be a reflection of different sensitivity and tuning as a function of gaze angle that is shaped by the effects of gaze angle on the input statistics.

Differences in model MT unit activity between actual motion patterns and matched simulated flat ground motion patterns provide a metric of terrain complexity that the visual system might easily extract, assuming there is a learned expected pattern for flat ground motion. This would be interesting to explore further. Ideally one could make use of an apparatus that allowed one to systematically vary terrain difficulty by adjusting the elevation variation, such that a full range of flat ground deviation values is swept over. This would provide a strong test of the idea that terrain complexity monotonically drives the average gaze angle lower. One could also simply collect data over a wider variety of terrains, or a stretch of terrain that has high variation in complexity along it and subdivide this into smaller sections. One could also manipulate the appearance of terrain such that this flat ground deviation signal is large, despite the terrain being relatively easily navigable by some other metric, although how to do this and what other metric one might use is not entirely

clear.

6.3.2 Gaze prediction

Gaze direction during this behavior was somewhat explainable by a linear regression model, using the center of mass relative body positions as input. It is difficult to assess the directionality of the relationship between the body positions and gaze directions, since it is not clear how one would influence the other. However one pattern was clear, which is that horizontal gaze angle and body orientation were related with horizontal gaze angle changes preceding body orientation changes. This had been observed in [125], where subjects direct gaze further along a curved path, preceding the turning of the body. This phenomenon explains the higher performance of horizontal angle regression. This does not mean that subjects fixate in location of future footholds that would be consistent with the current direction of the body, but rather that gaze identifies an approximate plan for the future path, and the body follows this plan. Other work shows that there is substantial dispersion of gaze (with a radius of about 10 deg) around the actual footholds.

The CNN result is difficult to interpret, however it at least demonstrates that there are visual features that identify where subjects direct gaze relative to the entire scene camera image. What is still unclear due to limitations of our performance metric is whether the CNN is finding features beyond those that identify the ground versus the sky or surrounding foliage. This would explain the above chance performance, and a CNN could likely learn the features necessary for differentiating these segments of the image. It may be possible to devise a method of pre-filtering these parts of the image out, and testing only within the region of walkable terrain. This would be more relevant to the research question of how is visual information used to guide foot placement.

Another approach could be attempting to understand what features the CNN is identifying and using through analysis of the learned weights and filters. These kinds of analyses were beyond the scope of the project, but would be an interesting way to revisit the work that was done in order to better understand what the CNN is doing.

6.3.3 Foothold selection

The difference in the distributions of average height change of the chosen paths, and the average height change of the randomly sampled paths, shows a preference of subjects for avoiding height changes when possible. However as noted previously, the overlap of the distributions makes this seem more like a preference than a hard constraint, although it is possible that there is some kind of hard constraint that is not captured by this method. Using the precomputed step locations and connecting steps between them, one might be able to discover step related metrics that increase the separation of these distributions. Again, access to some kind of deformable terrain apparatus could help in understanding this preference by adjusting the height change of different available paths to see how subject choice changes. This might also help elucidate what metric best captures subject decision making.

This kind of apparatus would also help in further exploring how subjects are trading off between straight paths and avoiding height changes. The correlation between straight path height change and the likelihood of taking a curvier path suggests that there is a trade off between the cost of a longer path compared to one with more height change, and this cost seems to be different for different subjects. This apparatus could be used to force subjects to decide between different paths, where parameters like tortuosity and average height change could be explicitly manipulated. Further understanding this

trade off, particularly how it relates to leg length, would be helpful. Leg length seems to decrease the cost of paths with more height change, enabling more straight paths, however the exact nature of this relationship is still unclear, and a metric for estimating energetic costs of the different path choices would be useful. This is where a biomechanical model of walking that accurately captures between subject differences could be usefully applied.

The results from our classification analysis of individual step locations show that average step slopes and average step slope derived features are predictive of foot placement. The predictive power of step slope derived features, which were computed across different scales, show that subjects may be incorporating information about foothold locations across multiple scales. This result could also be explored further using deformable terrain, however with a different approach where individual step locations are manipulated such that the average step slope to and from those locations changes. This would also be useful in better understanding the spatial extent that subjects are considering information over, since this could be explicitly manipulated.

The CNN result is also in need of further exploration. First, it would be useful to compare the performance of the CNN model in foothold location prediction to the classifiers, since it would give some indication of how much of subject choice comes from the fact that decisions are made from the subject's perspective (captured by the CNN) rather than a more omniscient one (classifier). One could derive some method of doing so, perhaps by projecting the CNN scores for each pixel location out onto the terrain, and then using the same performance metrics as the classifier. There are still some issues since the classifier approach engineered features that incorporate future step information in ways that are different from the depth image features the CNN uses, however it would be a start. Additionally, one might incorporate bio-

logical constraints into the model to see how this affects performance. The CNN assumes equal stereoacuity as a function of eccentricity, which is known to not be the case in humans ([122]). This degradation with eccentricity, and other known constraints on human visual systems could be incorporated into the model in order to test the degree to which particular constraints shape walker foot placement. The constraints could be systematically varied to see how they affect the predictions.

Another important insight from Chapter 5 was that paths appear to be planned over at least a 5 step window. This is consistent with the possibility that subjects choose a terrain-dependant policy, which may specify walking speed and step length. This plan could then be modulated on a more local basis to modulate step location and walking speed. The terrain complexity metric explored in Chapter 2, might then be useful for predicting gait parameters as described above in Section 6.3.1.

Bibliography

[1] Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

[2] Julie Epelboim, Robert M Steinman, Eileen Kowler, Zygmunt Pizlo, Casper J Erkelens, and Han Collewijn. Gaze-shift dynamics in two kinds of sequential looking tasks. *Vision research*, 37(18):2597–2607, 1997.

[3] Michael Land, Neil Mennie, and Jennifer Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999.

[4] Michael F Land and Mary Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41(25-26):3559–3565, 2001.

[5] Mary M Hayhoe, Anurag Shrivastava, Ryan Mruczek, and Jeff B Pelz. Visual memory and motor planning in a natural task. *Journal of vision*, 3(1):6–6, 2003.

[6] Laurence T Maloney and Hang Zhang. Decision-theoretic models of visual perception and action. *Vision research*, 50(23):2362–2374, 2010.

[7] David W Franklin and Daniel M Wolpert. Computational mechanisms of sensorimotor control. *Neuron*, 72(3):425–442, 2011.

[8] Daniel M Wolpert and Michael S Landy. Motor control is decision-making. *Current opinion in neurobiology*, 22(6):996–1003, 2012.

[9] Mary M Hayhoe. Vision and action. *Annual review of vision science*, 3:389–413, 2017.

[10] Michael F Land and David N Lee. Where we look when we steer. *Nature*, 369(6483):742–744, 1994.

[11] Michael F Land and Peter McLeod. From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, 3(12):1340–1345, 2000.

[12] Vidhya Navalpakkam, Christof Koch, Antonio Rangel, and Pietro Perona. Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences*, 107(11):5232–5237, 2010.

[13] Alexander C Schütz, Julia Trommershäuser, and Karl R Gegenfurtner. Dynamic integration of information about salience and value for saccadic eye movements. *Proceedings of the National Academy of Sciences*, 109(19):7547–7552, 2012.

[14] Moritz Kassner, William Patera, and Andreas Bulling. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pages 1151–1160, 2014.

[15] Aftab E Patla. Understanding the roles of vision in the control of human locomotion. *Gait & posture*, 5(1):54–69, 1997.

[16] William H Warren Jr. Visually controlled locomotion: 40 years later. *Ecological Psychology*, 10(3-4):177–219, 1998.

[17] Markus Lappe, Frank Bremmer, Martin Pekel, Alexander Thiele, and Klaus-Peter Hoffmann. Optic flow processing in monkey sts: a theoretical and experimental approach. *Journal of Neuroscience*, 16(19):6265–6285, 1996.

[18] Markus Lappe, Frank Bremmer, and AV Van den Berg. Perception of self-motion from visual flow. *Trends in cognitive sciences*, 3(9):329–336, 1999.

[19] William H Warren. Action-scaled information for the visual control of locomotion. In *Closing the gap*, pages 261–296. Psychology Press, 2007.

[20] Li Li and Joseph CK Cheng. Visual strategies for the control of steering toward a goal. *Displays*, 34(2):97–104, 2013.

[21] Kenneth H Britten. Mechanisms of self-motion perception. *Annu. Rev. Neurosci.*, 31:389–410, 2008.

[22] William H Warren and Brett R Fajen. From optic flow to laws of control. In *Optic flow and beyond*, pages 307–337. Springer, 2004.

[23] Jonathan Samir Matthis, Karl S Muller, Kathryn Bonnen, and Mary M Hayhoe. Retinal optic flow during natural locomotion. *BioRxiv*, pages 2020–07, 2021.

[24] Jean Pailhous, Anne-Marie Ferrandez, Michelangelo Flückiger, and Bernard Baumberger. Unintentional modulations of human gait by optical flow. *Behavioural brain research*, 38(3):275–281, 1990.

[25] Jürgen Konczak. Effects of optic flow on the kinematics of human gait: a comparison of young and older adults. *Journal of motor behavior*, 26(3):225–236, 1994.

[26] W Zijlstra, AWF Rutgers, AL Hof, and TW Van Weerden. Voluntary and involuntary adaptation of walking to temporal and spatial constraints. *Gait & Posture*, 3(1):13–18, 1995.

[27] Ying-hui Chou, Robert C Wagenaar, Elliot Saltzman, J Erik Giphart, Daniel Young, Rosa Davidsdottir, and Alice Cronin-Golomb. Effects of optic flow speed and lateral flow asymmetry on locomotion in younger and older adults: a virtual reality study. *Journals of Gerontology: Series B*, 64(2):222–231, 2009.

[28] Dimitrios Kastavelis, Mukul Mukherjee, Leslie M Decker, and Nicholas Stergiou. The effect of virtual reality on gait variability. 2010.

[29] Jessica R Berard, Joyce Fung, Bradford J McFadyen, and Anouk Lamontagne. Aging affects the ability to use optic flow in the control of heading during locomotion. *Experimental brain research*, 194(2):183–190, 2009.

[30] Anouk Lamontagne, Joyce Fung, Bradford J McFadyen, and Jocelyn Faubert. Modulation of walking speed by changing optic flow in persons with stroke. *Journal of neuroengineering and rehabilitation*, 4(1):1–8, 2007.

[31] Thomas Prokop, Martin Schubert, and Wiltrud Berger. Visual influence

on human locomotion modulation to changes in optic flow. *Experimental brain research*, 114(1):63–70, 1997.

[32] William H Warren, Bruce A Kay, and Emre H Yilmaz. Visual control of posture during walking: functional specificity. *Journal of Experimental Psychology: Human Perception and Performance*, 22(4):818, 1996.

[33] Hendrik Reimann, Tyler Fettrow, Elizabeth D Thompson, and John J Jeka. Neural control of balance during walking. *Frontiers in Physiology*, 9:1271, 2018.

[34] Charles J Duffy and Robert H Wurtz. Response of monkey mst neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience*, 15(7):5192–5208, 1995.

[35] Yong Gu, Paul V Watkins, Dora E Angelaki, and Gregory C DeAngelis. Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. *Journal of Neuroscience*, 26(1):73–85, 2006.

[36] Hyung-Kyu Kang, Young Kim, Yijung Chung, and Sujin Hwang. Effects of treadmill training with optic flow on balance and gait in individuals following stroke: randomized controlled trials. *Clinical rehabilitation*, 26(3):246–255, 2012.

[37] Trevor Drew, Stephen Prentice, and Bénédicte Schepens. Cortical and brainstem control of locomotion. *Progress in brain research*, 143:251–261, 2004.

[38] Patrick J Whelan. Control of locomotion in the decerebrate cat. *Progress in neurobiology*, 49(5):481–515, 1996.

[39] N Alberto Borghese, L Bianchi, and F Lacquaniti. Kinematic determinants of human locomotion. *The Journal of physiology*, 494(3):863–879, 1996.

[40] Ashutosh Kharb, Vipin Saini, YK Jain, and Surender Dhiman. A review of gait cycle and its parameters. *IJCEM International Journal of Computational Engineering & Management*, 13:78–83, 2011.

[41] Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Yokoi, and Hirohisa Hirukawa. The 3d linear inverted pendulum mode: A simple modeling for a biped walking pattern generation. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, volume 1, pages 239–246. IEEE, 2001.

[42] A Bertrand Arsenault, David A Winter, and Ronald G Marteniuk. Treadmill versus walkway locomotion in humans: an emg study. *Ergonomics*, 29(5):665–676, 1986.

[43] Darcy S Reisman, Robert Wityk, Kenneth Silver, and Amy J Bastian. Locomotor adaptation on a split-belt treadmill can improve walking symmetry post-stroke. *Brain*, 130(7):1861–1872, 2007.

[44] Darcy S Reisman, Heather McLean, Jennifer Keller, Kelly A Danks, and Amy J Bastian. Repeated split-belt treadmill training improves poststroke step length asymmetry. *Neurorehabilitation and neural repair*, 27(5):460–468, 2013.

[45] Alexandra S Voloshina, Arthur D Kuo, Monica A Daley, and Daniel P Ferris. Biomechanics and energetics of walking on uneven terrain. *Journal of Experimental Biology*, 216(21):3963–3970, 2013.

[46] Jenny A Kent, Joel H Sommerfeld, and Nicholas Stergiou. Changes in human walking dynamics induced by uneven terrain are reduced with ongoing exposure, but a higher variability persists. *Scientific reports*, 9(1):1–9, 2019.

[47] Jun Morimoto, Gen Endo, Jun Nakanishi, and Gordon Cheng. A biologically inspired biped locomotion strategy for humanoid robots: Modulation of sinusoidal patterns by a coupled oscillator model. *IEEE Transactions on Robotics*, 24(1):185–191, 2008.

[48] Azhar Aulia Saputra, János Botzheim, Indra Adji Sulistijono, and Naoyuki Kubota. Biologically inspired control system for 3-d locomotion of a humanoid biped robot. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(7):898–911, 2015.

[49] Marie-Claude L Grisé, Christiane Gauthier-Gagnon, and Georges G Martineau. Prosthetic profile of people with lower extremity amputation: conception and design of a follow-up questionnaire. *Archives of Physical Medicine and Rehabilitation*, 74(8):862–870, 1993.

[50] Clive Fraser. Slam, sfm and photogrammetry: What's in a name. In *Proceedings of the ISPRS Technical Comission II: Symposium*, 2018.

[51] James J Gibson. The perception of the visual world. 1950.

[52] Jan J Koenderink. Optic flow. *Vision research*, 26(1):161–179, 1986.

[53] Richard T. Born and David C. Bradley. Structure and function of visual area MT. *Annual Review of Neuroscience*, 28:157–189, 2005.

[54] C Daniel Salzman, Chieko M Murasugi, Kenneth H Britten, and William T Newsome. Microstimulation in visual area mt: effects

on direction discrimination performance. *Journal of Neuroscience*, 12(6):2331–2355, 1992.

[55] William T Newsome and Edmond B Pare. A selective impairment of motion perception following lesions of the middle temporal visual area (mt). *Journal of Neuroscience*, 8(6):2201–2211, 1988.

[56] Island Survivors, Gregory C DeAngelis, Bruce G Cumming, and William T Newsome. Cortical area MT and the perception of stereoscopic depth. *Nature*, 394(6694):677–680, 1998.

[57] Benedict Wild and Stefan Treue. Primate extrastriate cortical area mst: a gateway between sensation and cognition. *Journal of Neurophysiology*, 125(5):1851–1882, 2021.

[58] Wilson S. Geisler. Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, 59:167–192, 2008.

[59] Hilary W. Heuer and Kenneth H. Britten. Optic Flow Signals in Extrastriate Area MST: Comparison of Perceptual and Neuronal Sensitivity. *Journal of Neurophysiology*, 91(3):1314–1326, 2004.

[60] A. T. Smith, M. B. Wall, A. L. Williams, and K. D. Singh. Sensitivity to optic flow in human cortical areas MT and MST. *European Journal of Neuroscience*, 23(2):561–569, 2006.

[61] Ian E. Holliday and Timothy S. Meese. Optic flow in human vision: MEG reveals a foveo-fugal bias in V1, specialization for spiral space in hMSTs, and global motion sensitivity in the IPS. *Journal of Vision*, 8(10):1–24, 2008.

[62] Dirk Calow and Markus Lappe. Local statistics of retinal optic flow for self-motion through natural sceneries. *Network: Computation in Neural Systems*, 18(4):343–374, 2007.

[63] Dirk Calow and Markus Lappe. Efficient encoding of natural optic flow. *Network: Computation in Neural Systems*, 19(3):183–212, 2008.

[64] Jonathan Samir Matthis, Jacob L. Yates, and Mary M. Hayhoe. Gaze and the Control of Foot Placement When Walking in Natural Terrain. *Current Biology*, 0(0):1–10, 2018.

[65] Pieter Blignaut. Fixation identification: The optimum threshold for a dispersion algorithm. *Attention, Perception, & Psychophysics*, 71(4):881–895, 2009.

[66] Marcus Nyström and Kenneth Holmqvist. An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior research methods*, 42(1):188–204, 2010.

[67] Harry J Wyatt. Detecting saccades with jerk. *Vision research*, 38(14):2147–2153, 1998.

[68] Jonas Goltz, Michael Grossberg, and Ronak Etemadpour. Exploring simple neural network architectures for eye movement classification. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–5, 2019.

[69] Ethel Matin. Saccadic suppression: a review and an analysis. *Psychological bulletin*, 81(12):899, 1974.

[70] James M McFarland, Adrian G Bondy, Richard C Saunders, Bruce G Cumming, and Daniel A Butts. Saccadic modulation of stimulus pro-

cessing in primary visual cortex. *Nature communications*, 6(1):1–14, 2015.

[71] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE international conference on computer vision*, pages 1385–1392, 2013.

[72] Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 3, pages 2301–2306. IEEE, 2004.

[73] Jan J Koenderink and Andrea J van Doorn. Local structure of movement parallax of the plane. *JOSA*, 66(7):717–723, 1976.

[74] Michael Steven Graziano, Richard A Andersen, and Robert J Snowden. Tuning of mst neurons to spiral motions. *Journal of Neuroscience*, 14(1):54–67, 1994.

[75] AliceVision. Meshroom: A 3D reconstruction software., 2018.

[76] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 2021.

[77] Hugh Christopher Longuet-Higgins and Kvetoslav Prazdny. The interpretation of a moving retinal image. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 208(1173):385–397, 1980.

[78] Michael Beyeler, Nikil Dutt, and Jeffrey L Krichmar. 3d visual response properties of mstd emerge from an efficient, sparse population code. *Journal of Neuroscience*, 36(32):8399–8415, 2016.

[79] Daniel J Felleman and Jon H Kaas. Receptive-field properties of neurons in middle temporal visual area (mt) of owl monkeys. *Journal of Neurophysiology*, 52(3):488–513, 1984.

[80] Emilio Salinas. Context-dependent selection of visuomotor maps. *BMC neuroscience*, 5(1):1–22, 2004.

[81] Sima Mistry and Jose L Contreras-Vidal. Learning multiple visuomotor transformations: adaptation and context-dependent recall. *Motor control*, 8(4):534–546, 2004.

[82] F Bremmer, UJ Ilg, A Thiele, C Distler, and K-P Hoffmann. Eye position effects in monkey cortex. i. visual and pursuit-related activity in extrastriate areas mt and mst. *Journal of neurophysiology*, 77(2):944–961, 1997.

[83] Richard A Andersen, R Martyn Bracewell, Shabtai Barash, James W Gnadt, and Leonardo Fogassi. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *Journal of Neuroscience*, 10(4):1176–1196, 1990.

[84] Driss Boussaoud, Christophe Jouffrais, and Frank Bremmer. Eye position effects on the neuronal activity of dorsal premotor cortex in the macaque monkey. *Journal of neurophysiology*, 80(3):1132–1150, 1998.

[85] Pawel Zmarz and Georg B Keller. Mismatch receptive fields in mouse visual cortex. *Neuron*, 92(4):766–772, 2016.

[86] Ziad M Hafed and Chih-Yang Chen. Sharper, stronger, faster upper visual field representation in primate superior colliculus. *Current Biology*, 26(13):1647–1658, 2016.

[87] Sylvia Schröder, Nicholas A. Steinmetz, Michael Krumin, Marius Pachitariu, Matteo Rizzi, Leon Lagnado, Kenneth D. Harris, and Matteo Carandini. Retinal outputs depend on behavioural state. *bioRxiv*, 2019.

[88] Bernard Mariust Hart and Wolfgang Einhauser. Mind the step: Complementary effects of an implicit task on eye and head movements in real-life gaze allocation. *Experimental Brain Research*, 223(2):233–249, 2012.

[89] Matthew H Tong, Oran Zohar, and Mary M Hayhoe. Control of gaze while walking: task structure, reward, and uncertainty. *Journal of vision*, 17(1):28–28, 2017.

[90] F Javier Domínguez-Zamora, Shaila M Gunn, and Daniel S Marigold. Adaptive gaze strategies to reduce environmental uncertainty during a sequential visuomotor behaviour. *Scientific reports*, 8(1):1–13, 2018.

[91] Ruohan Zhang, Shun Zhang, Matthew H Tong, Yuchen Cui, Constantin A Rothkopf, Dana H Ballard, and Mary M Hayhoe. Modeling sensory-motor decisions in natural behavior. *PLoS computational biology*, 14(10):e1006518, 2018.

[92] H Dietrich and M Wuehr. Strategies for gaze stabilization critically depend on locomotor speed. *Neuroscience*, 408:418–429, 2019.

[93] Daniel Panfili, Karl Muller, Jon Matthis, and Mary Hayhoe. Gaze distribution around footholds in rough terrain. *Journal of Vision*, 21(9):2914–2914, 2021.

[94] Delphine Bernardin, Hideki Kadone, Daniel Bennequin, Thomas Sugar,

Mohamed Zaoui, and Alain Berthoz. Gaze anticipation during human locomotion. *Experimental brain research*, 223(1):65–78, 2012.

[95] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[96] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2161–2168. Ieee, 2006.

[97] Shiqi Li, Chi Xu, and Ming Xie. A robust o (n) solution to the perspective-n-point problem. *IEEE transactions on pattern analysis and machine intelligence*, 34(7):1444–1450, 2012.

[98] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[99] Chris Engels, Henrik Stewénius, and David Nistér. Bundle adjustment rules. *Photogrammetric computer vision*, 2(32), 2006.

[100] Hang Si and Klaus Gärtner. 3d boundary recovery by constrained delaunay tetrahedralization. *International Journal for Numerical Methods in Engineering*, 85(11):1341–1364, 2011.

[101] Michal Jancosek and Tomás Pajdla. Multi-view reconstruction preserving weakly-supported surfaces. In *CVPR 2011*, pages 3121–3128. IEEE, 2011.

[102] Michal Jancosek and Tomas Pajdla. Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces. *International scholarly research notices*, 2014, 2014.

[103] Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1124–1137, 2004.

[104] Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 2(4):217–236, 1983.

[105] Daniel Girardeau-Montaut. Cloudcompare. *France: EDF R&D Telecom ParisTech*, 2016.

[106] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992.

[107] Daniel Panfili, Karl Muller, and Mary Hayhoe. A real-time model of retinal stimulation in virtual environments. *Journal of Vision*, 20(11):1566–1566, 2020.

[108] Jason P Gallivan, Craig S Chapman, Daniel M Wolpert, and J Randall Flanagan. Decision-making in sensorimotor control. *Nature Reviews Neuroscience*, 19(9):519–534, 2018.

[109] David J Clark. Automaticity of walking: functional significance, mechanisms, measurement and rehabilitation strategies. *Frontiers in human neuroscience*, 9:246, 2015.

[110] Jessica C Selinger, Shawn M O'Connor, Jeremy D Wong, and J Maxwell Donelan. Humans can continuously optimize energetic cost during walking. *Current Biology*, 25(18):2452–2456, 2015.

[111] James M Finley, Amy J Bastian, and Jinger S Gottschall. Learning to be economical: The energy cost of walking tracks motor adaptation. *Journal of Physiology*, 591(4):1081–1095, 2013.

[112] David V. Lee and Sarah L. Harris. Linking gait dynamics to mechanical cost of legged locomotion. *Frontiers Robotics AI*, 5(OCT):1–11, 2018.

[113] Chase G. Rock, Vivien Marmelat, Jennifer M. Yentes, Ka Chun Siu, and Kota Z. Takahashi. Interaction between step-to-step variability and metabolic cost of transport during human walking. *Journal of Experimental Biology*, 221(22), 2018.

[114] Hikaru Yokoyama, Koji Sato, Tetsuya Ogawa, Shin Ichiro Yamamoto, Kimitaka Nakazawa, and Noritaka Kawashima. Characteristics of the gait adaptation process due to split-belt treadmill walking under a wide range of right-left speed ratios in humans. *PLoS ONE*, 13(4):1–14, 2018.

[115] Shawn M. O'Connor, Henry Z Xu, and Arthur D Kuo. Energetic cost of walking with increased step variability. *Gait and Posture*, 36(1):102–107, 2012.

[116] Koren Gast, Rodger Kram, and Raziel Riemer. Preferred walking speed on rough terrain: Is it all about energetics? *Journal of Experimental Biology*, 222(9), 2019.

[117] Jeff B Pelz and Constantin Rothkopf. Oculomotor behavior in natu-

ral and man-made environments. In *Eye Movements*, pages 661–676. Elsevier, 2007.

[118] Tom Foulsham, Esther Walker, and Alan Kingstone. The where, what and when of gaze allocation in the lab and the natural environment. *Vision research*, 51(17):1920–1931, 2011.

[119] Bernard Marius 't Hart, Hannah Claudia Elfriede Fanny Schmidt, Ingo Klein-Harmeyer, and Wolfgang Einhäuser. Attention in natural scenes: contrast affects rapid visual processing and fixations alike. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1628):20130067, 2013.

[120] J. M. Kinsella-Shaw, Brian Shaw, and M. T. Turvey. Perceiving "Walk-on-able" Slopes. *Ecological Psychology*, 4(4):223–239, dec 1992.

[121] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

[122] T Shipley and M Popp. Stereoscopic acuity and retinal eccentricity. *Ophthalmic Research*, 3(4):251–255, 1972.

[123] FFmpeg team. FFmpeg, 2021.

[124] Jérome Carriot, Mohsen Jamali, Maurice J Chacron, and Kathleen E Cullen. Statistics of the vestibular input experienced during natural self-motion: implications for neural processing. *Journal of Neuroscience*, 34(24):8347–8357, 2014.

[125] Colas Nils Authié, Pauline M Hilt, Alain Berthoz, Daniel Bennequin, et al. Differences in gaze anticipation for locomotion with and without vision. *Frontiers in human neuroscience*, 9:312, 2015.

Vita

Karl S. Muller was born in Mexico City, Mexico in 1993. He grew up in various countries in the Americas, Europe, and Africa. In 2012, he began attending the University of Texas at Austin for his undegraduate studies, obtaining a Bachelor's of Science in Neuroscience. During his junior and senior years as an undergraduate, he worked as an undergraduate research assistant in Dr. Mary Hayhoe's Vision, Cognition, and Action Virtual Reality Lab, on the human locomotion dataset that would later become the basis for this thesis. Through work on this dataset, he cultivated an interest in quantifying and analyzing natural behavior, and decided to pursue a PhD in Neuroscience at the Institute for Neuroscience at the University of Texas at Austin in 2017. Here he became a graduate student and researcher in the Hayhoe lab, conducting thesis work focusing on visually guided locomotion over complex terrain.

Permanent Address: karl.muller@utexas.edu

This dissertation was typeset with LaTeX 2$_\varepsilon$[1] by the author.

[1] LaTeX 2$_\varepsilon$ is an extension of LaTeX. LaTeX is a collection of macros for TeX. TeX is a trademark of the American Mathematical Society. The macros used in formatting this dissertation were written

by Dinesh Das, Department of Computer Sciences, The University of Texas at Austin, and extended by Bert Kay, James A. Bednar, and Ayman El-Khashab.