INTEGRATING GAZE, IMAGE ANALYSIS, AND BODY TRACKING: FOOTHOLD SELECTION DURING LOCOMOTION

A PREPRINT

Karl Muller
Center for Perceptual Systems
University of Texas at Austin
karl.muller@utexas.edu

Dan Panfili
Center for Perceptual Systems
University of Texas at Austin
dan.panfili@utexas.edu

Jonathan S. Matthis
Department of Biology
Northeastern University
j.matthis@northeastern.edu

Kathryn Bonnen
Department of Optometry
Indiana Universtity
k.bonnen@indiana.edu

Mary Hayhoe
Center for Perceptual Systems
University of Texas at Austin
hayhoe@utexas.edu

August 18, 2023

ABSTRACT

Relatively little is known about the way vision is use to guide locomotion in the natural world. What visual features are used to choose paths in natural complex terrain? How do walkers trade off different costs such as getting to the goal, minimizing energy, and satisfying stability constraints? To answer these questions, it is necessary to monitor not only the eyes and the body, but also to represent the three dimensional structure of the terrain. We used photogrammetry techniques to do this, and found substantial regularities in the choice of paths. Walkers avoid paths that involve changes in height and choose more circuitous and flatter paths. This stable tradeoff is related to the walker's leg length and reflects both energetic and stability constraints. Gaze data and path choices suggest that subjects take into account the terrain approximately 5 steps ahead, and so are planning routes as well as particular footplants. Such planning ahead allows the minimization

of energetic costs. Thus locomotor behavior in natural environments is controlled by decision mechanisms that attempt to optimize for multiple factors in the context of well-calibrated sensory and motor internal models.

1 Introduction

Actions shape the sensory input, and in turn, the sensory input determines actions. Consequently an understanding of even simple actions in the natural world requires that we monitor both the actions and the sensory input. While the technology for monitoring gaze and body during natural behavior is readily available, measurement of the visual stimulus has been limited. In this paper we use computer vision techniques to reconstruct 3D representations of the visual scene. Together with gaze data, this allows specification of the visual information that is used in action decisions in the context of a very basic human behavior, namely, locomotion.

Locomotion on flat ground requires very little visual guidance, and can be accomplished with minimal cortical control [1]. Locomotion over complex terrain, however, requires coordination between brainstem-mediated central pattern generators, and cortically-mediated modifications of leg and foot trajectories [2]. 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, or in laboratory settings, where it has been shown that humans converge towards energetic optima. Walkers adopt a preferred gait that constitutes an energetic minimum given their own biomechanics [3], [4], [5], [6]. The parameters over which this optimization principle holds include walking speed, step frequency, step distance, and step width [7], [8], [9].

There are a number of problems in generalizing these findings. Natural visually guided behaviors can be characterized as a sequence of complex sensorimotor decisions [10], [11], [12]. These decisions will be shaped by more complex cost functions than in treadmill walking, together with context-specific sensory and motor uncertainty. Consequently it is unclear how the optimization principles described above might play out in natural locomotion. Existing models focus primarily on optimization of the preferred gait cycle with respect to biomechanical factors. However, locomotion over rough terrain depends on both the biomechanics of the walker and visual information about the structure of the environment. When the terrain is more complex, walkers need to find stable footplants. Natural

behavior will also introduce factors such as the need to reach a goal or attend to the navigational context. Thus the visual demands in natural locomotion will be more complex. Another complication is that, in natural behavior, the movements of the observer shape the visual input. In locomotion, subjects stabilize the image as the body moves forward and then make rapid saccadic eye movements to the next location [13]. Consequently to understand what visual information is used for action choices we need to have a description of the terrain during these periods of stable gaze. Together with gaze location, this allows computation of the retinal image sequence. In this work we use computer vision algorithms to create a 3D representation of the terrain structure in addition to measuring eye and body movements. This allows a more complete characterization of the visuo-motor control loop in locomotion. In particular, it allows us to analyze the retinal image features that underlie the choice of footholds.

In this paper we ask how vision is used to identify viable footholds in natural environments, and what specific terrain features influence behavior. Ambiguity about the cost function creates a larger space of possible actions the subject can take, and is it not known how walkers use visual 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 patterns with the demands of the terrain [14], [15], [16], but foot placement was not measured, so it was not possible to analyze the relation between gaze and foot placement. Recent work by Matthis and by Bonnen and colleagues [17], [18] integrated gaze and body measurements in natural walking, but relied on the assumption of a flat ground plane to calculate gaze and foothold location. They 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 addition, 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. 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 that could be linked to gaze and gait data to shed light on how subjects use visual information to choose paths.

To construct a numerical representation of the environment, we took advantage of recently developed photogrammetry algorithms that use the sequence of camera views to reconstruct the 3 dimensional terrain structure, along with a representation of the 6 DOF camera path. These algorithms take advantage of the sequence of

images from a head mounted camera that are revealed by viewpoint changes in order to recover the depth structure of the scene. 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's body. This allowed much more accurate estimates of gaze and footholds on the ground surface than has previously been possible, 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. This strategy was also used in our recent work detailing the statistics of retinal motion using the same data set [13].

We first demonstrate that there are in fact regularities in the paths chosen by subjects when walking over the same terrain on a different occasion, and also that there are similarities between subjects in the chosen paths. Thus paths were not completely random, and must reflect some optimization principles. We developed techniques for comparing the features of chosen paths to other viable paths and found that subjects choose paths where the average height change in a short segment is less than neighboring possible paths, reflecting the role of energetic costs even in rough terrain. We also found evidence that height changes are evaluated over a set of several future steps, indicating planning of step sequences. Circuitous paths also incur energetic cost and we found that subjects deviate more from straight paths as height changes increase in a way that depends on leg length. Thus natural locomotion reflects planned sequences of complex sensory-motor decisions that are orchestrated by the local terrain features and the walker's individual cost function.

Results

Walking data were collected over a range of different terrains for 9 subjects collected both in Austin and Berkeley (see Methods). The most rocky segments of the terrain were selected for analysis here. Eye and body movements of participants were recorded using a Pupil Labs Core mobile binocular eye tracker and the Motion Shadow full body motion capture system.

Terrain Reconstruction: In order to estimate both the environmental structure and the relative camera position from the head mounted video, we used software package called Meshroom [19], that combines multiple image processing and computer vision algorithms in order to reconstruct the environment from an image sequence. The different viewpoints are integrated to create a depth map as in structure-from-motion or motion parallax calculations.This allowed a quantitative description of the 3D structure of the environment, together with RGB values, as well as estimates of head position relative to the environment. We then aligned the head orientation and position measured by the motion capture system with that of

the camera orientation and position estimated by Meshroom. In previous work [17], estimates of future foot locations relative to current body location were subject to noise resulting from drift in the Motion Shadow's IMU signal. By pinning the head position estimated from the IMU to the Meshroom estimates, we were able to eliminate this drift by fixing the body's reference frame to that of the environment. This alignment is illustrated in Figure 1. The elimination of this source of error and the parallax error from having to assume a flat ground plane allowed greatly improved estimates of both gaze and foot position, and the 3D representation allows an evaluation of the features of ground structure that might be involved in foothold selection. Estimates of the accuracy of the reconstructions are discussed in the Methods section. An example comparison of the original image and the reconstructed terrain are shown in Figure 2. A video of the walker situated in the terrain, together with an indication of the foot and gaze locations can be seen in the following video at https://youtu.be/TzrA_iEtj1s.

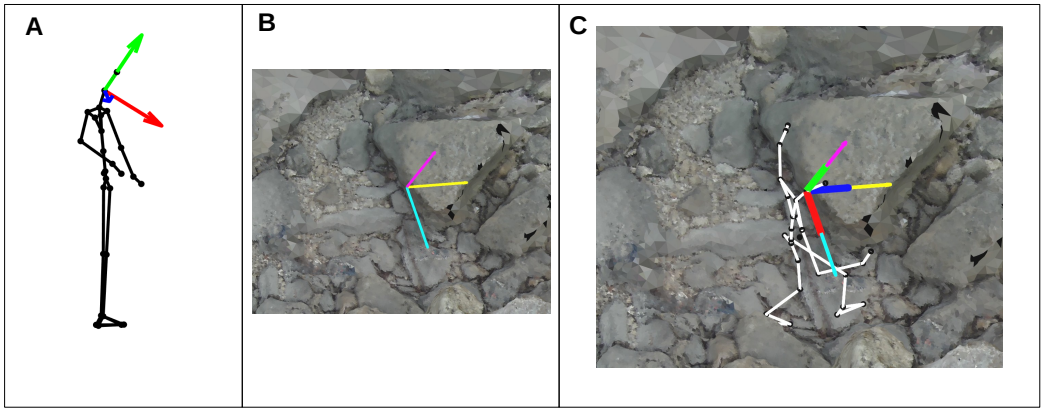

Figure 1: Alignment of motion capture data to environmental coordinates. Motion capture coordinate system (A) is aligned with Meshroom coordinate system (B) via a single rotation and translation that minimizes error between the mocap's camera axes and Meshroom's camera axes (C). The motion capture skeleton is then scaled such in order to minimize the distance from the closest point on the mesh to the locations of footfall frames, evaluated at each footfall frame. This scale factor is then applied to the motion capture data at every frame.

Figure 2: 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

Path Consistency

The first issue to address was whether there was any consistency in the paths chosen by different subjects and on different occasions. Without such regularities, we would be unlikely to find stable properties of the environment that determined foothold choice. Although we did not quantify this, we show some examples of both convergence and divergence of the chosen paths. Examples from the Berkeley data are shown in Figure 3. The degree of convergence between the 7 different subjects suggests that there are some common visual features that underlie path choices. Figure 4 shows the paths for the Austin data set. In Figure 4, the two different subjects are shown in different colors, each for 3 repetitions of the path, in both forward and backward traverses. Parts A and B of the Figure show paths in opposite directions along the same stretch of ground. The Figure shows regions where the different subjects take the same path and where the same path is taken on different occasions. The paths are not all identical, indicating that foothold choices are not tightly constrained, but considerable regularity exists across both repetitions and subjects.

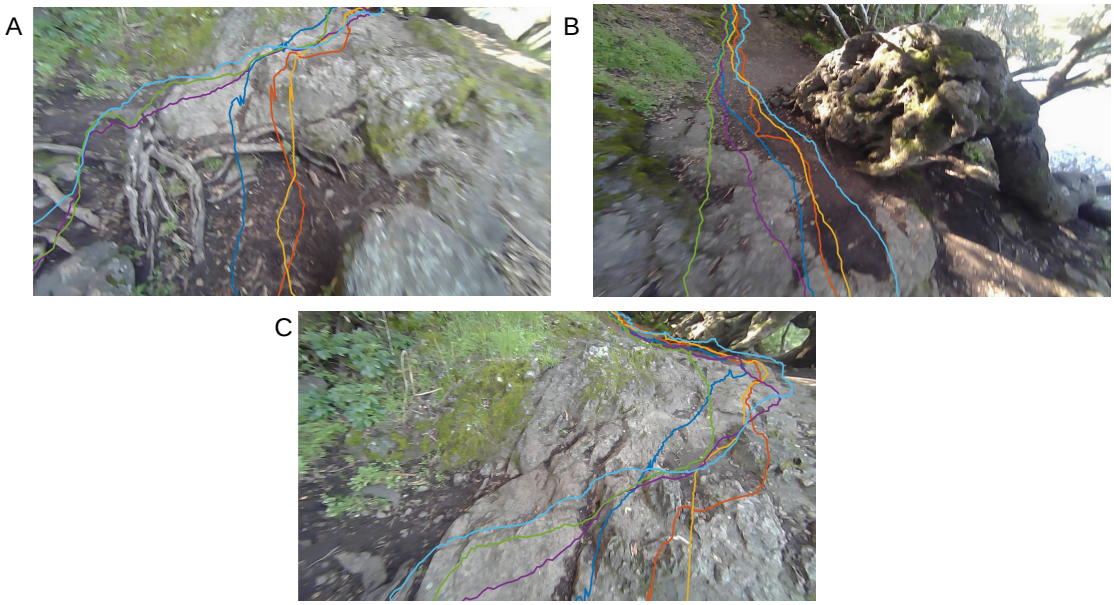

Figure 3: Examples of path convergence and divergence. The colors indicate different subjects. In (A), subjects diverge by choosing two different routes around a root, but then converge again. in (B) subjects paths converge to avoid the large outcrop. In (C) subject paths converge around a mossy section of a large rock.

Gaze location relative to 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

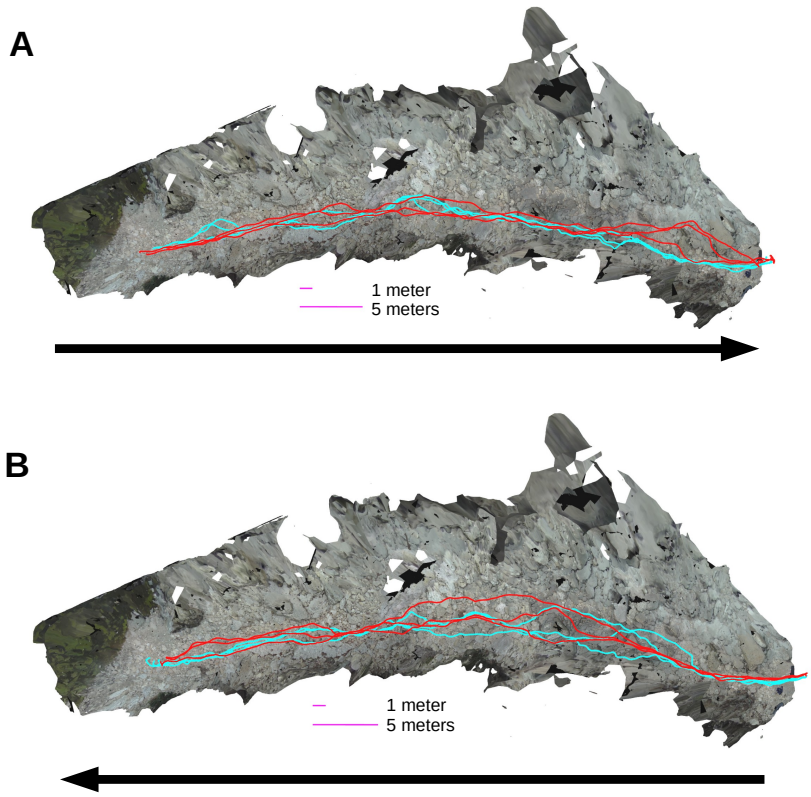

Figure 4: Overhead view of Austin data. Subjects walking from left to right (A) or right to left (B). Different colors correspond to different subjects, each traversing in each direction 3 times.

was the concern of our earlier investigation [17] , which showed that fixations were clustered most densely 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 a forthcoming paper. We show in that study that fixations are centered on future footholds, most commonly around 3 steps ahead, but ranging between 1 and 5 steps. The distributions around the footholds have a standard deviation of 25-30 cm which would correspond to roughly 5 deg of visual angle. Thus subjects look close to the locations where the feet are placed. This can be seen by the close relationship of the green dots showing gaze locations in Figure 5 and the pink dots showing foot placement. However, gaze does not always fall on future footholds. When the upcoming terrain is undesirable, walkers must change direction. Some examples of these direction changes are also visible in Figure 5. The blue lines show the current direction of the body, and the blue dots show gaze falling off the subsequent path. We infer that in these cases the walker changes direction to avoid some aspect of the upcoming terrain. These turns appear to be anticipatory in nature, since the blue lines in the figure indicate fixations ranging from 3 to 6 steps ahead. Although we

have not quantified these aspects of the data here, the close relation between gaze and paths were sufficient to justify analysis of the structural features of the terrain.

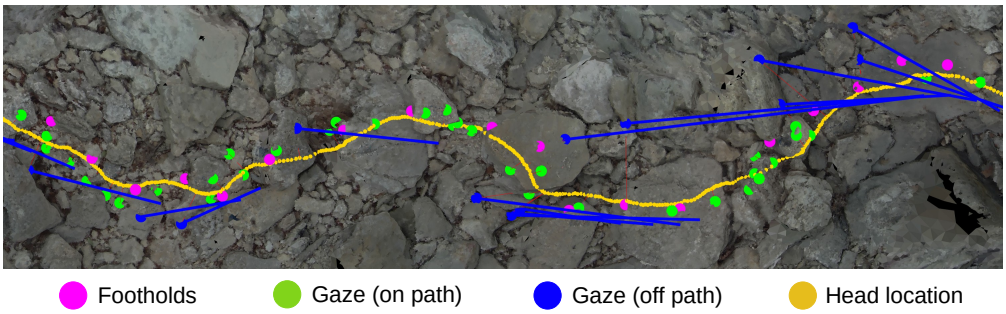

Figure 5: Gaze is used to select paths. Here we show a representative excerpt of data where gaze is directed further along the path, in this case at locations that are not travelled to. Gaze is apparently used to determine the viability of paths ahead of time, since fixations further ahead in straight directions often precede turns that deviate from the fixated locations. Other gaze locations, illustrated in green, fall close to the foothold locations shown in pink.

Role of height changes

Previous work in simpler environments has demonstrated that humans attempt to minimize energetic cost during locomotion [4], [5], [6]. We also know that monocular vision adds to uncertainty and leads walkers to look closer to the body, suggesting that depth judgments are important in foothold finding [18]. It is plausible, therefore, that changes in terrain height are relevant for walkers who might avoid stepping up and down over large rocks in order to reduce energy expenditure. This would also result in more stable locomotion, since it deviates less from flat ground which is most stable. In addition, 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. This required analyzing the terrain in a way that eliminated regions where footplants were impossible, and step transitions that were outside the distribution of those observed. We first excluded locations on the terrain where the average local surface slant exceeded 33 degrees, based on results by [20] who found this to be the approximate maximum slope of a foothold that subjects can walk on. This provides a collection of locations that we assume capture all viable step locations. We then needed to locate possible next steps, for any given foot location. For each of the subjects we calculated the distribution of step lengths (in terms of leg length), the distribution of step height (step height change either downwards or upwards), and the distribution of angular deviations from a step directly towards the subject's final step location. Distributions and schematics depicting these quantities can be seen in Figure 6. These three constraints define possible step transitions for each subject.

We then use the combination of viable step locations and possible steps in order to simulate possible paths across the terrain. (A more detailed description of this

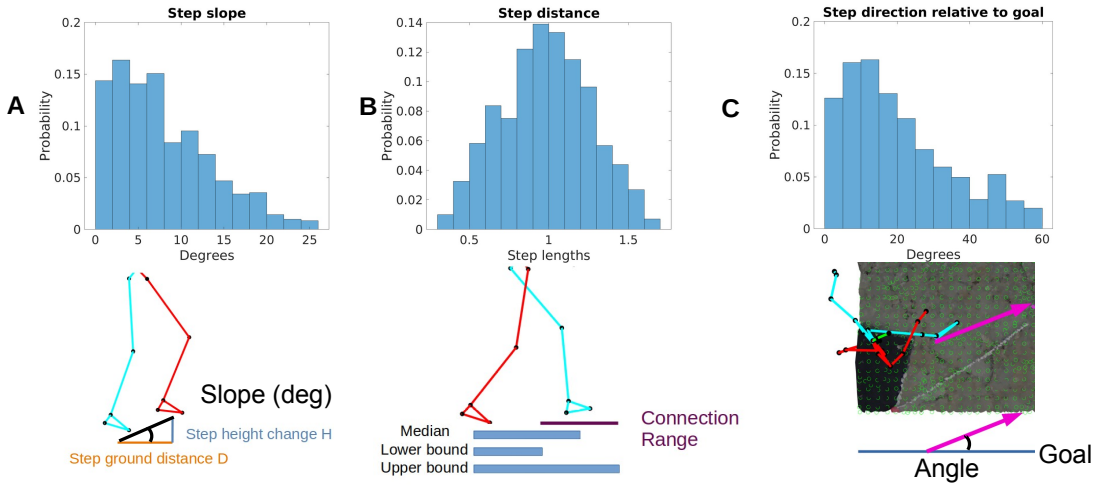

Figure 6: Step parameter distributions. The histograms show the distributions of (A) step slopes, defined as height change divided by the length of the step along the ground plane, (B) step lengths, and (C) direction changes. These deviations define set of feasible next steps for a given foothold, allowing the calculation of feasible alternative paths to the one actually chosen by the subject. The Figure shows histograms of these quantities pooled over subjects, although calculations of viable paths were done separately for individual subjects.

process is described in the Methods section.) 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 (6 step locations including the starting step location). We chose this value since an analysis of fixations showed that walkers adopt a strategy that alternates between looking in the distance to regions near the end of the path, most likely for guiding walking direction towards the goal, and looking near the body, presumably to guide foothold selection. Fixations close to the body were restricted primarily to the next 5 footholds. The 5-step sequences will be referred to here as paths. For each step location, there is thus an actual 5-step path, as well as a distribution of *possible* paths from that step location. The actual path was then compared to possible paths in order to determine the basis upon which it was selected when other paths were possible. An illustration of randomly sampled paths generated in this manner from a given footplant can be seen on the left hand side of Figure 7. The yellow lines are the randomly generated possible paths and the pink line is the actual path chosen. We first examined the average step slope of paths. For each possible path as well as the actual chosen path, we computed the average step slope (height change over step length) of all steps along the path. We call this ΔH. This results in two distributions of mean step slopes, those from the randomly sampled possible paths, and those from the chosen paths.

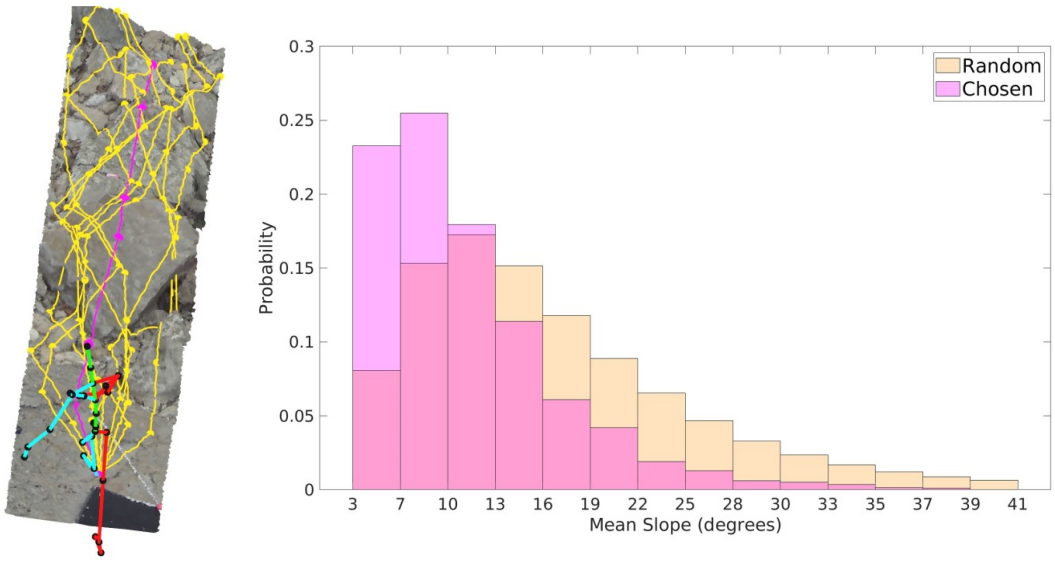

Figure 7: Chosen vs random path mean slope. Using the previously described method we can randomly sample available paths in order to compare them to the chosen path. (A) shows a subject's chosen path (magenta) along with a subset of randomly sampled paths (B) Shows histograms of the mean step slope, for paths that were chosen and for randomly sampled paths. The chosen path distribution is shifted to the left with far less rightwards skew.

The distributions of ΔH values for one subject for randomly generated versus the chosen paths are shown in Figure 7. The chosen path average slope distribution is biased to lower average slopes when compared to the randomly sampled path distribution. The median value of ΔH for the chosen paths, averaged over subjects, was 9.3 degrees, whereas for the randomly sampled paths it was 14.9 degrees. All subjects showed comparable difference and the difference in medians evaluated across subjects was highly significant ($p \ll 0.0001$). While there substantial overlap between the distributions, the chosen paths are clearly biased towards lower values of ΔH. This suggests a preference for nearby paths with lower average height changes between footholds. The existence of substantial overlap between the distributions indicates that subjects are flexible and sometimes choose paths with greater slopes, as might be necessitated by the local terrain structure. Analysis of the values of ΔH based on path computed over 3 or over 4 footholds revealed comparable differences, and the difference between the medians was greater for longer paths. An analysis of variance showed the effect for path length was significantly larger for the 5 steps paths ($p < 0.0001$). As might be expected, the distribution for the random paths shifted to the right as the paths became longer, but the distribution of ΔH for chosen paths remained constant. This provides evidence that subjects are taking into account the terrain structure for as much as 5 steps ahead.

Path Length

Frequent direction changes are a distinctive feature of locomotion in rocky terrains. This is presumably to avoid big height changes or other less desirable footplants. This feature is captured by the tortuosity metric, which is the length of the chosen path relative to a straight path. More tortuous paths are more energetically costly as they deviate more from a subject's preferred step width and are longer, so they presumably reflect a trade off with a less acceptable cost such as an obstacle or high value of upcoming ΔH. We examined the relation between tortuosity and ΔH. Randomly sampled paths with tortuosity less than the median tortuosity of all paths are classified as straight paths. (This was necessary because our calculations are all in terms of possible paths as defined above, and a straight line connecting two arbitrary locations may not be a viable path.) 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 8 and a more extensive description is given in the Methods section.

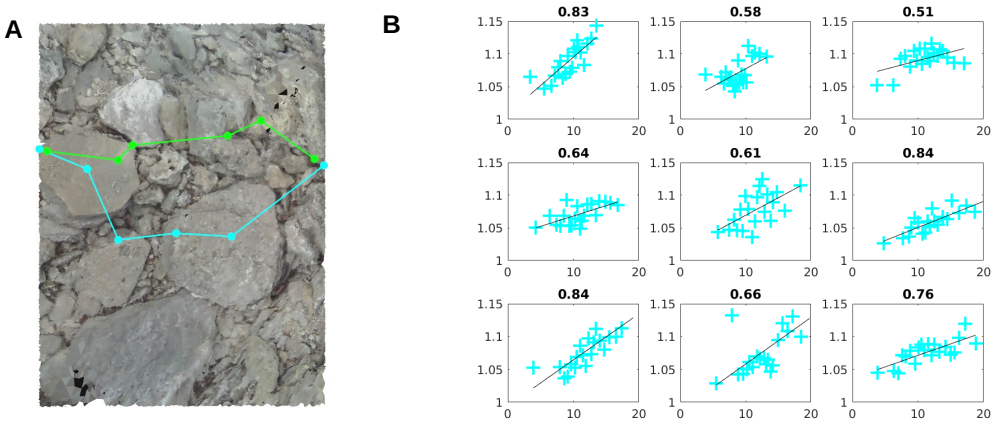

Figure 8: Turn probability vs straight path slope for 5 step sequences. For each sequence the distance of the straight line connecting the first and last footplant is computed, as well as the distance of the actual path. These are used to compute tortuosity of the chosen path. In addition, 10,000 paths are simulated that include 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, ΔH. This average straight path step slope is then compared to the tortuosity of chosen paths. Correlations are indicated at the top of each panel.

Chosen path tortuosity was positively correlated with straight path average ΔH with subjects choosing more tortuous paths for greater values of ΔH for straight paths. This relationship suggests a tradeoff between between the two and may reflect the point at which the energetic cost together with the stability cost of the longer path is less than the energetic plus stability cost of the straighter path. All subjects show his relationship although the steepness of the slope varies between the subjects.

It turns out that the slope of the relation between tortuosity and ΔH varies with the leg length of the subject. Subject leg lengths ranged from 810mm to 1035mm, with corresponding correlation values of 0.84 and 0.51, with longer leg lengths leading to shallower slopes. Figure 9 plots the correlation coefficient for the 9 subjects in Figure 8 against the subjects' leg lengths. This Figure shows that subjects with the shortest legs are more likely to choose longer paths as the direct path becomes less flat (that is, with increasing values of ΔH).

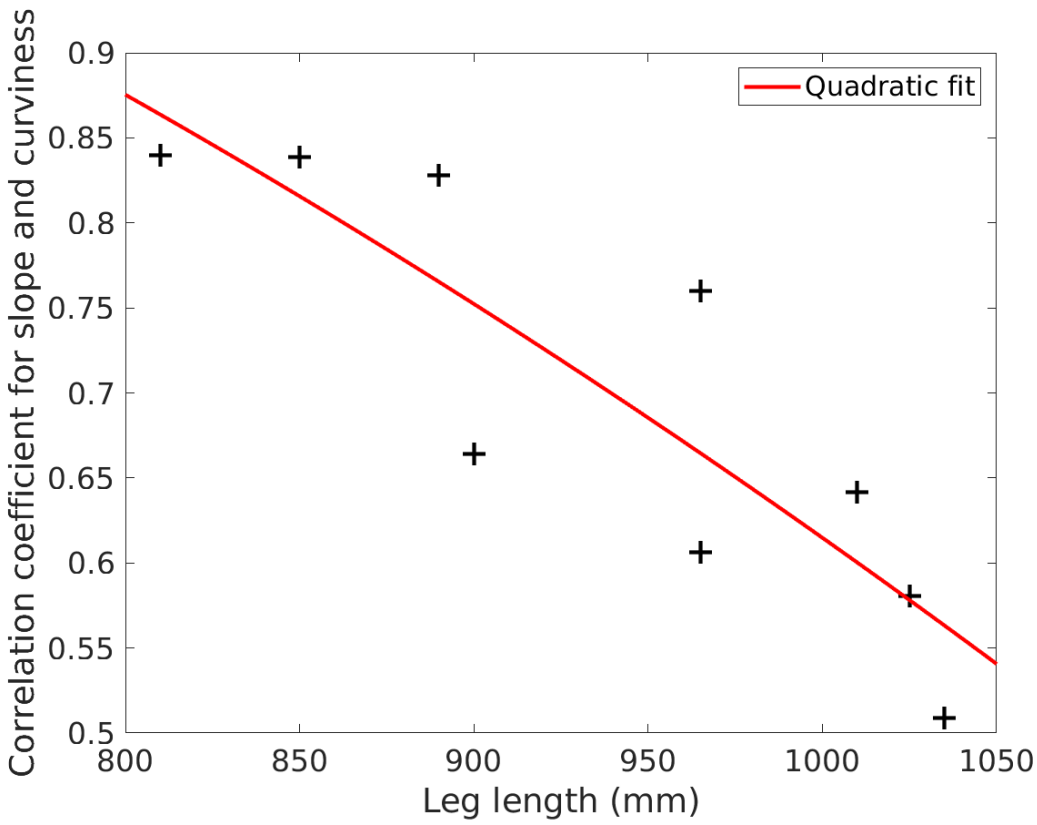

Figure 9: Relationship between leg length and correlation value between straight path step slope and path tortuosity. Subject length length (in millimeters) is plotted on the horizontal axis, against the correlation coefficients for each of the plots in Figure 8 plotted on the vertical.

Depth Features

One limitation of the previous analyses is that they implicitly assume that a subject has full information about the environment and the height changes associated with each location. However in reality subjects must make eye movements and acquire this information visually from their current viewpoint. 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 10. 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 for longer legged individuals, foot selection is more predictable on the basis of local structure features. It is possible that such individuals have a slightly different viewpoint that allows more accurate judgments.

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 better approximates how subjects might use such information. However, walkers are unlikely to have all the information present in the full resolution depth image, since depth perception falls off with eccentricity [21].

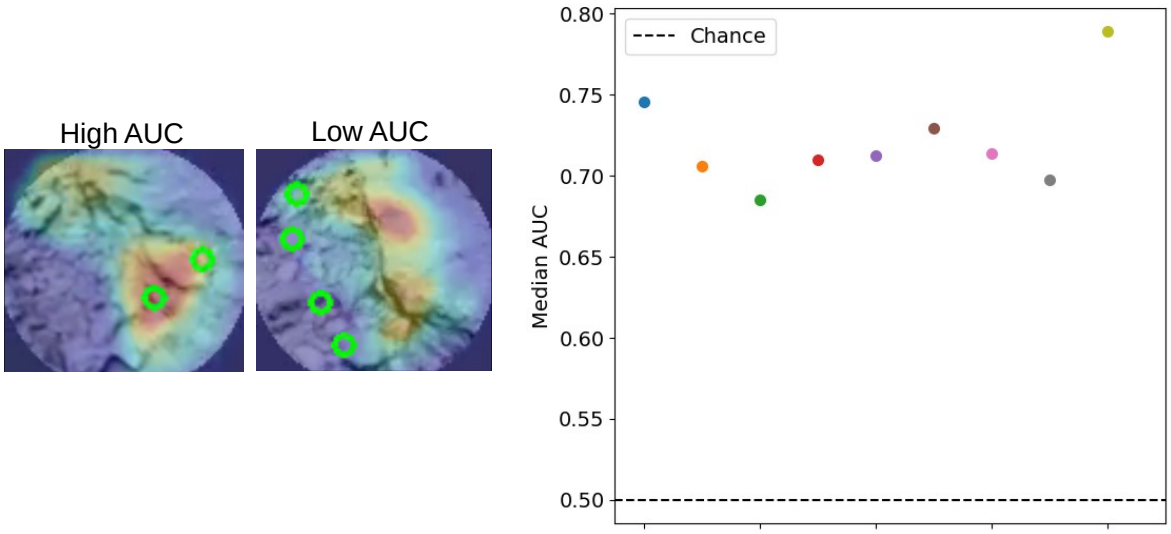

Figure 10: CNN based foothold location prediction. A CNN was trained to predict foothold locations in depth images from the subject's viewpoint. Depth images are acquired using Blender, where a virtual camera follows the same trajectory and orientation as the subject's eye. Foothold locations on the mesh are then projected back onto the retinal 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.

2 General Discussion

In this investigation we have presented a novel method for the analysis of natural terrain navigation by constructing accurate representations of the 3D visual input. This is a necessary component for understanding action decisions in natural environments. First it allowed more accurate calculations of both gaze and foothold locations than was possible in previous work in natural locomotion, where a flat ground plane was assumed, and drift in the motion capture signals added errors to the foothold estimates [22] [18]. In addition, it allowed us to analyze how the structure of the environment influences foothold selection. We first demonstrated, with examples, that there is in fact considerable regularity in the paths chosen by different walkers and by the same walkers at different times. This indicates that there are some terrain features that serve as a basis for path choice. The next challenge was to develop a strategy for finding what those features were. To do this it was necessary to generate a population of viable paths that could be compared with the chosen paths, since many locations do not permit a foothold, and for any given foot placement, leg length and other factors limit the next viable step. When comparing the population of viable paths with the chosen paths we found that walkers prefer flatter paths and avoid regions with large height irregularities. The median of the distribution of average height changes (relative to step length),

averaged over 5-step paths, was less than 10 degrees. This is quite a small change of a few inches for a normal step length.

The preference for flatter paths is presumably driven by both the energetic cost of height irregularities and by the increased stability associated with smaller height changes. As mentioned previously, humans converge to an energetic optimum consistent with their passive dynamics when on flat ground [3], [4], [5], [6] and deviations from this gait pattern, including turns and changes in speed, are energetically costly [23], [24]. Recent work by Kuo [25] also showed that subjects are able to minimize energetic cost on irregular ground planes, and do this by incorporating planning strategies to modulate speed over sequences of steps. Thus our findings indicate that minimizing energetic costs within the context of other constraints applies across a variety of locomotor contexts. Less is known about the way in which stability constrains foothold choices or how stability trades off with energetic constraints. Matthis et al [17] showed that subjects slow down as the terrain becomes more irregular, and Muller et al and Bonnen et al ([13], [18]) showed that the distribution of gaze locations moves closer to the body in irregular terrain. This strategy might address stability constraints by allowing both more time for visual input together with more accurate visual information. Both these factors might allow for course corrections without affecting energetic cost very much. The current work does not allow us to tease apart energetic costs from stability costs and presumably both are influenced by the choice of flatter paths.

A related finding was that walkers chose longer paths when the direct paths involved more height changes (Figure 8). The relationship was linear for all subjects. This suggests that the relationship defines the point at which the combined energetic plus stability costs of the more circuitous path are approximately equal to the energetic and stability costs of the straight path. The interpretation is supported by the finding that the slopes of the regression lines were more shallow for walkers with longer legs. Both the relative cost of height changes and the stability cost would be diminished as leg length increased. Another aspect of this result is that the correlations are quite high and the scatter around the best-fitting line is relatively modest. This suggests that walkers have remarkably good estimates of the costs, for both energy expenditure and stability, that are involved in the trade off. The dependence on leg length means that those costs are specific to their own bodies, and the evaluation of stability must include information about individual motor variability. This suggests that walkers use well defined internal models for state estimation and control during walking ([26], [25]).

An important finding of the present work is that even in rocky terrain, walkers appear to plan a substantial distance ahead. Not only is vision used for the two upcoming steps needed to preserve the inverted pendulum dynamics, but also

to locate flatter paths of 5 steps or more. There are several findings supporting this claim. In other data we show that chosen gaze locations cluster around the upcoming five footholds with the frequency of fixations on step N+5 being quite low, and fixations further along the path are clustered at distant locations, presumably for a different purpose such as steering towards the goal. Therefore our height change metric was calculated over these 5 step paths and the difference between random and chosen paths was greater for 5 step paths than for 3 or 4 step paths. The strength of the relationship between height changes and longer paths in Figure 8 also supports the claim that visual information about height changes is evaluated over the next 5 steps. This planning horizon is somewhat shorter than that found by Kuo et al. [25] who found evidence that walkers planned approximately 8 steps ahead to minimize energetic cost in irregular terrain. They made predictions from a dynamic model with a planning horizon of a variable number of steps. Depending on the planning horizon, the model calculations show that energy is minimized by speeding up prior to irregularities and slowing down subsequently. Human walkers' speeds were consistent with these predictions, and seemed best predicted by a planning horizon of 8 steps in the model. Although this is a longer planning horizon than we observe, their walking paths involved height changes of no more than 7.5 cm, the surfaces themselves were flat, and the path required no changes in direction. Our terrains involved greater height changes, irregular and sloping surfaces, and frequent direction changes based on visual information. More complex paths also impose a greater load on visual working memory especially in complex terrain. Thus a shorter planning horizon in our data might be expected. However, we did not explore paths over 5 steps, so the the planning horizon needs further exploration. It is clear, however, that it is at least 5 steps. This indicates underlying decision processes that optimize over sequences of movements.

While a 3D representation of the terrain is critical for evaluating the available sensory evidence, it does not precisely specify the information available to the walker. The calculation of mean height change used information from the mesh reconstruction, and it is not clear how well subjects can evaluate this quantity. We therefore used another approach to evaluating height changes that took into account the depth images available from the walker's viewpoint. The image seen by the walker is from an angle determined by the subject's height and distance along the path where they look, and visual resolution will fall off with distance. In order to take account of the viewpoint, we trained CNN's on sequences of footholds and depth images taken from the subject's viewing angle. The success of the CNN in predicting footholds supports the idea that depth changes or some other aspect of depth is used by walkers. Previous work by Bonnen et al [18], comparing gaze location in binocular and monocular vision, demonstrated that walkers gaze closer to the body in monocular vision. This supports the role of depth information, and

also indicates that the information closer to the body reduces visual uncertainty. It also suggests that walkers are sensitive to this uncertainty. Of course depth information is available using monocular cues, in particular motion parallax, so stereo may be only one source of depth information. Also, the CNN's do not reveal what aspect of the depth images is used in the prediction. For example local surface slant may be a factor in addition to height changes.

It should be noted that height changes over the next 5 steps can account for only a modest portion of the variance. While the median of the height change distribution is less than 10 deg, subjects are flexible, and sometimes choose paths with greater height irregularities, up to an average of 30 deg. This may reflect the available options within local regions of the path and the relative costs of stability, energy expenditure, and reaching the goal. Conversely, there may be a range of options that are roughly equivalent and this will limit the predictability of the paths. While choosing flatter paths can help optimize energetic and stability costs, other factors are likely to be important such as getting to the goal. Other visual features are also likely to have weight, such as the nature of the surface, for example gravel or slippery rock surfaces. Thus it would be difficult to say with confidence whether future height change is the dominant factor in path choices. However, the availability of a terrain representation allows examination of a wide variety of questions about the visual information used for locomotion.

Despite these limitations, the regularity of the paths chosen both within and between walkers is impressive. In particular, the well defined, leg-length specific, trade off between path flatness and path length is consistent with decision mechanisms that have well defined cost functions. Similarly, previous work showing the sensitive regulation of gaze location with relatively subtle terrain changes, and when binocular vision is compromised, indicates good internal estimates of sensory noise. In sum, despite the complexity of the sensory motor decisions in natural complex terrain, behavior appears fairly tightly orchestrated by decision mechanisms that are attempting to optimize for multiple factors in the context of well calibrated sensory and motor internal models.

3 Acknowledgements

This work was supported by NIH Grants EY05729 and K99 EY 028229.

---

4 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 eye and body

tracking devices. 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. Data from one subject in the Austin data set were not used because the camera angle did not give a good view of the terrain for the Photogrammetry, and two subjects in the Berkeley data set were not used because of poor quality eye tracking or poor terrain images.

Equipment: Eye and body movements of both groups of participants were recorded using a Pupil Labs Core 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 loosely defined 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 [17]. Since both terrains were rocky, it was necessary for subjects to use visual information in order to localize and guide foot placement [17].

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. 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. We segmented the image into saccades and fixations using an eye-in-orbit velocity threshold of 65deg/s and an acceleration threshold of 5deg/s^2. The velocity threshold is quite high in order to accommodate the smooth counter-rotations during stabilization. Saccadic eye movements induce considerably higher velocities, but saccadic suppression and image blur render this information less useful for locomotor guidance. For more detailed description of pre-processing of motion capture and eye tracking data, see [17] and [22].

Terrain Reconstruction: In order to estimate both the environmental structure and the relative camera position from the head mounted video, we used software package called Meshroom [19]. Meshroom [19] 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's 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 [27]. The individual frames are undistorted using a camera intrinsic matrix estimated by checkboard calibration [28]. This allows a pinhole assumption for the images, 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 pipeline described above, resulting in a 6D camera trajectory (3D position and 3D orientation), with one 6D vector for each frame of the original video (See Figure 2 for rendered image of textured Mesh output).

4.0.1 Motion capture to mesh data alignment

Meshroom provides pose (3D position, and 3D orientation) estimates corresponding to each of the input 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 Labs eye position and Motion Shadow body data with Meshroom's coordinate system. The 3D orientation of the world camera in the eye tracking and motion capture data is available from the procedure described above and in [17]. 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 1 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.

To evaluate the error in estimating foothold location we used the meshes estimated from different different terrain traversals and found different foothold estimates for the different meshes, The distribution of the differences is shown in part A of Figure 11. We also manually annotated the video images and compared the estimated foothold location with the location in the video image. This is shown in part B of Figure 11.

4.0.2 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 correspond-

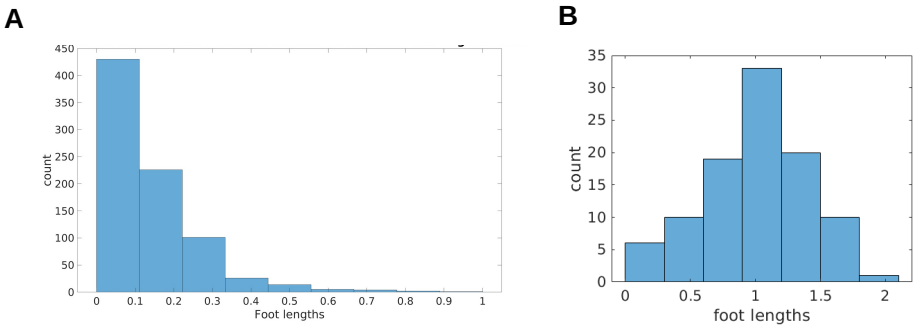

Figure 11: Foothold localization error. A. Distribution of between mesh errors of foothold location estimates (for the same subject traversal data. Foothold locations are estimated according to the same process described, but the terrain data used is interchanged, and the resulting different corresponding 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

ing 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 and in addition was useful for computing a cross mesh error metric.

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 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). 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.

4.1 Pre-processing

4.1.1 Motion capture data

For more detailed description of pre-processing of motion capture and eye tracking data, see [17] and [22].

4.1.2 Possible step and path simulation

In order to facilitate analysis of the data with respect to path planning and foot placement, all possible foot locations and steps between foot locations are predetermined using several constraints. The first is a constraint on possible step locations. Maximum walk-on-able slope was previously measured in [20]. 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 6). 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.

4.1.3 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 [29], 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 subject's gaze based on the aligned eye tracking data. The virtual camera was 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 coordinates (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 knowledge of the environment when choosing foothold locations.

4.2 Detailed Analysis

4.2.1 Possible path vs chosen path analysis

This analysis leverages the pre-computed possible step locations and possible steps connecting them from (4.1.2). 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 (6th 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.

4.2.2 Straight path slope vs. curved path probability

This analysis relies on the paths discussed in the 4.2.1. 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.

4.2.3 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 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 in diameter.

References

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

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

[3] AD Kuo, J. M. Donelan, and Ruina A. Energetic consequences of walking like an inverted pendulum: Step-to-step transitions. *Exerc. Sport Sci. Rev.*, 33:88—-97, 2005.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] F Dominguez-Zamora and D Marigold. Motives driving gaze and walking decisions. *Current Biology*, 31(8):1632–1642.e4, 2021.

[13] Karl S Muller, Kathryn Bonnen, Jonathan S. Matthis, Lawrence Cormack, Alexander Huk, and Mary Hayhoe. Retinal motion statistics during natural locomotion. *eLife*, 12:e82410, 2023.

[14] Jeff B Pelz and Constantin Rothkopf. Oculomotor behavior in natural and man-made environments. In *Eye Movements*, pages 661–676. Elsevier, 2007.

[15] 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.

[16] 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.

[17] 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.

[18] K Bonnen, Jonathan S Matthis, Agostino Gibaldi, Martin S Banks, Dennis M Levi, and Mary Hayhoe. Binocular vision and the control of foot placement whie walking in natural terrain environment. *Scientific Reports*, 11(1):20881, 2022.

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

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

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

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

[23] AS Voloshina, AD Kuo, MA Daley, and DP Ferris. Biomechanics and energetics of walking on uneven terrain. *Journal of Experimental Biology*, 216:3963—-3970, 2013.

[24] R. G. Soule and R. F. Goldman. Terrain coefficients for energy cost prediction. *Journal of Applied Physiology*, 32:706—-708, 2072.

[25] AD Kuo and O. Daricia. Humans plan for the near future to walk economically on uneven terrain. *Proceedings of the National Academy of Sciences*, 120(19):88—-97, 2023.

[26] Emanuel Todorov. Optimality principles in sensorimotor control. *Nature neuroscience*, 7(9):907–915, 2004.

[27] FFmpeg team. FFmpeg, 2021.

[28] 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.

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