Infant gait modifications: Integration of computer vision with human annotation provides accurate step classification and location as infants navigate varied terrain

Christina Hospodar*
Department of Psychology
New York University
New York, NY, US
christina.hospodar@nyu.edu
(*shared first authorship)

Tieqiao Wang*
School of EECS
Oregon State University
Corvallis, OR, US
wangtie@oregonstate.edu
(*shared first authorship)

Yasmine Elasmar
Department of Psychology
New York University
New York, NY, US
yasmine.elasmar@nyu.edu

Sinisa Todorovic
School of EECS
Oregon State University
Corvallis, OR, US
sinisa@oregonstate.edu

Karen Adolph
Department of Psychology
New York University
New York, NY, US
karen.adolph@nyu.edu

Abstract—Functional locomotion requires perception of affordances—the fit between body and environment that makes particular actions possible. To cope with changing affordances while walking over varied terrain, walkers must modify their steps as they approach and navigate each ground surface. Newly walking infants have the physical wherewithal to modify their gait by slowing down and taking shorter steps, but they do not do so systematically or prospectively. To test how gait modifications develop, we video recorded new and experienced infant walkers as they approached and crossed slopes and bridges. The long-term aim is to measure whether, when, and how infants modify their gait (relative to slope degree and bridge width, the location of the obstacle in space, and typical gait on flat, wide surfaces). Useful data require high precision in classifying steps and identifying the 3D location of infants' feet at each moment (despite the idiosyncrasies of infant movements and frequent occlusion and motion blur)—requirements beyond the capabilities of human annotation or computer vision alone. Thus, we built an integrated human-machine system to identify each step and its XYZ coordinates as infants approached and crossed the obstacles. We capitalized on the ability of human annotators to classify infants' movements into steps and of computer vision to identify the 3D coordinates of the feet in each video frame. We demonstrate the feasibility of this integrated, human-machine system to investigate the development of prospective infant gait modifications.

Keywords— computer vision, infant, walking, locomotion, gait, gait modifications, perception

I. Introduction

Walking over irregular terrain requires perception of affordances (e.g., whether a slope is too steep for walking) [1], [2]. However, in most situations where walking is possible, gait modifications are required. People must tailor walking patterns to features of the environment (e.g., decrease step length and speed to brake forward momentum on steep slopes). Moreover, gait modifications provide insight into the continual interactions between perception and action because functional locomotion in a real-world environment requires visually-guided control of walking [3]. Exploratory looking and touching generate perceptual information to guide upcoming action, and feedback from the last step guides planning for the next step. Thus, plans are continually updated and are reflected in gait modifications.

How do prospective gait modifications develop? From their first walking steps, infants' gait is not uniform. Babies can take slower or faster, shorter or longer, and wider or narrower steps. The range of possible gait modifications increases with walking experience, but even novice walkers can slow down and shorten their steps. Thus, by the time infants have sufficient skill to walk across a room, they are physically able to modify their gait. But can infants modify gait systematically to suit variations in terrain? Prior work suggests the ability to modify gait develops with walking experience. New walkers do not modify their gait prospectively to cope with slopes, bridges, and so on [4]. But after several months of experience, infants take smaller, slower steps to walk down steep slopes [5], across narrow bridges [6], [7] and over small obstacles [8]; they carefully lower themselves to step down a high drop- off [9]; and turn sideways to squeeze through narrow apertures [10].

However, the recording technologies used in prior work limited researchers' ability to determine whether, when, and how infants modify their gait with high resolution. Motion tracking markers and electromyography (EMG) require infants to wear recording devices and thus limit the number of test trials and constrain testing to a small area [4], [8]. Video-based measures are less intrusive, so infants can produce more trials in larger areas on more varied ground surfaces. But prior video measures relied on human annotation with relatively low resolution such as the total number of steps or time per trial [11]. Higher resolution measures are needed to understand planning and implementation of gait modifications within and across trials—where in the trial infants modify their gait, step trajectories, and so on.

Here we present a new solution by integrating computer vision and human annotation. Humans can easily identify each step, but cannot accurately identify the location of the foot in the air or on the ground. The opposite is true of computer vision: Identifying foot location is relatively trivial, but step classification is inaccurate. Our solution is a human-in-the-loop integration for accurate, high-resolution gait measures.

This research was supported by Defense Advanced Research Projects Agency Grant (N66001-19-2-4035) to KEA and ST and a National Institute of Child Health and Human Development Grant (F31-HD107999) to CH.

II. Related Work

Gait modifications can be understood only relative to typical gait patterns on flat, open ground. A century of work details development of infant walking over straight paths on flat, open ground [12]-[15]. However, infants' typical walking patterns provide no insight into their ability to modify their gait to cope with changes in the environment [16].

A. Measuring Infant Walking on Flat, Open Ground

To record gait, infants typically have to wear something (magnetic markers, inertial sensors, etc.) or walk over something (instrumented mat, force plate, etc.). But infants are not compliant participants, and often remove devices attached to their body. When encouraged to walk a specific path, babies veer off-course. The practical demands of typical recording methods (trailing wires, limited calibration space, floor sensors in particular locations, etc.) and the need to record a series of gait cycles leads most researchers to study gait as infants step on a treadmill or walk over flat, uniform ground. Perhaps the most robust finding from the standard "straight-path" test is that elapsed time since walk onset predicts infants' gait maturity . Younger, less experienced walkers display shorter, wider, slower, more variable steps compared to infants with more months of walking experience.

B. Challenges in Measuring Gait Modifications

Technical constraints make it difficult to test infants in situations that require them to modify their gait (e.g., instrumenting a sloping walkway is difficult). Moreover, measuring gait modifications requires greater precision than measuring gait. Overall gait patterns are still recognizable despite a missing step, but gait modifications are not robust to missing steps [12], [16]. Computer vision can obviate the need for wearable recording devices or limitations in terrain. But it cannot solve the problem of accurate step classification alone.

C. Computer Vision for Gait Analysis

Gait analysis is a prominent focus in the computer vision community [17], [18]: gait recognition [19] to identify individuals by their walking patterns, and gait analysis to detect abnormal gait patterns for intervention or rehabilitation [20]. Such research typically involves coarse predictions, with only one classification per video. Notably, such work features videos with a limited number of well-separated adult walkers.

Our work is distinct because it addresses the unique challenges posed by infant walkers. The most critical design features of existing computer vision systems expect motions to be temporally smooth. But infants' footwork includes non-smooth, jittery movements. Moreover, compared to adults, infants' feet are small, creating low pixel resolution and large motion blur and frequent occlusion by nearby experimenters or infants' changes in posture. Thus, no existing work produces fine-grained, location-specific gait analysis.

III. Considerations In Experimental Design

We tested infants walking on adjustable slopes and bridges. Here we focus on the relevant considerations of the study design in relation to the human-machine interface.

A. Selecting the Participants

To test effects of age and walking experience on gait modifications, we tested younger, less experienced infants (12-14 months, ≤ 6 weeks of walking experience) and older, more experienced infants (17-19 months, ≥ 21 weeks of walking). Caregivers reported infants' walk onset age (first day they saw infants walk 3m independently without stopping or falling). We expected older, more experienced infants to modify their gait to cope with steeper slopes and narrower bridges, whereas younger, less experienced infants would not.

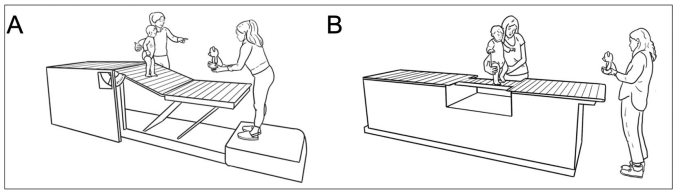

Fig. 1. Adjustable slope and bridge. (A) Slope apparatus (86 cm wide × 276 cm long): Flat starting and landing platforms (86 cm wide × 92 cm long) flank a sloping middle section (0-90° in 1° increments). (B) Bridge apparatus (76 cm wide × 288 cm long): Wide starting and landing platforms (76 cm wide × 106 cm long) flank a 76-cm long gap spanned by bridges of varying widths (1-76 cm in 1-cm increments) that can be quickly inserted and removed. Experimenter follows alongside infants to ensure their safety. Caregivers at the far end of the landing platform encourage infants to walk.

B. Selecting the Terrain

We selected slopes and bridges (Fig. 1) because they are novel and challenging, can be adjusted incrementally, entail a series of steps over the surface, and involve similar penalties for error. Most critical, on both apparatuses, infants must modify their gait prospectively to prevent falling. The surfaces of each apparatus were striped (5 cm apart) to calibrate the physical space for computer vision. See databrary.org/volume/1685/slot/69987?asset=474902 for an exemplar video.

C. Collecting a Range of Trials

Sessions lasted ~90 minutes. Infants walked barefoot wearing a colorful onesie so movements were unimpeded and limbs clearly visible on video. The experimenter wore a long-sleeved shirt of a contrasting color to make the infant's body more distinct. Half of the infants started with the slope and half with the bridge. Trials began with infants standing on the starting platform. Caregivers at the far side of the landing platform encouraged infants to cross, while the experimenter followed alongside infants to ensure their safety (Fig. 1).

Infants completed 4 baseline trials on flat, wide ground on the first apparatus (0° slope or 60-cm wide bridge). Then they were presented with a more difficult increment (8° slope or 52-cm wide bridge). We used an adaptive staircase procedure, increasing or decreasing the difficulty of each trial based on the outcome of the previous trial. Baseline trials were interspersed as needed to renew infants' motivation to walk and to ensure that gait modifications on challenging increments were not due to fatigue. The procedure was repeated with the other apparatus. Thus, infants were presented with slopes 0° to 50° (~30-50 total trials) and bridges 4 cm to 60 cm wide (~30-50 total trials).

D. Collecting "Gold-Standard" Data

Infants walked over a pressure-sensitive mat on the floor (Protokinetics, 1.2 m wide × 4.9 m long, protokinetics.com, 120 Hz, 4 sensors/in^2) to provide a gold-standard comparison for computer vision location estimates. The experimenter placed infants at one end of the mat and caregivers at the other end encouraged infants to walk toward them as quickly as possible. If infants careened off the mat, stopped, or fell, we repeated the trial, aiming for at least 3 good trials.

E. Recording in High-Resolution From Multiple Angles

To produce reliable estimates of foot location, the feet must be visible in at least three camera views (30 fps), so each task (slope, bridges, gait mat) had four overhead camera views exported in high-resolution (1920×1080 pixels). The four views were also synced into one video frame.

IV. Computer Vision

Our computer vision system aimed to estimate 3D foot locations in each frame.

A. Video Pre-Processing

Our goal was to accurately segment small objects—the baby's feet. Thus, the detector model was run on the four high-resolution views, but two synchronization challenges arose. First, although the four individual views began recording simultaneously, a potential 0.4-s offset may exist between views, posing a substantial disparity for quickly moving objects like infant feet. Therefore, despite identical timestamps, the four high-resolution views must be synchronized. So, the experimenter flashed the room lights during each recording, and we identified the first frame of darkness in each high-resolution view as follows:

$$\begin{cases} 1, & if \ \sum F_i < dm\_thred \\ 0, & if \ \sum F_i \geq dm\_thred \end{cases} \quad (1)$$

where $\sum \mathrm{F_i}$ represents the summation of all RGB values within the current frame F_i, and dm_thres denotes a predefined threshold for "dark moment" determination.

Second, time stamps for each trial were human-annotated on the low-resolution mixed views. Thus, after creating the high-resolution mixed video, it was synchronized with the low-resolution mixed video to identify the corresponding video segments of interest. Video synchronization was achieved using a distinctive frame from the moving baby. After a unique frame, denoted as F^h, was pinpointed within the high-resolution view, we determined the optimal matching with the following optimization:

$$\arg \min_{\mathrm{i,i\in[1,T]}} \left|\left|F_i^l - \bar{F}^{\mathrm{h}}\right|\right|_2 \quad (2)$$

where T represents the total number of frames, F_i^l denotes the i-th frame within the low-resolution video, and $\bar{F}^h$ signifies the resized high-resolution frame.

B. Camera Calibration

Camera calibration [21] determines the parameters that articulate the relations between the 3D world and the 2D image—imperative for the 3D reconstruction of foot locations. In the 3D space, the walkway surface serves as the x-y plane. To calibrate the slope, we varied the slant and used the stripes on the slope for camera calibration. To calibrate the bridge, we affixed two checkerboards onto 3×3ft rigid boards, each comprising 12×12 units. We placed one checkerboard on the walkway and the other perpendicular to the walkway, moving them forward systematically to cover the 3D space. Thus, we guaranteed that the stripes or checkerboards covered the entire space for potential foot locations, as even a minor error could result in a substantial misestimation of a tiny foot.

C. Human Pose Estimation

We considered using pose estimation to estimate infants' foot location such that the entire body could provide contextual information [22]. But it was unsuitable. Ensuring the baby's safety on slopes and bridges—where infants often fall—takes precedence. Thus, the experimenter must stand with an arm on each side of the infant. Consequently, the two bodies overlapped in the image, leading to significant occlusions. We tested the most recent models for occlusion-aware 3-D pose estimation models, but they did not consistently distinguish experimenter from infant, and are not designed or trained to handle such challenging scenarios. Thus, state-of-the-art human pose estimation models did not return precise foot keypoints. Improving human pose estimation to fine-tune existing models by compiling a suitable training dataset with manual annotations of infant body keypoints from our videos would be prohibitively costly.

D. Foot Detection

Thus, we used an alternative approach based on object detection to deliver precise foot keypoint locations. Although state-of-the-art object detection models underperform on small objects (here, infants' feet) [23], they allow for more convenient fine-tuning than human pose estimation. The manual annotations for fine-tuning an object detector consist of outlining ground-truth bounding boxes or polygons around objects (infants' feet), which can be done more quickly and accurately than annotating the body keypoints.

To identify and detect infants' right and left feet in a video frame, we used the state-of-the-art Mask2Former model [23] with weights pre-trained on the public COCO dataset of images [24]. Following the standard transfer-learning practice for fine-tuning a pre-trained model [25] we fine-tuned the entire model with a modified classification head (adapted to our dataset) to predict foot instances under low resolution, motion blur, and partial occlusion. The model was fine-tuned in 20,000 iterations with a batch size of 2 and learning rate of 5e-5, on 1089 random frames across the slope, bridge, and gait mat videos from 13 randomly-selected infants, manually labeled with LabelMe [26]. However, after fine-tuning, our Mask2Former still frequently confused the two feet in the image, due to their similar appearance under low resolution and motion blur. We improved identification of left and right feet by extending Mask2Former with an additional module to reason about foot trajectories. However, in cases when infants step backward or rotate or raise their feet up and down in place, this module fails to improve foot classification. Note that our model is only trained on detecting bare feet; using the model to detect shod feet would require additional training.

E. 3D Foot Estimation

After detecting the right and left feet in each camera view, we mapped the estimated image locations to the 3D world coordinates. The center of a predicted bounding box serves as the foot location in the image. Given our comprehensive calibration of the entire 3D space of possible feet locations, any errors in estimating the 3D coordinates of the feet can originate only from mistakes in our 2D object detection. The 3D foot location estimation first considers pairs of detections with the same label (e.g., right foot detection pairs) identified in every pair of camera views, and for each detection pair estimates the corresponding 3D location of the foot using the standard stereo geometry [21]. It then removes outlier detection pairs whose 3D reconstructions significantly differ

from those of the other detection pairs, and finally estimates the 3D location of the foot using the least squares algorithm [27] from the remaining inlier pairs of detections.

V. Human Annotation Tool And Workflow

Even the most advanced deep-learning models may fail when handling edge cases underrepresented in the training data. Thus, to ensure the accuracy of the data, human annotators must classify steps and correct foot detections.

A. User Interface

Our tool is publicly available (github.com/tqosu/Infant-Gait-Modifications). The user interface (Fig. 2) consists of three windows: a control panel, video player, and top-down visualization of the steps and foot locations. The control panel allows users to select participants and trials, control video playback forward and backward at various speeds, navigate videos frame-by-frame, alter foot labels, toggle the display of detected foot boxes, delete foot boxes, insert steps, and so on. Users can activate these functions with button clicks, sliders, and keyboard shortcuts. Users can select what camera views to display in the video player, and the video player includes frame and timing information. The top-down visualization allows users to see the locations of user-inserted steps in the top panel and the locations of the feet in every frame in the bottom panel.

B. Annotating Steps in the User Interface

Human annotators identify each step event. Of note, we originally tested a workflow where the interface provided suggestions for steps using a hierarchical clustering algorithm to group foot locations (clusters surpassing a predefined threshold of 4 were retained as a step cluster, with a step suggested at the earliest frame within the cluster and the averaged location of all values in the cluster). However, the suggestions required human review and considerable correction, as the algorithm frequently added or missed steps, or inserted steps at incorrect times (correcting one 90-min session took ~ 4-5 hrs). Thus, humans independently insert each step event (annotating one session takes < 2 hrs).

Humans add a step (labeled left or right) in the first frame where the foot fully landed in all 4 views after a displacement of ~1 inch or more (despite the resynchronization with the light flash, camera views are not perfectly synchronized at every moment).

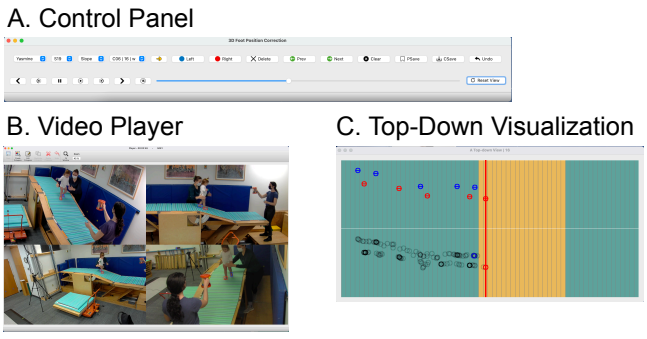

Fig. 2. User interface. (A) Control panel for user navigation. (B) Video player. Figure displays all camera views but any views can be selected. Feet shown surrounded by detection boxes. (C) Top-down visualization. Top panel shows user-inserted steps. Bottom panel shows foot locations in all frames. Vertical red line indicates the step currently displayed in the video.

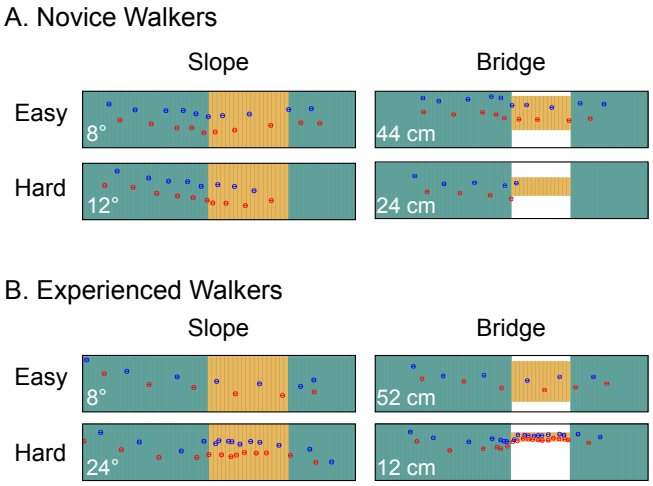

Fig. 3. Example footprints from (A) novice infant walkers and (B) experienced infant walkers. Blue circles denote steps with the left foot and red circles denote steps with the right foot. (A) On easy trials, novice walkers take wide, short, irregularly spaced steps. On hard trials, footprints end midway on the obstacle because infants did not modify their steps and fell. (B) On easy trials, experienced walkers take narrow, long, regularly spaced steps. On hard trials, footprints cluster while approaching and crossing the obstacle, showing gait modifications on steep slopes and narrow bridges.

Of note, on steep slopes and narrow bridges, infants' steps occasionally do not have a clear swing phase (e.g., sliding, shuffling steps) and/or clear landing (e.g., slow landing of heel, scrunched foot that does not flatten, toes lifted after foot landing), making step identification challenging. Nonetheless, Fig. 3 and the exemplar video demonstrate the feasibility of the system for annotating steps and comparing gait modifications across infants, apparatuses, and trials (databrary.org/volume/1685/slot/69987?asset=474902).

C. Correcting Foot Detections in the User Interface

As the video plays, infants' feet are outlined with foot detection boxes (blue for the left foot, red for the right). After identifying frames for each step, human annotators verify that the feet were accurately detected and labeled so the resulting locations are correct. Thus, human annotators swap detection boxes if the algorithm incorrectly labeled the feet or delete detection boxes if the relevant foot was not identified. If a correction is made, the software automatically re-estimates foot locations, updating the detection boxes in the videos and the location of the foot in the top-down visualizations.

D. Outlier Removal for Remaining Foot Detections

After steps are identified and foot detection boxes adjusted, we use the verified data as constraints to identify outliers for foot detection boxes in all remaining frames, as the foot detection boxes must be correct in every frame to calculate the trajectory of the foot between steps and to determine the maximum height of each step. Within each segment defined by two steps of the same foot, the stepping foot moves through the air (or slides along the ground) and the other foot remains stationary. For the moving foot, we predict a trajectory based on boundary annotations. If the distance is less than a threshold of 25 pixels (considering the potential movement of a baby's foot between two consecutive frames), we retain the detection and update the trajectory; otherwise, we discard it. For the stationary foot, we preserve the foot estimation if the Intersection-over-Union (IOU) score between the prediction and the ground truth at the segment boundary is greater than 0.6.

VI. Results

Here we evaluate the reliability of human annotation to accurately identify steps and evaluate the ability of computer vision to accurately locate and label infant feet.

A. Reliability of Human Annotation of Steps

Two humans annotated steps for one exemplar novice and one exemplar experienced walkers' videos, selected quasi-randomly to ensure infants' behavior was representative. As shown in columns two and three of Table I, the coders detected similar number of steps across apparatuses and infants, with a total difference of 4 steps across both videos. (The computer vision algorithm we chose to ignore suggested 134 additional steps compared to human annotators.)

To determine where human annotators differed, we created a 4-frame window around each annotator's step events (2 frames prior to the step event and 2 frames after) and checked that the windows overlapped. We used a relatively conservative window of +/- 2 frames to ensure high temporal precision (a larger window could encompass more than one step if two steps happened in quick succession). If the windows did not overlap, we counted those detections as disagreements. Thus, disagreements could result if coders detected step events at different times (outside of the 4-frame window, thus resulting in 2 disagreements for the same event) or if one coder did not count a step that the other coder counted. As shown in the right column of Table I, disagreements were low, M = 3.1% overall. Disagreements were higher for the bridge than the slope, and higher for the experienced baby than the novice baby, likely due to the difficulty of identifying very small, short, low-to-the-ground steps on narrow bridges. Additionally, 38 of the 49 disagreements for the experienced baby on the bridge were timing disagreements (resulting in 2 disagreements for the same event), which can occur due to imperfections in video syncing or to difficulty identifying steps with a limited swing phase and/or muddled landing—not detection disagreements. Thus, human step annotation is highly reliable.

B. Evaluation of 3D Location Using Gait Mat Data

To evaluate the accuracy of 3D location estimates, we compared our computer vision estimates to ground-truth data from the instrumented gait mat for one novice and one experienced infant (Sec. Ⅲ-D). Table II provides the precision and recall between the predicted and ground truth centroids at two different thresholds. Our foot detection results reported in Table II are significantly better than performance of the latest instance-segmentation models on small-size objects in the benchmark public datasets. For example, the latest, best AP50 results (Average Precision at IOU threshold of 50, equivalent to Precision at the threshold of 1" in Table II) are 30.3 [28] and 30.7 [29] for small objects in the SOD4SB dataset [30] and the SODA-D dataset [31], respectively.

TABLE I. Reliability of Human Annotation of Steps

	Human 1 Step #	Human 2 Step #	Disagreements
Novice: Slope (21 trials)	271	270	1 (0.3%)
Novice: Bridge (49 trials)	598	600	8 (1.3%)
Experienced: Slope (33 trials)	452	453	3 (0.6%)
Experienced: Bridge (38 trials)	692	686	49 (7.1%)
Total (141 trials)	2013	2009	62 (3.1%)

*Disagreements occurred if Human 1 identified a step that Human 2 did not or vice versa. Thus, if Human 1 and Human 2 both identified a step, but did so outside of the 4-frame window, 2 disagreements resulted. Percentages in parentheses are estimated by dividing the number of disagreements by the total number of steps identified by Human 1.

TABLE II. 3D Location Evaluation

Threshold	Precision	Recall
1.5"	97.4	99.0
1.0"	84.3	91.2

TABLE III. Evaluation of Foot Fetection (Phase 1 Model)

	Swap	Delete	# Steps	1-2 Outliers	3-4 Outliers	# Frames
Novice: Slope (3 infants, 79 trials)	261 *24%*	73 *7%*	1081	4085 *29%*	416 *3%*	13,894
Novice: Bridge (3 infants, 118 trials)	82 *5%*	118 *7%*	1634	3117 *14%*	353 *2%*	21,607
Experienced: Slope (3 infants, 125 trials)	374 *23%*	162 *10%*	1593	5222 *33%*	738 *5%*	15,793
Experienced: Bridge (3 infants, 111 trials)	116 *6%*	231 *11%*	2040	19,199 *52%*	2094 *6%*	37,241
Total (433 trials)	833 *13%*	584 *9%*	6348	31,623 *36%*	3601 *4%*	88,535

*Table III summarizes foot detection data from 6 infants (3 novice, 3 experienced), processed with Phase 1 of the model. The "Swap" and "Delete" columns provide counts of the number of foot detection boxes where the human annotator had to swap the left and right labels or delete the detection boxes entirely, out of all the frames where steps were added (and across all 4 camera views). The "1-2 Outliers" and "3-4 Outliers" columns provide counts of the number of frames where 1-2 or 3-4 outliers were removed from the full-trajectory data, using the step-level data as constraints (see Sec. VD). Italicized numbers are % of steps (swaps, deletions) or frames (outliers).

TABLE IV. Evaluation of Foot Detection (Phase 2 Model)

	Swap	Delete	# Steps	1-2 Outliers	3-4 Outliers	# Frames
Novice: Slope (81 trials)	257 *19%*	62 *5%*	1367	6326 *33%*	1065 *6%*	19,268
Novice: Bridge (101 trials)	132 *7%*	5 *<1%*	1949	19,420 *49%*	2195 *6%*	39,409
Experienced: Slope (107 trials)	374 *28%*	45 *3%*	1337	5466 *39%*	416 *3%*	14,040
Experienced: Bridge (103 trials)	85 *6%*	7 *<1%*	1538	16,114 *76%*	511 *2%*	21,220
Total (392 trials)	848 *14%*	119 *2%*	6191	47,326 *50%*	4187 *4%*	93,937

*Table IV summarizes foot detection data from 6 additional infants (3 novice, 3 experienced), processed with Phase 2 of the model (trained on "Delete" examples from Phase 1). Columns are the same as Table III.

C. Evaluation of Foot Detection

To evaluate the accuracy of foot detection, we first used the 1089 manually labeled frames (Sec. IV-D) to evaluate Mask2Former's ability to correctly detect the foot in 2D. Average Precision (AP) was 93.7 at the IOU threshold of 50% and 76.9 at the IOU threshold of 75%, suggesting that the algorithm performed reasonably well at drawing bounding boxes around infants' feet, especially given the challenges introduced by occlusion, motion blur, and small object sizes.

Next, we counted the number of times humans had to swap detection boxes (where the algorithm misassigned left-right labels) or delete detection boxes entirely (where the algorithm misidentified objects as "feet") for frames where they added steps. Table III provides counts for swaps and deletions from 6 infants (3 novice, 3 experienced) from Phase 1 of the algorithm, and Table IV provides counts for swaps and deletions from 6 additional infants (3 novice, 3 experienced) from Phase 2 of the algorithm. Phase 2 was trained on "delete"

examples from Phase 1. Across both phases, swapping functions were more common than deletions, suggesting that the algorithm usually identifies both feet, but occasionally confuses the left-right labeling (Sec. IV-D). Moreover, the number of deletions decreased from Phase 1 to 2, indicating improvement of the algorithm with additional training data, and highlighting the iterative nature of our approach. We expect the algorithm to continue to improve with more training data. If users wish to improve the number of swaps, the algorithm could be trained on swapping data.

The outlier columns of Tables III and IV show the number of frames with 1-2 or 3-4 outliers, identified and removed as in Sec. V-D. Note, we prioritize high accuracy to detect outliers, so the algorithm eliminates outliers of temporal inconsistency or low confidence in cases with severe occlusion or high motion blur. Removing 1-2 outliers improves the accuracy of the identified location of the foot and saves considerable human effort of manually swapping or deleting outliers from the full trial (humans only manually correct box detections for frames where steps were inserted). Removing 3-4 outliers similarly saves human effort, but deletes the foot detection from that frame, as the foot must be detected in at least 2 views to estimate location. Nonetheless, only ~4% of frames were removed due to outliers; trajectory data can be inferred by assuming smooth temporal motion between steps, or users can manually correct detection boxes for trials with high numbers of deleted frames. Compared to human correction, our approach significantly reduces the need for human intervention while maintaining high detection precision and recall in every frame.

VII. Conclusion

We demonstrated the feasibility of an integrated human-machine system to evaluate infant walking on varied terrain, where humans annotate step events and computer vision identifies the location of infants' feet in every frame. This system expands opportunities to study unconstrained infant walking, providing a user-friendly, accurate annotation tool and a streamlined pathway for future research endeavors. As long as the system is used in a calibrated space and the feet are clear in videos, the system could be used in many settings and for many tasks (including walking along curved and winding paths). Ongoing work uses these tools to investigate the development of infant gait modifications.

References

[1] E. J. Gibson and M. A. Schmuckler, "Going somewhere: An ecological and experimental approach to development of mobility," *Ecological Psychology,* vol. 1, pp. 3-25, 1989.

[2] E. J. Gibson, "The concept of affordances in development: The renascence of functionalism," in *The concept of development: The Minnesota Symposia on Child Psychology*, vol. 15, W. A. Collins Ed. NJ: Lawrence Erlbaum Associates, 1982, pp. 55-81.

[3] J. J. Gibson, *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin, 1979.

[4] N. Dominici, Y. P. Ivanenko, G. Cappellini, M. L. Zampagni, and F. Lacquaniti, "Kinematic strategies in newly walking toddlers stepping over different support surfaces," *Journal of Neurophysiology,* vol. 103, pp. 1673-1684, 2010, doi: 10.1152/jn.00945.2009.

[5] K. E. Adolph, "Learning in the development of infant locomotion," *Monographs of the Society for Research in Child Development,* vol. 62, no. 3, Serial No. 251, pp. 1-140, 1997.

[6] K. S. Kretch and K. E. Adolph, "No bridge too high: Infants decide whether to cross based on the probability of falling not the severity of the potential fall," *Developmental Science,* vol. 16, pp. 336-351, 2013.

[7] K. S. Kretch and K. E. Adolph, "The organization of exploratory behaviors in infant locomotor planning," *Developmental Science,* vol. 20, p. e12421, 2017.

[8] G. M. Mulvey, M. Kubo, C. L. Chang, and B. D. Ulrich, "New walkers with Down syndrome use cautious but effective strategies for crossing obstacles," *Research Quarterly for Exercise and Sport,* vol. 82, pp. 210-219, 2011.

[9] K. S. Kretch and K. E. Adolph, "Cliff or step? Posture-specific learning at the edge of a drop-off," *Child Development,* vol. 84, pp. 226-240, 2013.

[10] J. M. Franchak and K. E. Adolph, "What infants know and what they do: Perceiving possibilities for walking through openings," *Developmental Psychology,* vol. 48, pp. 1254-1261, 2012.

[11] S. V. Gill, K. E. Adolph, and B. Vereijken, "Change in action: How infants learn to walk down slopes," *Developmental Science,* vol. 12, pp. 888-902, 2009.

[12] K. E. Adolph and J. E. Hoch, "Motor development: Embodied, embedded, enculturated, and enabling," *Annual Review of Psychology,* vol. 70, pp. 141-164, 2019.

[13] Y. P. Ivanenko, N. Dominici, and F. Lacquaniti, "Development of independent walking in toddlers," *Exercise and Sport Sciences Reviews,* vol. 35, pp. 67-73, 2007.

[14] M. M. Shirley, *The first two years: A study of twenty-five babies. Postural and locomotor development*. Minneapolis, MN: University of Minnesota Press, 1931.

[15] M. B. McGraw, *The neuromuscular maturation of the human infant*. New York, NY: Columbia University Press, 1945.

[16] C. M. Hospodar and K. E. Adolph, "The development of gait and mobility: Form and function in infant locomotion," *Wiley Interdisciplinary Reviews: Cognitive Science (WIREs),* p. e1677, 2024.

[17] S. A. Niyogi and E. H. Adelson, "Analyzing and recognizing walking figures in XYT," in *1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994: IEEE, pp. 469-474.

[18] C. Wan, L. Wang, and V. V. Phoha, "A survey on gait recognition," *ACM Computing Surveys,* vol. 51, pp. 1-35, 2018.

[19] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, "Gait recognition in the wild with dense 3d representations and a benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20228-20237.

[20] Z. Zhu *et al.*, "Gait recognition in the wild: A benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14789-14799.

[21] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[22] C. Zheng *et al.*, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys,* vol. 56, no. 1, pp. 1-37, 2023.

[23] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290-1299.

[24] T. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV*, Zurich, Switzerland, 2014: Springer, pp. 740-755.

[25] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE,* vol. 109, pp. 43-76, 2020.

[26] W. Kentaro. "LabelMe: Image Polygonal Annotation with Python." github.com/wkentaro/labelme (accessed.

[27] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust region methods*. SIAM, 2000.

[28] H.-Y. Hou *et al.*, "Ensemble fusion for small object detection," in *2023 18th International Conference on Machine Vision and Applications (MVA)*, 2023: IEEE, pp. 1-6.

[29] X. Yuan, G. Cheng, K. Yan, Q. Zeng, and J. Han, "Small object detection via coarse-to-fine proposal generation and imitation learning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 6317-6327.

[30] Y. Kondo *et al.*, "MVA2023 Small Object Detection Challenge for Spotting Birds: Dataset, Methods, and Results," in *2023 18th International Conference on Machine Vision and Applications (MVA)*, 2023: IEEE, pp. 1-11.

[31] G. Cheng *et al.*, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2023.