Infant Video Interaction Recognition using Monocular Depth Estimation

Christopher Rasmussen, Amani Kiruga, Julie Orlando, and Michele A. Lobo

University of Delaware, USA {ras, akiruga, jorlando, malobo}@udel.edu

Abstract. We present a system for automatic analysis of videos of infant actions and infant-caregiver interactions during in-home play sessions. Variables examined include the posture of the infant (sitting, standing, prone, supine, etc.) and whether they are supported in that position by an inanimate object or assisted by a caregiver and at what body location. Leveraging recent advances in neural monocular depth estimation, human body keypoints are lifted from 2-D to 3-D to compute metric distance and angle features, and 3-D scene properties such as the floor plane are estimated to put detections in a global spatial context. We demonstrate strong performance on related pose estimation and posture benchmarks as well as vs. state-of-the-art methods on a challenging new naturalistic video dataset featuring complex interactions in cluttered scenes. We believe that this approach shows promise as a tool for scaling up infant motor developmental studies and is extensible to other developmental domains and age groups.

Keywords: Interaction recognition · pose estimation · monocular depth estimation · object detection

1 Introduction

In this work we describe a novel approach to automatically analyzing videos of household scenes of infants and caregivers at play that were gathered for pediatric physical therapy research. Formal assessments of capacity (what a child can do) and performance (what they actually do) are a crucial part of developmental monitoring, but they can be very labor intensive [10]. The protocol for the study from which our primary dataset (described in detail in Sec. 2) was gathered, for example, requires trained human analysts to annotate video recordings with timestamped "codes" of infant posture, location, object manipulation, parent proximity, and so on. Unfortunately, coding just a few variables over 1 minute of video often takes an experienced analyst 20+ minutes, making scaling up to more subjects, longer interactions, and richer sets of variables impractical.

Object/person detection [2,22] and human pose estimation (aka body keypoint detection) [1,4,23] in images and videos are classical computer vision problems which are closely tied to higher-level tasks like pose classification and action recognition [6,30]. Recently there has been interest in applying these techniques to infants and children for therapeutic and developmental purposes [5,11,15,26].

Different body proportions, posture distributions, and action types vs. adults have led to a variety of approaches, including the development of a linear infant body model [13] and efforts to fine-tune foundation models on smaller infant datasets [9,14,16,31].

Here we focus on extracting from a video the following two time-indexed variables of interest for child development. First, posture (aka *position*): is the infant sitting, standing, prone, or in one of several other possible positions? Second, what is providing *support* in that pose and where is it being provided? Is a caregiver holding them by their torso or upper arms? Are they lying with their head on the floor or sitting independently, unsupported above their hips? Are they leaning against a pillow? (1) can be regarded as a traditional action recognition problem but (2) depends on the *interaction* of the infant with other people and fixed objects in the environment.

It seems self-evident that the essential information required to accurately assess these variables is *spatial*. How far are the infant's body parts from one another? What are the angles of their joints? What is their orientation in a global coordinate frame? Are an adult's hands in contact with them? In each image such relationships are only directly measurable in 2-D, leaving critical ambiguities [16]. Fitting 3-D body models to detected 2-D keypoints can overcome some of these limitations [13,15,16], but is still underconstrained without 3-D information such as from a RGB-D camera [12,16], which may not be available or was not used to collect legacy data. Instead, we seek to demonstrate that recent neural models for inferring scene depth from monocular images [28] have become accurate enough to consider them as aids for this problem. Pseudo-depth images allow coarse 3-D spatial reasoning about distances and angles and also enable reasoning about scene *context* in 3-D. Is the infant standing on the floor or lying on it? Certain camera angles can make it impossible to tell in just 2-D, but with 3-D information there is hope of differentiating these and other tricky situations.

This paper describes a modular, explainable system for combining semantic and structural information, as well as temporal, to reason about infant-caregiver and infant-scene spatial relationships in order to make accurate inferences about actions and interactions. We further introduce a video dataset collected in subject homes that is wide-ranging, naturalistic, and difficult, with cluttered images, bad lighting and framing, and mutual occlusions. We demonstrate strong performance by our system on some related benchmarks as well as this new dataset, and show that it compares very favorably for this task to a range of other approaches including state-of-the-art image understanding systems such as Google Gemini 1.5 Pro and OpenAI ChatGPT-4o.

2 Datasets

PLAY-92 We created a dataset from 92 roughly 10-minute videos of in-home infant-caregiver interactions spanning 41 subjects (twins and triplets were excluded) ranging in age from 1-9 months old. The videos were made – up to several months apart if of the same subject – by researchers as well as parents,

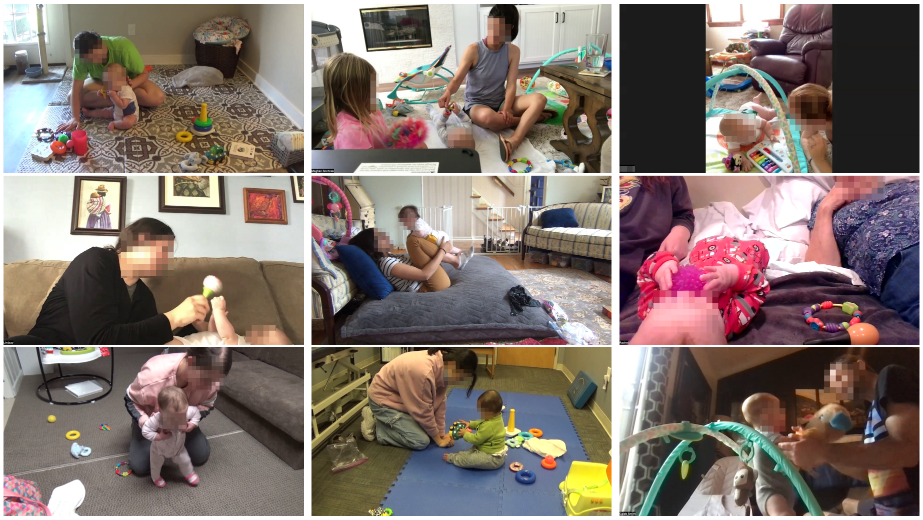


Fig. 1. `PLAY-92` `train_10s` sample images

some from tripod-mounted cameras, some from propped phones, and some as recorded Zoom calls. They exhibit a wide range of quality in terms of scene framing, lighting, and resolution[1]. Example images are shown in Fig. 1.

Two key variables are "coded" (i.e. annotated): (1) infant *position* and (2) infant level of *support*.[2]. Position is overall body posture with 7 categories: `supine`, `prone`, `sit`, and `stand` (as in [16]), but also "in-between" poses `side` (lying but neither prone nor supine) and `reclined` / `inclined` (trunk neither vertical nor horizontal). "All fours" (aka crawling) from [11] is coded here as `prone`.

Level of support is a combination of two factors: the "highest" point on the infant body that is supported, and what provides the support. From high to low, there are 5 coded levels: head and neck, arm and hands, upper trunk, lower trunk, and hips. Two possible sources of support are coded: "p" for another *person* (either actively or passively), and "o" for an inanimate *object* such as the floor, a chair, pillow, exersaucer, etc. Putting these together, there are 10 categories: `headp`, `heado`, `armp`, `armo`, `upp`, `upo`, `lowp`, `lowo`, `hipp`, and `hipo`.

Ground-truth coding consists of start and stop times for action variable values. For example, "the infant position is `prone` from 0:00 to 0:32; then `sit` from 0:33 until 1:07," and so on. Roughly the middle 4 minutes of each session were coded by trained personnel according to a written protocol, yielding 6+ hours of coded video.

[1] 11 videos are 640 × 360, 28 are 1280 × 720, and 53 are 1920 × 1080

[2] Several other variables (infant location, arm and leg mobility, and hand/foot toy interaction; caregiver toy interaction; and infant-caregiver gaze interaction) were coded, but we do not study them here as this is preliminary work

A 60/20/20 split yielded 56 training videos, 18 validation videos, and 18 testing videos. For purposes of training and evaluation here, the videos were sampled at 10-second intervals, resulting in 25 images per video → 1400 training images (`train_10s`), 449 validation images (`val_10s`), and 449 testing images (`test_10s`) (not 450 because the coded range for one video each in `val` and `test` was non-trivially shorter than 240 seconds). Category frequencies for `train_10s` and `val_10s` are shown in Figure 2.

All `PLAY-92` images shown in this paper contain only select study participants who consented to their depiction and dissemination (meaning that the entire raw dataset, unfortunately, cannot currently be distributed). We have further pixellated faces where possible for anonymity.

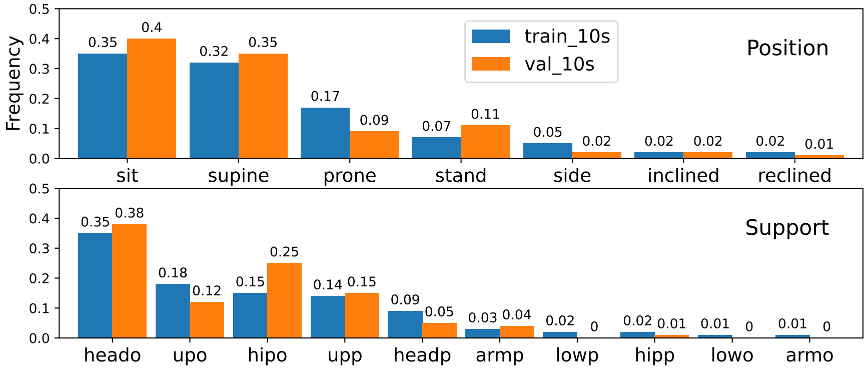


Fig. 2. `PLAY-92` `train_10s` and `val_10s` category frequencies

`SyRIP` [14] The Synthetic/Real Infant Pose dataset includes 700 real infant images ("newborn to one year old") collected from internet photos and video clips, plus 1000 synthetically generated images. The training set consists of 200 real images plus the synthetic images, and the full test set is 500 real images. 100 of the test images are designated a "challenging subset" called Test100. Each image contains exactly one infant and no other people; annotations include a bounding box and 17 ground-truth COCO 2017 keypoints. Some images are from the same videos and thus have the same subject, some images are studio-type photographs with homogeneous backgrounds, some are outdoors, some are pillarboxed vertical video, and some have prominent text and other graphical overlays.

`SyRIP_Posture` [16] This dataset is a re-split of `SyRIP` into 600 real training images and 100 validation images which are the same as `SyRIP`'s Test100, with the addition of a posture label annotation chosen from {`supine`, `prone`, `sit`, `stand`}.

3 Methods

Our overall video processing system, which analyzes interactions between relevant scene entities based on spatial proximity, is diagrammed in Fig. 3. There are two main pathways flowing from the stack of RGB input images in the upper left: *semantic*, in which relevant objects are detected and segmented; and *structural*, in which 3-D depths are estimated per pixel. The current frame is denoted t in a video consisting of or sampled down to T total frames. Several forms of batch temporal analysis are performed in order to ensure consistent interpretations across the T frames, but when images are known to be independent or video membership information is not given (as in the case of `SyRIP`/ `SyRIP_Posture`), such steps are dropped.

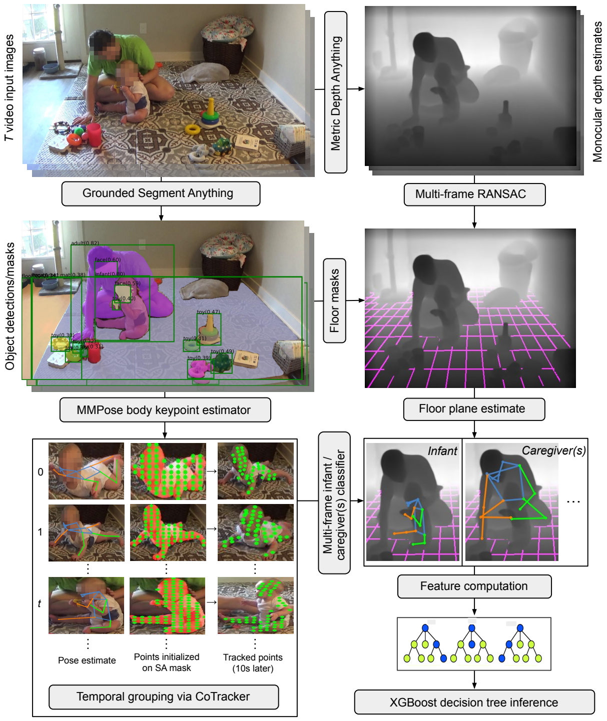

T video input images
Metric Depth Anything
Monocular depth estimates
Grounded Segment Anything
Multi-frame RANSAC
Object detections/masks
Floor masks
MMPose body keypoint estimator
Floor plane estimate
0
1
t
Pose estimate
Points initialized on SA mask
Tracked points (10s later)
Temporal grouping via CoTracker
Multi-frame infant / caregiver(s) classifier
Infant
Caregiver(s)
...
Feature computation
XGBoost decision tree inference

Fig. 3. Steps in video processing pipeline. Grounded Segment Anything [21], Depth Anything [28], MMPose [23], Segment Anything (SA) [20], CoTracker [19], and XGBoost [3] are external libaries. The crops in the temporal grouping section have been scaled to align for display.

For initial semantic analysis, we detect objects of interest in each of the T frames of the video. In this work these are people, specifically infants and potential caregivers including adults and older children; and regions associated with the floor/ground plane. For flexibility in specifying sets of pertinent objects, we use the Grounded Segment Anything (GSA) [21] open object detector, which does not require retraining to ground natural language queries in an image with detection Transformer (DETR) [2] bounding boxes. We use the Swin-B (base) DETR variation, as Swin-L (large) weights that work with GSA were not publicly available. The prompt is a simple noun list – see Sec. 4 for the specific words used. Each detection has a confidence and a segmentation mask obtained by running Segment Anything (SA) [20] on the bounding box. Sample object detection boxes and masks are shown in the middle left image of Fig. 3.

To gain insight into the 3-D scene structure of the input video images, we apply the monocular depth estimation network of Depth Anything [28] to every frame to obtain a stack of T depth images. Specifically, we use the NYUv2-trained *indoor* version of the network to infer nominal metric depths for all pixels. We call this step Metric Depth Anything (MDA) and an example output is shown in the top right of Fig. 3. [28]'s depth estimation network accepts any size input image, but its output is always scaled to 518 × 392, so subsequent steps in the pipeline work in this coordinate system.

3.1 Floor/ground plane estimation

The floor/ground plane establishes the gravity vector, which enables global orientation inference for each person detected in the scene. At the most basic level, plane parameters are estimated through a least-squares fit on the MDA-derived (x, y, z) coordinates of every pixel that the GSA semantic segmentation step deems to belong to the floor or ground. Simply searching for the "floor" or "ground" region is not always sufficient, so we add keywords like "rug" and "carpet" (see Sec. 4 for exact word list) and take the union of all category masks returned to yield the *floor mask*. There are few false positives for the majority of videos (even pillarboxes in vertical videos such as the top right image of Fig. 1 are usually ignored), but misclassified regions necessitate outlier handling. For individual images, robust fitting is achieved through *Random Sample Consensus* (RANSAC) [7], which concurrently classifies 3-D points as inliers or outliers and fits a plane to the inliers. Furthermore, we optionally constrain solutions to "realistic" angular ranges to filter out situations where the plane is erroneously fit to a wall or other vertical surface.

For multi-frame videos, even when the camera is not moving, applying RANSAC to each frame individually can result in variable plane estimates due to several factors: (1) GSA may change entire region classifications based on small pixel changes or occlusions due to people moving around the scene; (2) RANSAC itself is non-deterministic even with the same floor masks; and (3) MDA's depth estimates, while similar, can vary based on camera noise or exposure changes. It is also possible that due to thresholds on the minimum number of inliers or the angle limits mentioned above, the ground plane may not be detected in one frame of a video while it is in the rest.

Thus, we extend our RANSAC scheme to obtain *consistent* multi-frame plane estimates as follows (assuming that the camera is completely static or only repositioned discretely). First, the video is broken into segments in which the camera view is static. This may be part of the dataset annotations or it can be done automatically via shot boundary detection [18]. Next, single-frame RANSAC is applied to all video frames. For each static segment, we measure the support for every frame's candidate plane by summing the number of inliers of other frames' planes whose pitch and roll angles are within some ϵ, and pick the "best" plane solution to use for the entire segment as the one with maximal support. In addition to eliminating noise and outliers vs. single-frame floor plane estimates, this approach also allows accurate floor plane interpolation to individual frames where it is impossible – because of a transient occlusion, for example.

A sample estimated floor plane is visualized as a grid drawn only on the floor mask pixels in the middle right image of Fig. 3. We also note that when the camera pitch angle is level or slightly upward, as is the case in the middle left and bottom right images in Fig. 1, the floor may never be visible and this is explicitly detected.

3.2 Lifting 2-D keypoints to 3-D

The COCO Keypoints 2017 dataset [4] consists of 17 landmarks on the body and face as shown in Fig. 4. Keypoint 0 is the nose, 1-2 are the eyes, 3-4 the ears, and so on. We tested a variety of 2-D body keypoint detection models from the popular open-source library MMPose [23] and observed superior performance with ViTPose-H [27]. Although this and other models (we also tested RTMPose-M and RTMPose-L [17]) performed quite well on infants without any fine-tuning, they were all sensitive to orientation to some degree. That is, upright views of infants often garnered the best keypoint detections, while upside-down and sideways views (associated with prone/supine positions) were correlated with lower quality, presumably because such postures were underrepresented in an adult-dominated training set. To mitigate this, we rotate every infant detection crop 0°, 90°, 180°, and 270° before presenting it to the keypoint detector, select the rotation whose keypoints have the highest median confidence, and use those rotated back to the original orientation. Although the detector gives a location for every keypoint, we treat keypoints with confidences below a threshold τ as "missing."

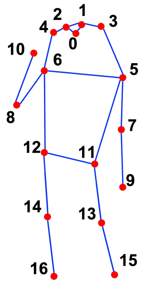


Fig. 4. COCO keypoints [4]

Nominal 3-D keypoint locations are inferred by looking up their depth values in the MDA monocular depth estimate at the 2-D keypoint pixel coordinates (x_i, y_i), as illustrated in Fig. 5. The neurally-inferred metric values themselves have some errors (quantified for general scenes in [28]), but we do not need them to be extremely accurate for our inference to work. Nonetheless, this approach has other issues which may lead to errors. First, the MDA depth image is fairly low resolution and body part details may be blurred or lost. Second, slight misalignments between keypoint locations and body part depth edges may result in depth values being read from the background rather than the body part itself

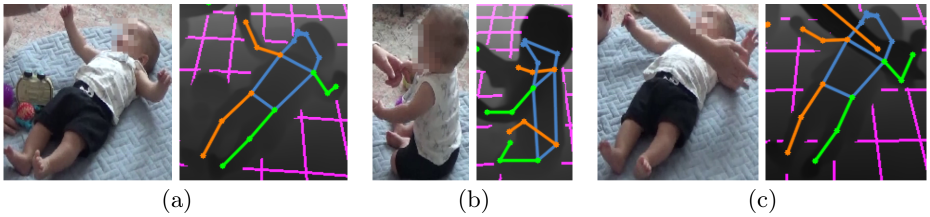


Fig. 5. Sample infant detections from a `PLAY-92` training video with inferred ViTPose-H keypoints [27] and ground plane estimates overlaid on MDA-estimated depth images [28]. Caregiver arm keypoints are visible in (c)

(e.g., the infant's nose in Fig. 5(a)). Finally, the depth value read may not correspond to the actual body part but instead a nearer surface like the front side of the body (e.g., most right-side body parts in Fig. 5(b)) or another person or scene element (e.g., the caregiver's arm blocking the infant's left shoulder and elbow in Fig. 5(c)).

3.3 Consistent infant/caregiver discrimination

In order to correctly compute infant-specific and infant-caregiver interaction features, it is necessary to identify each category of person detected explicitly. We assume that scenes contain a maximum of one infant per image (there are no twins or triplets in `PLAY-92`) and that all other people, including toddler siblings (e.g., in pink in top center of Fig. 1), are non-infants and therefore potential caregivers. Our GSA detector prompt includes keywords "infant" and "adult" (a term which we use interchangeably with "caregiver" and "non-infant"), so it's straightforward to pick the maximum likelihood infant based on the highest confidence Swin-B infant detection. However, this approach yields a small but unacceptably high number of errors. First, the detector sometimes returns multiple bounding boxes on the same person. Second, it occasionally returns false positives on dolls, mirror reflections, and small children or adults in hunched postures.

We had good success filtering out the first type of error with a form of non-maximum suppression based on computing the SA [20] masks $\mathbf{M}_i^t$ for all infant detection bounding boxes i in a frame t and identifying problem pairs $\{\mathbf{M}_i^t, \mathbf{M}_j^t\}$ whose masks have a large overlap relative to their area (e.g., ≥ 0.9). For such pairs, the detection with the lower overlap fraction is discarded, as this tends to eliminate "looser" fits. To reduce the second type of error, we trained an infant/caregiver classifier combining the detection label with 12 3-D body part size features measured from the keypoints and MDA depth estimate. These features were the bilateral lengths of the lower leg, upper leg, forearm, upper arm, and torso (shoulder to hip), plus the shoulder to shoulder and hip to hip distances. The classifier was gradient-boosted decision trees [3], which is discussed more in Sec. 3.4.

If the input is a video (possibly sampled at intervals as with `PLAY-92`'s `train/val_10s`), we temporally group infant detections via tracking to make this classifier's predictions *consistent* vs. running it on each frame independently. Suppose after filtering and per-frame classification we have a set of candidate infant detections and their SA masks $\{\mathbf{M}_i^t\}$ in frame t, and a set $\{\mathbf{M}_j^{t+1}\}$ in the next sampled frame $t+1$. If there are still false positives after filtering and per-frame classification, $i = j$ does not necessarily mean that these are the same infant. So we match as follows: first, grid sets of points $\mathcal{P}_i^t$ are initialized by Co-Tracker [19] from each mask (see middle column of temporal grouping section in Fig. 3). Next, these point sets are tracked *forward* to $t+1$ through intermediate frames at a higher temporal resolution, yielding $\overrightarrow{\mathcal{P}}_i^t$ (shown in the right column of Fig. 3's temporal grouping section). Some tracked points are lost and some are marked non-visible. If enough points "survive," a match score consisting of the fraction inside each infant mask j in the next frame $f(\overrightarrow{\mathcal{P}}_i^t, \mathbf{M}_j^{t+1})$ is computed, and if the highest score is over a threshold, infant detections i in frame t and $\text{argmax}_j f(\overrightarrow{\mathcal{P}}_i^t, \mathbf{M}_j^{t+1})$ in $t+1$ are grouped together. Finally, all nominal infant detections grouped together in a video vote for their infant/caregiver classification and the winning category is imposed on all group members.

3.4 Feature computation and inference

The core of both the *position* and *support* variables of the `PLAY-92` dataset concerns the infant's overall posture, which we believe can be discriminated from 3-D body part distances and angles. We use the following 31 infant-derived features for *position*:

- Height off floor (17): 3-D point to floor plane distance for each body part keypoint, in meters. This should indicate grossly vertical vs. horizontal or bent configurations.
- Torso orientation (4): Angle in degrees between the floor plane normal and normal of plane fit to front of torso. As a proxy for a robustly-fit single plane, we compute this for each of 4 planes defined by triplets taken from the set $\{\text{left hip}, \text{right hip}, \text{left shoulder}, \text{right shoulder}\}$. This can be useful for differentiating between prone and supine, as floor heights alone may still be ambiguous.
- Joint angles (10): 3-D angle in degrees made by the following 5 triplets of keypoints: shoulder-elbow-hand, hip-shoulder-elbow, shoulder-hip-knee, hip-knee-foot, and hip-shoulder-ear, bilaterally. These features provide some posture information even when the floor plane cannot be estimated.

The *support* variable further depends on whether a caregiver is in contact with the infant and at what body "level" that support occurs. To learn this, we start with the 31 *position* features above and add 17 more measuring the proximity of every infant body part to the nearest caregiver hand position, in meters, for a total of 48. By "nearest," we mean hand rather than the entire person. It could be either hand of a single caregiver, or any hand of multiple caregivers in the scene.

To learn from these features, we use the gradient-boosted decision trees paradigm of XGBoost [3]. There are several reasons for this. First, decision trees are very flexible about heterogeneous feature types, including accepting `NaN`'s to represent missing features. Because a missing keypoint means that any feature which references it cannot be computed, and all features that reference the floor plane cannot be computed when a floor plane estimate is unavailable, missing data is an unavoidable characteristic of both *position* and *support* as well as the infant/caregiver classifier of Sec. 3.3. Second, we have relatively little training data compared to the scale typically expected for deep learning. Finally, the learned trees are more easily explainable and amenable to modification than deep neural network weights, for example.

4 Results

4.1 `SyRIP` and `SyRIP_Posture`

Accuracy on COCO 2017 body keypoints is quantified by how many are found within a threshold distance of their ground truth 2-D image locations. In [14] the average precision (AP) of a variety of methods on their own `SyRIP` dataset's Test100 are assessed, and the highest score is achieved by their DarkPose + FiDIP ("fine-tuned domain-adapted infant pose") with an AP of 0.936. In contrast, our MMPose-based ViTPose-H network achieves an AP of 0.987 on the test set with no fine-tuning.

Taking as input the 2-D image coordinates of 12 body keypoints from the 17 produced by the DarkPose + FiDIP keypoint detector of [14], normalized to a common scale, [16] trains a 4-layer fully-connected classifier network on `SyRIP_Posture`'s 4 posture categories {`supine`, `prone`, `sit`, `stand`}. They report a 90.0% test set classification accuracy.

Using our system outlined in Sec. 3, we trained on the 600 training images of `SyRIP_Posture` – with the same gradient-boosted trees and hyperparameters as for the `PLAY-92` *position* results reported in the next section. Because outdoor images are part of this dataset, we modified the GSA prompt to the following: "floor . rug . carpet . play mat . ground . grass . dirt . pavement". We manually masked out color pillarboxes, which sometimes confused GSA, on the assumption that they could easily be detected with a separate filter. The highest confidence infant detection was used and no angular constraints were placed on plane estimates, as there are a number of top-down views not present in `PLAY-92`. We obtained a test set accuracy of 92.0%, and note that 2 of the 8 incorrectly-classified images were studio photographs with all-white backgrounds that resulted in erroneous floor plane estimates.

4.2 `PLAY-92`

We assessed the performance of our full system on `PLAY-92`'s validation subset `val_10s` (`test_10s` is still withheld for future work), which as described in Sec. 2

consists of 18 sequences of 25 frames (4-minute videos sampled at 10 s intervals). Training of the *position*, *support*, and infant/caregiver XGBoost classifiers was carried out on the `train_10s` subset, which is similarly sampled from 56 videos.

The full GSA prompt for object detection was `"infant . adult . floor . rug . carpet . play mat . toy . book . cup . ball . rattle . face . hand . foot"`, and the Swin-B object detection confidence threshold was 0.4. For a number of images the infant was completely or partially out of frame, and the detector found at least one infant in 1342/1400 (95.9%) `train_10s` images and 446/449 (99.3%) `val_10s` images. With a ViTPose-H keypoint confidence threshold of 0.4, an average of 15.43/17 "good" body keypoints were found in each `val_10s` image containing an infant detection.

The angular range for plane fitting was pitch in $[-55°, 25°]$ and roll in $[-15°, 15°]$, and the multi-frame RANSAC angle ϵ was 5°. Tracking in the temporal grouping step of Sec. 3.3 was carried out at 6 fps, the minimum number of surviving tracked points to calculate a match score was 5, and the minimum score to match infant detections was 0.5. The temporal group-based infant/caregiver classifier achieved 99.7% accuracy on `val_10s`.

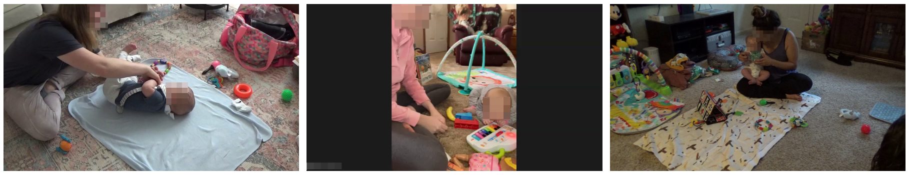
As explained in Sec. 3.4, the feature sets used to train XGBoost on *position* and *support* were slightly different. XGBoost hyperparameters for both had default values except $(\texttt{n\_estimators}, \texttt{eta}, \texttt{max\_depth}, \texttt{min\_child\_weight}) = (50, 0.1, 3, 4)$ for *position* and $(50, 0.1, 2, 5)$ for *support*.

Table 1. Accuracies for all methods on `PLAY-92` `val_10s`. F1 score is weighted.

	Position			Support		
Method	Top-1	Top-2	F1	Top-1	Top-2	F1
ZeroR	40.1	–	22.9	37.6	–	20.6
DINOv2-base [25] MLP	70.8	86.2	62.4	59.5	75.9	49.9
Gemini 1.5 Pro [8]	77.7	–	79.3	59.7	–	61.6
ChatGPT-4o [24]	**83.1**	–	**81.5**	62.6	–	60.3
Ours	82.2	**90.0**	81.0	**73.1**	**87.1**	**71.7**

Classification accuracy results for *position* and *support* are given in Table 1, along with a number of other alternative methods (explained below) that we ran for comparison since `PLAY-92` is a new dataset without established benchmarks. Our system exhibits strong performance in an absolute sense given the difficulty of the images, and is at or near the top of the rankings of all methods tested. Only ChatGPT-4o [24], a state-of-the-art end-to-end text, vision, and audio understanding network with hundreds of billions of parameters that took months to train, was better at anything, scoring 0.9% higher on top-1 *position* accuracy.

We report top-2 accuracies because the ambiguity of intermediate *position* categories like `side`, `inclined`, and `reclined` means that they are often confused with "nearby" categories such as `prone`, `supine`, `sit`, and `stand`. A similar



side : supine

sit : prone

headp : upp

Fig. 6. Selected failure cases on ambiguous/subtle PLAY-92 val_10s images, captioned with ground-truth : prediction. In the middle image the infant's chest is on the floor but they are doubled over as never seen in training. In the right image, the infant's head is touching the caregiver's shoulder, superseding the hands on their torso

issue arises with nearby body levels in *support* – e.g., the difference between the caregiver holding the infant by their upper torso (upp) or lower torso (lowp) may be just a few centimeters. Several such failure cases are shown in Fig. 6.

Alternative methods The so-called "Zero Rule" (ZeroR) classifier, which just picks the highest frequency category in the training data for the variable to be inferred (as seen in Fig. 2), represents minimal performance as it uses no information from the images themselves.

In order to gauge what is possible with a neural network classifier that works from image pixels directly, we trained a multi-layer perceptron (MLP) with 3 layers and 1024 hidden units on DINOv2-base [25] embeddings of the infant detection crops (after infant/caregiver classification). We reasoned that using the infant crops rather than the entire image would simplify the task and still convey sufficient information for accurate inference. This is clear enough for *position*, but we justify excluding the caregiver crops from *support* training by asserting that caregiver hands, if touching the infant, would likely be visible in the infant crop. The input to the MLP was the 768-D class token concatenated to the mean of the 768-D patch tokens → 1536-D. Training was carried out for 50 epochs with a 0.02 learning rate and the Adam optimizer.[3]

Taking this farther with state-of-the-art image understanding systems, we ran OpenAI's flagship multimodal large language model ChatGPT-4o [24] and Google's Gemini 1.5 Pro [8] on each image in val_10s individually. Here we submitted the entire image rather than a crop, allowing a wide range of contextual cues to be analyzed. The prompt included comprehensive definitions of the variables of interest and a request for the inferred coding values in structured JSON format. Both systems had a few spurious responses that were not in the requested categories, and Gemini 1.5 Pro refused to answer for 10 (2.2%) of the images in val_10s because its safety filters blocked the prompt with no explanation. All such responses were simply counted as wrong rather than being corrected with ad-hoc post-processing.

[3] These parameters were the best after a sweep over different numbers of hidden units, learning rates, optimizers, and augmentations

5 Conclusion

This paper has presented promising preliminary results on automatically analyzing infant interaction videos as a tool for monitoring early childhood motor development. We believe that elements of our system could easily be adapted to other early childhood developmental domains, as well as aging in place issues like fall detection, memory assistance, and in-home stroke rehabilitation.

We are currently investigating additional variables from those listed in Sec. 2, particularly toy manipulation/awareness and infant/caregiver gaze directions. Ongoing work includes converting video analysis from batch to online, using higher-resolution monocular depth estimation approaches [29], and integrating 3-D SMIL model fitting with assistance from the depth estimates.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, Z., Martinez, G.H., Simon, T., Wei, S., Sheikh, Y.: OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Analysis & Machine Intelligence (2019)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: IEEE European Conf. Computer Vision (2020)
3. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Int. Conf. Knowledge Discovery & Data Mining (2016)
4. COCO Contributors: COCO 2017 keypoint detection task (2017), https://cocodataset.org/#keypoints-2017
5. Dechemi, A., Karydis, K.: E-babynet: Enhanced action recognition of infant reaching in unconstrained environments. IEEE Trans. Neural Systems & Rehabilitation Engineering (2024)
6. Duan, H., Zhao, Y., Chen, K., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. In: IEEE Conf. Computer Vision & Pattern Recognition (2022)
7. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
8. Google: Gemini Pro (2024), https://deepmind.google/technologies/gemini/pro/
9. Groos, D., Adde, L., Støen, R., Ramampiaro, H., Ihlen, E.: Towards human-level performance on automatic pose estimation of infant spontaneous movements. Computerized Medical Imaging & Graphics (2022)
10. Harbourne, R., Dusing, S., Lobo, M., McCoy, S., Koziol, N., et al.: START-Play physical therapy intervention impacts motor & cognitive outcomes in infants with neuromotor disorder: A multisite randomized clinical trial. Physical Therapy (2021)
11. Hatamimajoumerd, E., Daneshvar, K., Huang, X., Luan, L., Somaieh, S.A., Ostadabbas, S.: Challenges in video-based infant action recognition: A critical examination of the state of the art. In: WACV Workshop on Computer Vision with Small Data: A Focus on Infants and Endangered Animals (2024)

12. Hesse, N., Schröder, A., Müller-Felber, W., Bodensteiner, C., Arens, M., Hofmann, U.: Body pose estimation in depth images for infant motion analysis. In: Int. Conf. IEEE Engineering in Medicine & Biology Society (2017)
13. Hesse, N., Pujades, S., Romero, J., Black, M.J., Bodensteiner, C., et al.: Learning an infant body model from RGB-D data for accurate full body motion analysis. In: Int. Conf. Medical Image Computing & Computer-Assisted Intervention (2018)
14. Huang, X., Fu, N., Liu, S., Ostadabbas, S.: Invariant representation learning for infant pose estimation with small data. In: IEEE Int. Conf. Automatic Face & Gesture Recognition (2021)
15. Huang, X., Luan, L., Hatamimajoumerd, E., Wan, M., Kakhaki, P., Obeid, R., Ostadabbas, S.: Posture-based infant action recognition in the wild with very limited data. In: IEEE Conf. Computer Vision & Pattern Recognition Workshops (2023)
16. Huang, X., Liu, S., Wan, M., Fu, N., Pino, D., Modayur, B., Ostadabbas, S.: Appearance-independent pose-based posture classification in infants. In: Int. Conf. Pattern Recognition Workshop (2022)
17. Jiang, T., Lu, P., Zhang, L., Ma, N., Han, R., Lyu, C., et al.: RTMPose: Real-time multi-person pose estimation based on MMpose. arXiv:2303.07399 (2023)
18. Kar, T., Kanungo, P., Mohanty, S., Groppe, S., Groppe, J.: Video shot-boundary detection: issues, challenges and solutions. AI Review **57**(104) (2024)
19. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. arXiv:2307.07635 (2023)
20. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al.: Segment anything. arXiv:2304.02643 (2023)
21. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., et al.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. arXiv:2303.05499 (2023)
22. MMDetection Contributors: OpenMMLab detection toolbox and benchmark (2024), https://github.com/open-mmlab/mmdetection
23. MMPose Contributors: OpenMMLab pose estimation toolbox and benchmark (2024), https://github.com/open-mmlab/mmpose
24. OpenAI: ChatGPT-4o (2024), https://openai.com/index/hello-gpt-4o
25. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., et al.: DINOv2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)
26. Tafasca, S., Gupta, A., Obobez, J.: ChildPlay: A new benchmark for understanding children's gaze behaviour. In: IEEE Int. Conf. on Computer Vision (2023)
27. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: ViTPose: Simple vision transformer baselines for human pose estimation. In: Advances in Neural Information Processing (2022)
28. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Computer Vision & Pattern Recognition (2024)
29. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv:2406.09414 (2024)
30. Yu, X., Liu, Y., Zhang, X., Zhong, S., Chan, C.: MMNet: A model-based multimodal network for human action recognition in rgb-d videos. IEEE Trans. Pattern Analysis & Machine Intelligence (2022)
31. Yurtsever, M., Eken, S.: BabyPose: Real-time decoding of baby's non-verbal communication using 2d video-based pose estimation. IEEE Sensors (2022)