Infants' Developing Environment: Integration of Computer Vision and Human Annotation to Quantify Where Infants Go, What They Touch, and What They See

Danyang Han*
Department of Psychology and Behavioral Sciences
Zhejiang University
Hangzhou, Zhejiang, China
danyang.han@zju.edu.cn
(*shared first authorship)

Nicolas Aziere*
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, US
azieren@oregonstate.edu
(*shared first authorship)

Tieqiao Wang
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, US
wangtie@oregonstate.edu

Ori Ossmy
School of Psychology
Birkbeck, University of London
London, UK
ori.ossmy@bbk.ac.uk

Ajay Krishna
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, US
krishnaj@oregonstate.edu

Hanzhi Wang
Department of Psychology
New York University
New York, NY, US
hw3629@nyu.edu

Ruiting Shen
Department of Psychology
New York University
New York, NY, US
rs8422@nyu.edu

Sinisa Todorovic
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, US
sinisa@oregonstate.edu

Karen Adolph
Depts. of Psychology, Neuroscience, and Child & Adolescent Psychiatry
New York University
New York, NY, US
karen.adolph@nyu.edu

***Abstract*—Infants learn through interactions with the environment. Thus, to understand infants' early learning experiences, it is critical to quantify their natural learning input—where infants go, what they touch, and what they see. Wearable sensors can record locomotor and hand movements, but cannot recover the context that prompted the behaviors. Egocentric views from head cameras and eye trackers require annotation to process the videos and miss much of the surrounding context. Third-person video captures infant behavior in the entire scene but may misrepresent the egocentric view. Moreover, third-person video requires machine or human annotation to make sense of the behaviors, and either method alone is sorely lacking. Computer-vision is not sufficiently reliable to quantify much of infants' complex, variable behavior, and human annotation cannot reliably quantify 3D coordinates of behavior without laborious hand digitization. Thus, we pioneered a new system of behavior detection from third-person video that capitalizes on the integrated power of computer vision and human annotation to quantify infants' locomotor, manual, and egocentric visual interactions with the environment. Our system estimates a real infant's interaction with a physical environment during free play by projecting a "virtual" infant in a "virtual" 3D environment with known coordinates of all furniture, objects, and surfaces. Our methods for using human-in-the-loop computer vision have broad applications for reliable quantification of locomotor, manual, and visual behaviors outside the purview of standard algorithms or human annotation alone.**

Keywords—computer vision, infant, behavior, visual, manual exploration, locomotion

I. Introduction

Many researchers propose that infants actively generate input for learning through interactions with the environment—where infants go, what they touch, and what they see [1-5]. Thus, accurate quantification of infants' natural locomotor, manual, and visual activity is critical to understand their learning experiences and the nature of their self-generated learning input. Infants are not bombarded with the "blooming buzzing confusion" [6] of the surrounding environment because much of the surrounding environment is not immediately available. Babies can only interact with their accessible environment. The accessible environment determines infants' learning input, but what determines access? A long-standing hypothesis is that infants' growing bodies and sensorimotor skills expand infants' accessible environment and alter infants' self-generated learning experiences, but no research systematically quantified the real time and developmental changes in infants' locomotor, manual, and visual inputs. We aim to document how infants' immediate, accessible environment "develops" alongside infants' developing bodies and motor skills. Specifically, we seek to quantify how infants' locomotor, manual, and visual interactions with the environment unfold from moment to moment during natural activity and how the input changes in amount, type, and temporal structure over development.

To achieve this goal, our procedure and method have several desiderata: (1) To capture infants' natural activities, babies should play in a large, fun, complex space in which they can move freely, rather than in a constrained space or while seated at a table. To compare change across development and infants, the play space must be constant and fully calibrated. We chose to quantify infants' locomotor,

This research was supported by Defense Advanced Research Projects Agency Grant N66001-19-2-4035 to Sinisa Todorovic and Karen E. Adolph.

manual and visual experiences because these are infants' primary learning inputs. (2) To record infants' self-generated behaviors, they should play alone with caregivers nearby, but without caregiver intervention to alter infants' behavior. Thus, we omitted language input and social interactions. (3) Our method must capture infants' interactions with the accessible environment rather than recording locomotion, manual actions, or visual behaviors in isolation. We aim to record the places infants visit, the location and identity of objects and surfaces they touch, and the array of places, surfaces, and objects they look at. (4) Our method must record all three modalities simultaneously to reveal their temporal relations. (5) Finally, to record developmental changes from pre-rolling neonates to accomplished infant walkers, our method must be safe, reliable, and valid for all ages from 2-18 months.

A. Wearable Technologies

State-of-the-art, wearable recording technologies do not meet the requirements of our desired dataset. Inertial sensors and machine learning, for example, can estimate the quantity of infant postures and locomotion during natural play and the trajectories of infant movements. But inertial sensors cannot recover infant interactions with the surrounding environment that prompt behavior (e.g., surfaces infants touched, destinations they visited).

Head cameras and head-mounted eye trackers record infants' egocentric visual experiences but require machine-learning algorithms or human annotation to identify the scenes and objects in the videos. Recent machine-learning algorithms did not achieve high accuracy in detecting faces and hands in egocentric videos [7, 8]. Practically, it is impossible to manually annotate the entire array of objects, background scenes, and human bodies in every video frame. Therefore, previous research scored down-sampled subsets of video frames for a few targets of interest (e.g., hands, faces, and a few select objects) [9-11]. Moreover, objects, scenes, and social partners outside infants' field of view are not recorded, so critical information about the surrounding environment is missing, and thus selective attention to the environment is also missing. In addition, head-mounted cameras and eye trackers may constrain infants' natural activity or alter their behavior. Eye trackers also pose safety concerns for infants in a prone posture with their faces and the device near the floor.

B. Computer Vision and Human Annotation Based on Third-Person Video

Third-person videos uniquely capture infants' behavior and the subtle details of the entire scene but require machine or human annotation. Each method alone is severely lacking. Computer vision estimations of body poses frequently err in detection (misses and false alarms); existing algorithms are typically trained on adult datasets and so are less accurate for infant behaviors. Performance of computer-vision algorithms suffers from occlusion, blending effects from similar colors, and so on. Moreover, computer-vision detection from third-person video typically focuses on behaviors separated from the environmental context—precluding analyses of places visited, objects explored, and egocentric visual input.

Human annotation of infant behaviors can include the surrounding environment (e.g., surfaces infants walked on, objects they touched) but cannot reliably quantify 3D coordinates of locomotion and manual interactions without laborious hand digitization (e.g., where infants traveled) [12-14]. Manual annotation cannot reliably estimate infants' entire field of view with existing annotation tools.

C. Current Study

The current study pioneers new forms of infant behavior detection by representing the real infant's behavior in a physical environment with a virtual infant in a virtual environment. We created a 3D virtual environment scaled 1:1 to the physical environment with known 3D coordinates of all objects, furniture, and surfaces. We innovated an integrated system to capitalize on the respective powers of computer vision and human annotation to detect infant behaviors from 2D video and project the behaviors in the corresponding locations of the 3D virtual environment (Fig. 1A). Thus, we know the location of infants' virtual bodies in the surrounding environment, the objects infants' virtual hands touch, and what infants' virtual eyes see from moment to moment. Our current system is also an experiment of the precision of third-person video analysis without wearable technologies to detect locomotor, manual, and visual behaviors.

The critical task is the moment-by-moment detection of the 3D coordinates of infant body keypoints (to track infants' whole-body and hand locations) and infants' face orientation

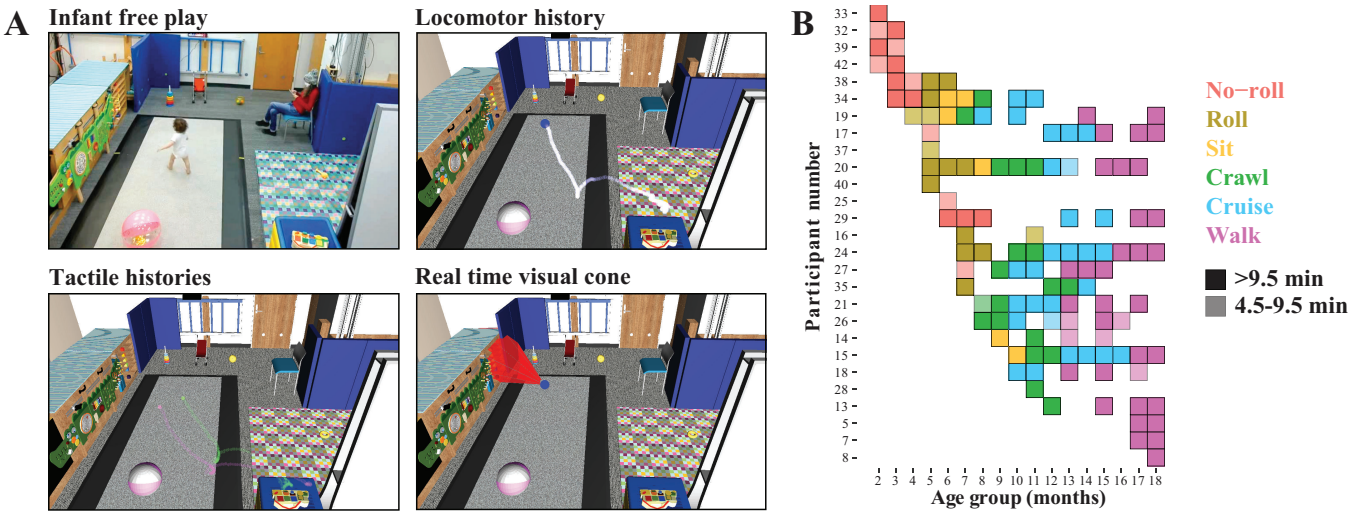

Fig. 1. (A) Top left panel: Real infant in physical playroom. Infants began each session lying supine on a gridded play mat (bottom right of frames). Caregivers sat on chair occupied with phone. Top right and bottom panels: "Virtual" infant in virtual playroom (caregiver not represented). Locomotor history shows 22s of infant head path. Manual histories show 22s of left and right wrists' paths. Red cone represents infant's field of view in 1 video frame. See corresponding video clips at databrary.org/volume/1684/slot/69933. (B) Longitudinal dataset. Each row shows one infant's data, each symbol represents a session, and symbol color denotes locomotor skill in that session. Transparency denotes observation duration.

(to estimate infants' eye gaze direction) in the 3D space based on multiple camera views. Although results from machine annotation alone align with standards for computer vision, accuracy falls short for the intended developmental study.

To increase accuracy, we developed a sophisticated user interface for semi-automated error correction by human annotators. This interface is designed to identify and prompt corrections for the most uncertain results. Human annotators are guided to rectify selected outcomes, and these corrections are propagated to neighboring video frames. This approach ensures high accuracy while minimizing human effort.

II. Dataset and Procedure

A. Dataset

We collected 115 sessions from 27 infants (9 boys, 18 girls). We aimed to test infants monthly from 2 to 18 months, ±1 week of infants' monthly birthday, but infants missed sessions due to holidays, illness, moving, and family schedules. Six infants contributed 1 session and 21 infants contributed 2 to 11 sessions (Fig. 1B). Infants' race and ethnicity were White (48.1%), Black (3.7%), Asian (14.8%), multiracial (33.3%); Hispanic or Latino (18.5%), non-Hispanic or Latino (70.4%) or unknown (11.1%).

As shown in Fig. 1B, infants displayed various postural and locomotor skills: *pre-rolling* if infants laid on their back for the entire session (18 sessions), *rolling* if they rolled from supine to prone (15 sessions), *sitting* if they sat up (6 sessions), *crawling* if they crawled 3m on hands and knees without stopping or falling (18 sessions), *cruising* if they moved 2m sideways holding furniture or wall for support without stopping or falling (27 sessions), and *walking* if infants took upright steps 3m without stopping or falling (31 sessions).

We aimed to obtain 10 min of infant play. Although some *pre-mobile* (pre-rolling, rolling, sitting) sessions were cut short due to fussiness, infants' behavior was less variable relative to *mobile* infants (crawling, cruising, walking). For mobile infants, we analyzed 76 *full-data* sessions (9.5+ min). For pre-mobile infants, we analyzed 28 *full-data* sessions and 11 *partial-data* sessions (4.5 to 9.5 min).

B. Procedure

Infants played in a large laboratory playroom (5.97m × 9.42m) with abundant surfaces and objects at various locations and heights—6 pieces of furniture, 17 movable toys, a wall of toys for display, and varied floor and ceiling surfaces (Fig. 1A). A video tour of the playroom is available at databrary.org/volume/1684/slot/69934. The entire playroom was calibrated with centimeter accuracy from floor to ceiling.

Each session began by placing babies supine in the same location on a large play mat. Caregivers sat on a chair 2m away from infants' starting location and were occupied on their phone. We asked caregivers to refrain from interacting with infants to ensure infants' spontaneous, natural behavior. Infants wore a solid-color onesie so their body and limb movements were clearly visible in cameras for human annotators and easily detected by computer vision due to high color contrast with the environment.

C. Third-person camera recording

We recorded infants with eight fixed, synchronized cameras (standard non-fisheye RGB cameras with 1920 × 2160 resolution for the synchronized videos with 8 views; i.e., 960 × 540 for each view) on the ceiling to cover the entire playroom and to ensure infants were visible in at least two views at each moment. Four camera views were the same across sessions. The other four views differed for mobile and pre-mobile infants. Four views covered the playroom area for mobile infants and four focused on the baby play mat for pre-mobile infants. Camera placement and views are available at databrary.org/volume/1684/slot/69936.

III. Virtual Room and 2D-3D Calibration

Our current system integrates the power of computer vision and human annotation to track infants' locomotor, manual, and visual interactions with the environment solely from third-person video. (1) We built a "virtual" playroom with 1:1 scale relative to the physical playroom with known 3D coordinates of all objects, furniture, and surfaces. (2) We used computer vision to detect infants' bodies and behaviors from the third-person video recordings. (3) We custom built an annotation tool for easy, efficient human correction assisted by computer vision to ensure high accuracy of body-behavior detection with minimal human effort. (4) We projected the virtual infant into the virtual playroom with their bodies and visual cones "colliding" or "touching" the objects and surfaces in the virtual playroom at every frame to reconstruct the real infant's interactions with the real environment (Fig. 1A).

A. Virtual room

We constructed the virtual room in Autodesk Maya. Creating a virtual room to scale with accurate dimensions for floor, ceiling, walls, furniture, and objects is critical. Computer vision algorithms require an accurate 3D representation of the physical playroom to accurately map the virtual room with the third-person video data. We measured the size of the physical playroom with a laser device. We hand-measured the size of objects, furniture, and wall hangings, and the location of all room features to locate them in the virtual room. For ceiling features (e.g., metal frames, lights) not accessible to hand measurement, we used a Matterport 3D camera (go.matterport.com) to measure their size and location and validated them by known distances on the floor. The virtual room matched the colors, patterns, and lighting effects of the physical room to facilitate human annotation. See Fig. 1A for the image of the virtual and physical playroom and databrary.org/volume/1684/slot/69935 for a video tour of the virtual playroom.

B. Calibration between 2D videos and 3D virtual room

We calibrated the 2D pixels in the third-person videos to their corresponding locations in the 3D virtual room. We created a grid (91.2 × 91.2 cm for each unit) on the floor using visually distinctive tapes and then moved a checkerboard with 12 × 12 alternating black and white squares (7.6 × 7.6 cm for each square) along the grid to cover the horizontal and vertical spaces in the room. See video of the calibration process at databrary.org/volume/1684/slot/69937.

We used Microsoft Paint to hand-select pixels on the video recordings that represent a specific location in the physical playroom and mapped the 2D pixel coordinates to the 3D virtual space. We used the camera calibration methods from the widely used OpenCV API [15]. We validated the 2D to 3D mapping algorithm by "back-projecting" known 3D locations to 2D video images. We repeated the process of generating 2D to 3D mapping algorithms with varying sets of 2D to 3D mapping relations until we achieved effective calibration—when the reprojection error remains low for a set of new 3D

points that were not included as input to the calibration algorithm. We observed errors ranging from 0.5 to 5.5 pixels across all 12 views due to slight lens distortion on the edge of the room or noise in the calibration process. Although the ceiling was not captured by the cameras and thus not included in the 2D to 3D mapping sets, the precisely constructed 3D space retains the relative relations between the mapped and unmapped space and thus the entire 3D space was calibrated.

IV. Computer Vision Estimation of Infants' Bodies and Behaviors Prior to Corrections by Human Annotators

We leveraged recent advances in computer vision and Deep Learning algorithms to detect infants and their body pose for each video frame. We used the state-of-the-art body keypoint detection method DeepHighNet [16] to detect infants and their body pose represented as 17 keypoints (nose, left and right eyes, left and right ears, left and right shoulders, left and right elbows, left and right wrists, left and right hips, left and right knees, and left and right ankles). DeepHighNet was initially designed to work on adults, so we used InfantPose [17] specifically trained to work with infants' body proportions. To avoid finetuning a specialized adult/infant detector that would involve annotating a set of additional bounding boxes, we used the pretrained faster-rcnn to extract detections that would be classified later as infant or adult.

Both the infant and caregiver are visible in the videos. DeepHighNet is designed to detect all human bodies in the video frames, so we need to track who the detections belong to across time. We leveraged the inherent constraint that each view contains one adult (caregiver seated on a chair) and one infant in each frame. To keep tracking simple and accurate, we designed a specialized classifier to differentiate adult and infant, and trained it on 3000 manually annotated video images from our own dataset. The training set included 1400-1600 instances of adults and infants, with an additional 200 instances of each class in the test set. Accuracy on the test set was 94.3%. See databrary.org/volume/1684/slot/69933 for an exemplar video of infant behavior with body keypoints.

A. Locomotor and manual interactions

We used infants' head to represent their whole-body location. We extracted the midpoint between left and right eye keypoints to estimate real-time head location. We extracted the left and right wrist keypoints to estimate infants' hand location. See exemplar video of infant behavior with head and wrist trajectories at databrary.org/volume/1684/slot/69933. We compared the computer vision results to the human annotation on 3216 frames. The average error for head location was 12 +/- 38cm, with 78.5% of frames within 10 cm. The wrist locations were evaluated on a different manually-corrected set of 860 instances. The average error of our automated wrist location estimation was 17 +/- 36 cm, with 68.1% of frames within 10 cm. Thus, for the purpose of quantifying infants' interactions with the environment, errors were relatively large based solely on computer vision.

B. Visual interactions

With the current video recording technologies, it was impossible to record infants' eyeball and pupil movement from fixed cameras on the ceiling due to low resolution. In theory, we could record infants' faces with more close-up camera views. However, there is a trade-off between recording the room from a distance to cover wider areas of the room and recording the room with close-up views to ensure high resolution. Because infants' eye movements are typically aligned with their head movement [18-20], we use infants' face orientation to approximate their eye gaze direction.

Within facial analysis, 3D head pose estimation from a single image is pivotal and has applications across various domains. Existing work falls into landmark-based [21], or landmark-free categories [22-27]. Head pose estimation from multiple camera views is a relatively unexplored area in computer vision. Some work uses techniques prior to the Deep Learning era [28-30]. 3D head pose estimation methods primarily targeted adults, assuming smooth head motions in video. However, adapting these methods to infants, with their distinct characteristics and non-smooth movements, poses a significant challenge. Additionally, the common practice of reaching consensus among annotators to establish ground truth in annotation tools within computer vision is well-established, but is largely tailored to adults. The nuanced requirements for infants highlight the need for specialized approaches—a gap our research aims to bridge.

To estimate infants' visual access, we first estimated infants' head location as the middle point of the left and right eyes extracted from body keypoint estimation. Then we estimated infants' face orientation. We employed the most advanced deep head pose estimator 6DRepNet [27] and represented infants' face orientation using 3 Euler angles to describe all orientations—the same format as face pose estimation algorithms using Deep Learning. The model 6DRepNet takes an image representing a person's face as input and produces 3 Euler angles representing the face orientation as output. We cropped the image of the infant face from the original videos as input for 6DRepNet. The dimensions of the cropped face image were adjusted proportionally to the dimensions of the infant's body within the view.

6DRepNet encounters challenges in accurate head pose estimation, especially when infants' faces are not directed toward the camera. Initially, we attempted to use 6DRepNet as a standalone head pose estimator and aggregate predictions from each view by computing the average head pose in the room coordinate system. However, the limitations of monocular head pose models and the difficulty in selecting relevant views led to unsatisfactory results. To address this issue, we harnessed the advantages of our multi-view setup and developed our own multi-view head orientation network. Our model consisted of (1) bounding box feature extractions and (2) multi-view feature fusion. Feature extraction was done using our DeepHighNet network, where each bounding box representing an infant is passed through the network such that the resulting deep features containing body pose information is the concatenation of the set of normalized 2D keypoints ($f_{i,0} \in \mathbb{R}^{(17\times 2)}$) and the d-dimensional deep feature $f_{i,1} \in \mathbb{R}^{d}$ extracted before the body keypoints inference layer. We experimentally found that the concatenation of both features yielded the best results on the test set. For each view i, we have feature $f_i = [f_{i,0}, f_{i,1}] \in \mathbb{R}^{(17\times 2+d)}$, where f_i is a zero vector when view i does not contain any bounding box representing the infant. The set of multi-view features $\{f_i\}_{i=1}^{8}$ is then passed to the fusion network. The fusion network is composed of one linear layer followed by one masked multi-head self-attention layer [31], where attention is computed only between muti-view features containing a valid infant detection. The output of the attention layer is then passed to 2 linear layers. The output of the last layer is the head pose in 3D. It was trained using a L2 norm loss

minimizing the difference between the predicted and ground-truth head poses.

Our multi-view head pose model was trained on manually annotated head poses for multiple rounds. The model provides the initial results of the head poses, human annotators corrected the results, the model was retrained on the corrected annotations, and we repeated the process until we obtained satisfactory results. The annotation procedure was assisted by the custom-built interface to improve the efficiency and accuracy of head pose annotation. The human annotators were trained to reach consensus in annotating the head pose based on files with a wide variety of infant behaviors. See Section VI for details about the annotation process. The manually annotated dataset was composed of 37057 frames for training and 2528 for test. All the frames were selected to provide a robust foundation for accurate head pose detection in diverse scenarios. Our current multi-view model for head pose estimation has an error of 30.6 +/- 26.5 deg on the test set, with 31% within 15 deg. Thus, for the purpose of quantifying infants' visual access to the environment, errors were large based solely on computer vision.

After detecting infants' face orientation, we estimated infants' field of view as a "cone," with the axis as the vector of infants' face orientation and the opening angle of the cone as a narrow field of vision (here, 30 deg, but the field of view is adjustable). See databrary.org/volume/1684/slot/69933 for an exemplar video of infant behavior with the real-time and accumulated history of infants' visual cone. See Fig. 2 for the multiple steps in computer vision estimation.

Occasionally infants moved to areas in the room that were captured by less than two cameras (13.9% of all frames). For frames that capture the infant from one view, the human annotators fill in the missing data of the head and wrist locations and correct the head pose provided by the model. For frames that did not capture the infant from any view (e.g., infant momentarily occluded by furniture), missing data were interpolated based on the nearest known past and future estimations. Full details on the models, networks, and implementation for infant body detection, face orientation detection, and object tracking are publicly available at github.com/azieren/DevEv-BackEnd.

C. *Object tracking*

We tracked the locations of the movable objects with computer vision to obtain real-time 3D coordinates of all objects in the virtual room. The movable toys were tracked in 2D videos using a faster-rcnn detector network [32]. We finetuned the network on a set of toy classes representing the set of movable objects in the room. The ground-truth bounding boxes for moveable objects were annotated on 1008 images. The 2D centroids of the toy bounding boxes were then projected in 3D to estimate toy location in 3D. The locations of a set of small movable objects not captured by the detector will be manually annotated.

V. Reliability of Computer Vision

We encountered several challenges that affect the accuracy of state-of-the-art models—low resolution, instances of partial occlusion and motion blur, and occasional failure to record infants' face. It is crucial to note that the models used in this study were originally trained on extensive public datasets featuring adults. To adapt to the unique characteristics of infants, we fine-tuned our models on a significantly smaller dataset specifically annotated for infants. However, despite these adjustments, the performance still falls short of the requisite accuracy level essential for subsequent data analysis by developmental scientists.

In the computer vision and AI community, it is common to present average performance metrics that include both easy and challenging videos, leading to more optimistic evaluations. However, our developmental study requires high accuracy on every video, not merely on average. This underscores a key gap in collaboration between computer vision and developmental scientists, where the former focus on statistical averages, and the latter value individual performance on each video. Bridging this conceptual gap is essential for effective interdisciplinary collaboration.

Although statistical analyses could potentially remove some noise in the computer vision results, such methods typically require large datasets of samples governed by the same distribution which cannot be satisfied in our case because infants differ in age and behaviors. Thus, corrections by human annotators are currently necessary to compensate for errors in computer vision.

VI. Manual Correction Assisted by Computer Vision

To increase the accuracy of the results, it is essential to leverage human annotation for correcting and quality-assuring computer vision outputs. However, manual correction of 3D information from multiple 2D camera views for every frame is not feasible for humans without assistance. Human coders are not capable of identifying the (x,y,z) coordinates in a 3D environment with their naked eyes. Moreover, the sheer volume of images that need correction makes it a mission

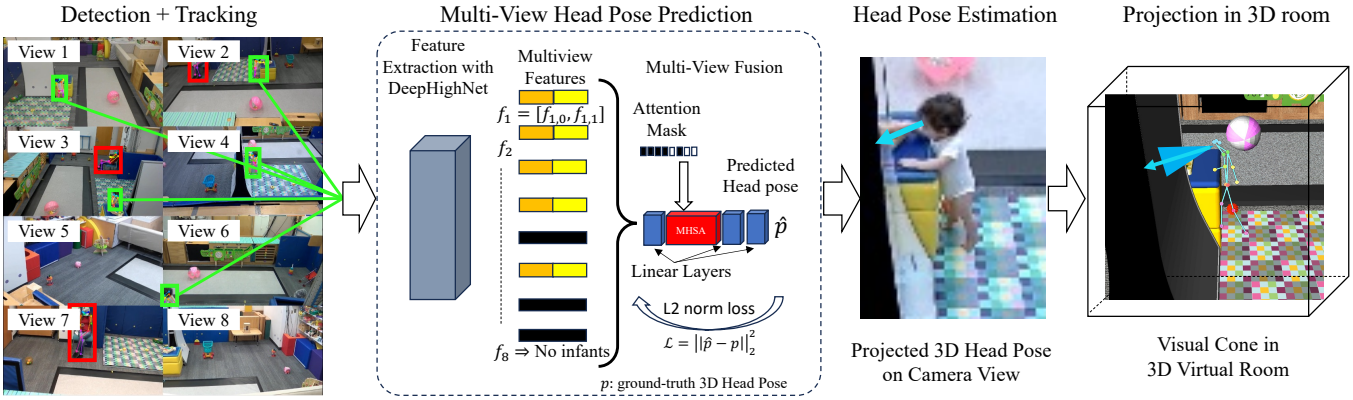

Fig. 2. Computer vision pipeline for multi-view head pose estimation. The adult and infant body keypoints are recognized and tracked across frames for each view (Left). Our multi-view head pose prediction model first extracts features from infant bounding boxes from each view and passes them to the Multi-View Fusion network for predicting the head pose in 3D (Center). The infants' head, wrists, and face orientation are projected into the 3D room (Right).

impossible—each 10-min session has 18,000 frames to correct at 30 frames per second.

We developed a tool that integrates the powers of human annotation and computer vision and compensates for the drawbacks of relying solely on each alone. The annotation tool time-locks the third-person video frames with the corresponding computer vision estimation of the virtual infant in the virtual room. We created a suite of functions in the annotation tool to allow for easy examination and correction of the computer vision results with simple mouse clicks and keyboard presses. The annotation tool propagates manually corrected results to neighboring frames which significantly reduced the number of needed corrections.

Numerous general-purpose image and video labeling software solutions exist, such as the widely adopted LabelMe [33] and VIA [34]. Recognizing the costliness of manual annotation, techniques have emerged to semi-automate the labeling process. For instance, SeTa [35] employs a Cascaded Regressor for eye tracking, training on a small dataset to label a larger one. Similarly, [36] emphasizes minimal human intervention for high accuracy in all circumstances. In our approach, inspired by [35], we select a subset of head poses for correction, leveraging both refinement of the model and temporal smoothness in video to propagate corrections to nearby frames. In domains with high inter-expert variability, like healthcare, collaborative approaches to increase consensus have proven beneficial [37].

A. *Easy, effective user-interface*

Our annotation tool allows users to examine and adjust the 3D projection of estimation results in the virtual room to the correct location based on the 2D video. The tool shows the 2D videos and virtual room side by side. For every video frame, the tool shows the corresponding 3D projection of the estimation results in the virtual room. Coders can watch all 8 views to examine the infant from all angles and get a holistic judgment of the baby's location and face orientation, or can zoom in to one or two views for close-up examination. Annotators can use keyboard and mouse to navigate the virtual room by rotating 360 degrees, scaling, and zooming in and out to examine the 3D location of a point or a vector in the virtual room.

Annotators have two options to set the 3D location of a point (for head and wrists) or a vector (for face orientation) in the virtual room. (1) Annotators can select pixels in the 2D video images that represent a point (e.g., infant's left wrist) in two or more camera views and the annotation tool creates corresponding 3D points or vectors (by specifying the start and end points) in the virtual room based on calibration. (2) Annotators can click any location on objects in the virtual room to set new end points for vectors. They can also use mouse scrolls for fine adjustments of the 3D location of points. To aid in estimating infants' face orientation, annotators can show a sphere with guidelines around the baby's head to provide benchmarks for angle estimation.

After the annotator sets the location of a point or a vector, the annotation tool can visualize the points and vectors back on the 2D videos to verify accuracy. We pre-specified 12 perspectives of the virtual room that were rotated and zoomed in to provide the same views of the room as the 12 video cameras. Coders can use a button click to set the virtual room to any of the 12 views to compare the result in the virtual room and the 2D videos. For a video clip of using the manual annotation tool, see databrary.org/volume/1684/slot/69938.

B. *Human correction assisted by computer vision*

Computer vision estimations achieved high accuracy for head locations relative to wrist locations and face orientation. The errors typically occurred because a background object was erroneously classified as the infant that resulted in excessive movement of the head location between consecutive frames. To minimize manual correction, the annotation tool proposed a set of frames for review based on trajectory smoothness assessment. Annotators examined each identified frame and correct erroneous estimations. Because computer vision achieved relatively low accuracy for wrist locations and face orientation, we adopted a different strategy: The annotation tool proposed 1 frame every second for manual review and the annotator corrected results as needed.

After correction, the annotation tool propagated the corrected results to 15 neighboring frames before and after the corrected frame. Moreover, the tool ensured that frames were sparsely distributed to maximize the power for post-correction propagation. Finally, annotators visually examined results over the entire session with the estimation projected on the 2D video and corrected any remaining errors. Note, the propagation function may remove fast manual movements (e.g., banging).

See databrary.org/volume/1684/slot/69939 for two exemplar infant sessions from one mobile infant and one pre-mobile infant and the estimation of infants' locomotor, manual, and visual behavior based on computer vision and human correction. We showed both the accumulated amount of infant behavior and the process of how the behaviors unfold across the session. With parents' permission, we will openly share raw videos and processed data from other sessions with authorized researchers on Databrary.com.

C. *Reliability of manual correction*

To assess the quality of corrections and the effectiveness of the interface, we engaged multiple annotators to correct the same set of frames from the same videos. Two humans independently annotated 300 frames from 5 sessions (3 from mobile infants and 2 from pre-mobile infants) and reached high agreement with M = 3.03 cm (SD = 0.94) difference in infants' head locations, M = 5.32 cm (SD = 2.23) difference in infants' wrist locations, and M = 13.69 deg (SD = 1.34) difference in infants' face orientation, with 45.3% having less than 10 deg of difference and 6.3% having more than 30 deg of difference.

VII. Collision function: Virtual infant interacting with virtual environment

Finally, we can examine real infants' interactions with the physical environment by estimating the virtual infants' interactions with the virtual environment. For locomotor interactions, we create a sphere representing the infant head with the center as the real-time head location and the diameter as the infant head size. By projecting the sphere to the floor and computing the relative distance between infants' head and the rest of the virtual room, we can examine the unique areas infants visit, surfaces travelled, objects or caregiver near the start and stop of locomotor bouts, and so on. For manual interactions, we create a half-sphere with the center as the wrist and the radius as the infant hand size, with the direction extending from elbows to wrists to represent infants' hands.

We use a collision function to assess what infants' virtual hands collide with in the virtual room—defined as the overlap of the 3D coordinates—to represent the objects and surfaces that infants touch. Likewise, for visual interactions, we assess what the visual cone collides with in the virtual room to represent the entire array of objects and surfaces within infants' field of view.

VIII. Conclusion

Our work offers methodological breakthroughs with a new system that integrates the power of computer vision and human annotation to estimate infants' locomotor, manual, and visual interactions with the environment moment-by-moment solely based on third-person videos. The creation of an accurate virtual 3D environment allows us to quantify infants' behaviors in relation to the environment. We provide a set of procedures and an annotation tool that allow for infant behavior detection without interfering with infants' natural activity or the need for expensive equipment. Our method can be used in other studies that aim to obtain locomotor, manual, and visual behaviors in infant, child, and adult natural, unconstrained activities.

Acknowledgments

We thank Minxin Cheng for her beautiful and precise construction of the "virtual room" and members of New York University Infant Action Lab for data collection.

References

[1] J. Piaget, *The origins of intelligence in children*. New York, NY: International Universities Press, 1952.

[2] L. B. Smith, S. Jayaraman, E. Clerkin, and C. Yu, "The developing infant creates a curriculum for statistical learning," *Trends in Cognitive Sciences,* vol. 22, pp. 325-336, 2018, doi: 10.1016/j.tics.2018.02.004.

[3] K. E. Adolph, "An ecological approach to learning in (not and) development," *Human Development,* vol. 63, pp. 180-201, 2019.

[4] J. J. Campos, D. I. Anderson, M. A. Barbu-Roth, E. M. Hubbard, M. J. Hertenstein, and D. C. Witherington, "Travel broadens the mind," *Infancy,* vol. 1, pp. 149-219, 2000.

[5] E. J. Gibson, "Exploratory behavior in the development of perceiving, acting, and the acquiring of knowledge," *Annual Review of Psychology,* vol. 39, pp. 1-41, 1988, doi: 10.1146/annurev.ps.39.020188.000245.

[6] W. James, *The principles of psychology*. London: Macmillan, 1890.

[7] B. L. Long, G. Kachergis, K. Agrawal, and M. C. Frank, "A longitudinal analysis of the social information in infants' naturalistic visual experience using automated detections," *Developmental Psychology,* vol. 58, pp. 2211-2229, 2022.

[8] B. L. Long, A. Sanchez, A. M. Kraus, K. Agrawal, and M. C. Frank, "Automated detections reveal the social information in the changing infant view," *Child Development,* vol. 93, pp. 101-116, 2022.

[9] C. M. Fausey, S. Jayaraman, and L. B. Smith, "From faces to hands: Changing visual input in the first two years," *Cognition,* vol. 152, pp. 101-107, 2016, doi: 10.1016/j.cognition.2016.03.005.

[10] S. Jayaraman, C. M. Fausey, and L. B. Smith, "Why are faces denser in the visual experiences of younger than older infants?," *Developmental Psychology,* vol. 53, pp. 38-49, 2017.

[11] W. Jayaraman, C. M. Fausey, and L. B. Smith, "The faces in infant-perspective scenes change over the first year of life," *PloS One,* 2015, doi: 10.1371/journal.pone.0123780.

[12] J. E. Hoch, S. M. O'Grady, and K. E. Adolph, "It's the journey, not the destination: Locomotor exploration in infants," *Developmental Science,* vol. 22, p. e12740, 2019.

[13] J. E. Hoch, O. Ossmy, W. G. Cole, S. Hasan, and K. E. Adolph, "'"Dancing" together: Infant-mother locomotor synchrony," *Child Development,* vol. 92, 2021.

[14] J. E. Hoch, J. Rachwani, and K. E. Adolph, "Where infants go: Real-time dynamics of locomotor exploration in crawling and walking infants," *Child Development,* vol. 91, pp. 1001-1020, 2020.

[15] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal: Software Tools for the Professional Programmer,* vol. 25, pp. 120-123, 2000.

[16] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[17] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021, pp. 1-8.

[18] K. S. Kretch and K. E. Adolph, "Active vision in passive locomotion: Real-world free viewing in infants and adults.," *Developmental Science,* vol. 18, pp. 736-750, 2015.

[19] S. Bambach, D. J. Crandall, and C. Yu, "Understanding embodied visual attention in child-parent interaction," in *Proceedings of the IEEE Conference on Development and Learning*, 2013.

[20] H. Yoshida and L. B. Smith, "What's in view for toddlers? Using a head camera to study visual experience," *Infancy,* vol. 13, no. 3, pp. 229-248, May 1 2008, doi: 10.1080/15250000802004437.

[21] D. Strazdas, J. Hintz, and A. Al-Hamadi, "Robo-hud: Interaction concept for contactless operation of industrial cobotic systems," *Applied Sciences,* vol. 11, p. 5366, 2021.

[22] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 2074-2083.

[23] H. W. Hsu, T. Y. Wu, S. Wan, W. H. Wong, and C. Y. Lee, "Quatnet: Quaternion-based head pose estimation with multiregression loss," *IEEE Transactions on Multimedia,* vol. 21, pp. 1035-1046, 2018.

[24] B. Huang, R. Chen, W. Xu, and Q. Zhou, "Improving head pose estimation using two-stage ensembles with top-k regression," *Image and Vision Computing,* vol. 93, p. 103827, 2020.
[25] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," in *31st British Machine Vision Conference*, 2020.
[26] T. T. Yang, Y. T. Chen, Y. Y. Lin, and Y. Y. Chuang, "Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1087-1096.
[27] T. Hempel, A. A. Abdelrahman, and A. Al-Hamadi, "6d Rotation representation for unconstrained head pose estimation," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 2496-2500.
[28] Y. Yan, E. Ricci, R. Subramanian, O. Lanz, and N. Sebe, "No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1177-1184.
[29] Z. Zhang, Y. Hu, M. Liu, and T. Huang, "Head pose estimation in seminar room using multi view face detectors," *Multimodal Technologies for Perception of Humans. CLEAR 2006. Lecture Notes in Computer Science*, R. Stiefelhagen and J. Garofolo, Eds., Berlin: Springer, Berlin, Heidelberg, 2007, pp. 299-304.
[30] M. Voit and R. Stiefelhagen, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in *Proceedings of the 10th international conference on Multimodal interfaces*, 2008, pp. 173-180.
[31] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems 30,* 2017.
[32] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *Proceedings of the IEEE International Conference on Computer vision,* pp. 2961-2969, 2017.
[33] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision,* vol. 77, pp. 157-173, 2008.
[34] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, 2019, pp. 2276-2279.
[35] A. Larumbe-Bergera, S. Porta, R. Cabeza, and A. Villanueva, "SeTA: Semiautomatic tool for annotation of eye tracking images," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, Denver, Colorado, USA, 2019, pp. 1-5.
[36] S. D. Beugher, G. Brône, and T. Goedemé, "A semi-automatic annotation tool for unobtrusive gesture analysis," *Language Resources and Evaluation,* vol. 52, pp. 433-460, 2018.
[37] C. Mata, P. Walker, A. Oliver, J. Martí, and A. Lalande, "Usefulness of collaborative work in the evaluation of prostate cancer from MRI," *Clinics and Practice,* vol. 12, pp. 350-362, 2022.