Posture-based Infant Action Recognition in the Wild with Very Limited Data

Xiaofei Huang[1], Lingfei Luan[1], Elaheh Hatamimajoumerd[1], Michael Wan[1], Pooria Daneshvar Kakhaki[1], Rita Obeid[2], Sarah Ostadabbas[1*]
[1]Northeastern University, MA, USA
[2]Case Western Reserve University, OH, USA
*Corresponding author's email: Ostadabbas@ece.neu.edu

Abstract

Automatic detection of infant actions from home videos could aid medical and behavioral specialists in the early detection of motor impairments in infancy. However, most computer vision approaches for action recognition are centered around adult subjects, following datasets and benchmarks in the field. In this work, we present a data-efficient pipeline for infant action recognition based on the idea of modeling an action as a time sequence consisting of two different stable postures with a transition period between them. The postures are detected frame-wise from the estimated 2D and 3D infant body poses and the action sequence is segmented based on the posture-driven low-dimensional features of each frame. To spur further research in the field, we also created and release the first-of-its-kind infant action dataset—InfAct—consisting of 200 fully annotated home videos representing a wide range of common infant actions, intended as a public benchmark. Among the ten more common classes of infant actions, our action recognition model achieved 78.0% accuracy when tested on InfAct, highlighting the promise of video-based infant action recognition as a viable monitoring tool for infant motor development[1].

1. Introduction

Human action recognition from videos has become an active area of research in recent years due to advancements in computer vision [20, 47]. Typical videos involve human subjects carrying out day-to-day activities in indoor and outdoor settings. While most research has focused on adult subjects—due in part to application objectives such as video surveillance, human-computer interaction, or robotic design—some recent work has centered around children and adolescent subjects, as part of efforts to characterize behavioral or movement disorders [2,7,8,19,27,37,38]. Most recently, our lab has developed *infant*-domain specific computer vision techniques to enable further understanding and characterization of infant development [16, 39, 40, 49]. We aim to extend the benefits of unobtrusive vision-based tools to the domain of *video*-based infant *actions*.

Research on pediatric development has consistently shown links between early motor development in infancy and subsequent cognitive, social, and linguistic development in childhood [18, 22]. For instance, as early as 6 to 9 months of age, infant gross motor movements are synchronized with their early vocalizations [17]. Specifically, their babbling is in-rhyme with their limb activity suggesting that these movements set the stage for speech development. Links have also been found between poor childhood motor skills and developmental delays, including but not restricted to autism spectrum disorders (ASD) and developmental language conditions [10, 32].

However, the majority of the research is conducted with school-aged children or adults due to the nature of the tasks that require children to understand task instructions. There are some studies that look at infant motor development, nonetheless, this work remains scarce. The implication of video-based action recognition for understanding and characterizing infant development holds immense promise for improving medical diagnoses and treatment plans. This technology can help identify at-risk infants, assess the effectiveness of behavioral programs, and promote more meaningful caregiver-infant interactions.

In general, video-based human action can be recognized from multiple vision centered modalities, such as appearance [11, 28], depth [6, 23, 33, 43], optical flow [5, 13, 45], and body skeletons [30, 35, 41]. Within each modality current state-of-the-art recognition networks have deep structure and require large-scale labeled action datasets with sufficient variations to produce robust performance. Building such datasets of videos is much harder than doing so for images, so popular benchmarks for action recognition are

[1]InfAct dataset and infant action recognition model code available at https://github.com/ostadabbas/Infant-Posture-based-Action-Recognition

Table 1. An overview of the existing infant-specific image/video datasets used in computer vision tasks.

Dataset	Content	Purpose	Age Range	# of Samples	Frame Size	Annotations	Public
SyRIP [14]	Real RGB images of infants collected from web and synthetic RGB images	Pose/Posture Recognition	Infant	1,700 images	Varies	17 2D & 3D joints location, 4 posture classes	✓
MINI-RGBD [12]	Synthetic RGB-D videos captured in hospital	Pose Estimation, Medical Infant Motion Analysis	Infant up to 7 months	12 videos (12,000 frames)	640 × 480	24 2D & 3D joints location	✓
BabyPose [26]	Depth videos of preterm infants in cribs hospitalized in NICU	Pose Estimation, Preterm infants' movement pattern recognition	Preterm infants	16 videos (16,000 frames)	640 × 480	12 joints location	✓
XJTU-IDP [44]	Depth videos of infants hospitalized in NICU	Pose Estimation	Infant up to 5 months	27 videos (54,724 frames)	350 × 350	13 joints location	✗
AggPose [4]	RGB videos of infants in supine position	Pose Estimation	Infant	5187 videos (20,748 frames)	Unknown	21 joints location	✓
InfAct (Ours)	RGB videos and images of infants collected from web	Posture/Action Recognition	Infant	200 videos (38126 frames)& 400 images	Varies	5 posture classes, 20 action classes, transition state segmentation	✓

smaller in size[2], having video samples only in the order of 10^3 as supposed to 10^6 in image-based benchmarks. These challenges are magnified in the case of infant action recognition, with no public dataset of infant actions to date.

In this paper, we introduce a novel infant action recognition algorithm that deals with data limitations by representing each infant action as a sequence of a start posture state to a transition state to an end posture state. These postures are the main milestone positions that an infant takes in the first year of their life, defined by the Alberta infant motor scale (AIMS) [29]. Using postures as low-dimensional representations of actions allows the model to be conservative with data usage during supervised steps of the model training. We also present a video segmenter to detect the onset and offset of the transition state and then using the segmentation results and the frame-wise posture-based probabilities, the action label can be determined. To help push the field forward, we have also curate the first-ever infant action dataset comprised of 200 infant videos, called InfAct, with accurate posture state and transition segment annotations.

2. Related Work

Despite the significant progress made in human pose estimation and action recognition, most work is exclusively centered around adult subjects, as evinced by the latest survey [36] published by IEEE transactions on pattern analysis and machine intelligence (TPAMI) in 2022[3]. Infant action recognition is particularly challenging due to the data scarcity, caused by privacy concerns, as well as high variability in infant movements and difficulty in labeling them by non-experts. In this section, we focus on infant-specific computer vision work, first reviewing recent research on capturing infant movements from videos and then listing existing datasets created for these tasks.

Infant Pose, Posture, and Action Recognition– Over the last few years, several approaches ranging from classical image processing techniques to deep learning-based methods have been developed for infant pose estimation. In previous research, we built a domain-adapted infant pose network [14] from a pre-trained adult pose estimation network, trained and tested on a new image-based dataset, called synthetic and real infant pose (SyRIP). [44] proposed a joint feature coding model with a ResNet-50 backbone and key point positional encoding to get high-resolution heatmaps of infant poses. However, this model only focuses on infant poses in supine positions. In [4], the authors proposed a deep aggregation vision transformer framework for infant pose estimation. By leveraging a new large-scale infant dataset, called AggPose, with pose labels and clinical labels, their transformer model could detect infant supine position pose from movement frames in video. Authors in [48] proposed a hierarchical posture classifier based on 3D human pose estimation and scene context information. They combined ResNet-50, stacked hourglass network, and 3D pose estimation scheme for posture classification, and used estimated 3D keypoints to predict infant postures. Nonetheless, the aforementioned studies have been merely developed for image-based infant pose or posture prediction, and there have been limited studies on infant action recognition. [9] proposed BabyNet to capture infant reaching action. BabyNet uses long short-term memory (LSTM) structure to model motion correlation of different phases of a reaching action, but does not cover other infant action types.

Infant Pose, Posture, and Action Datasets– Recently, several infant-specific image/video datasets have been realised, each with their own unique characteristics and applications (see Table 1 for an overview): (1) babyPose [26] contains over 1000 videos of preterm infants aged between 2 and 6 months, captured using a depth-sensing camera along with annotations of 12 limb-joint positions for each frame. However, it only contains the data of newborns with limited supine pose and one-fold background. (2) SyRIP

[2]KTH [21] with 2391 video sequences for 6 actions, NTU-60 [34] with 56880 sequences for 60 actions, and Northwestern-UCLA [42] multi-view action 3D dataset with 1494 video clips for 10 actions.

[3]In several of the existing human pose datasets, such as MPII Pose [3] and MS COCO [24] there are some samples of infant images, nonetheless they are so sparse and not categorized as infant images.

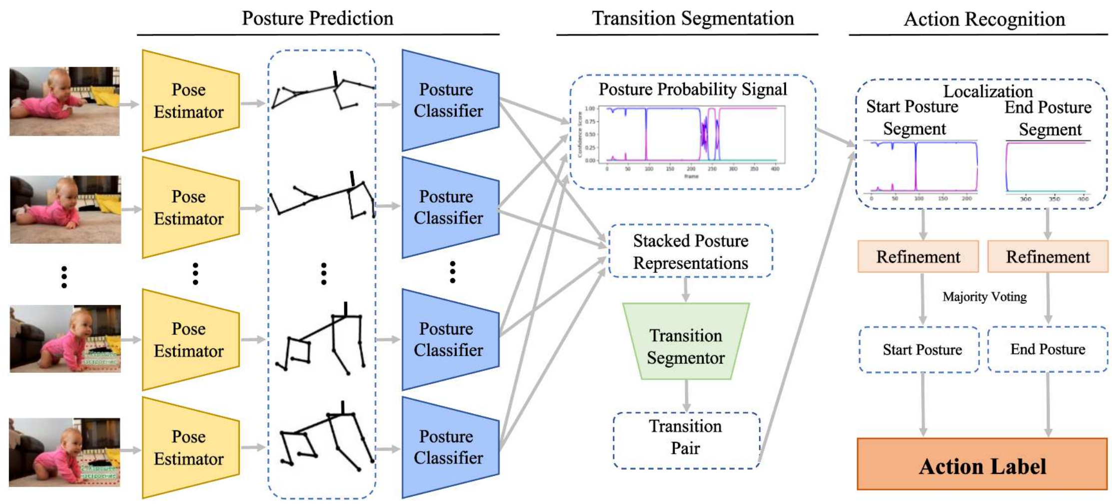


Figure 1. Overview of our proposed infant action recognition method. It mainly consists of (1) Posture Prediction: employing an infant pose estimator to predict the pose of the infant at each frame of the video, and then according to the inferred poses, utilizing an infant pose-based posture classifier to estimate the series of infant postures, (2) Transistion Segmentation: deploying an infant transition segmentator to extract start and end stable posture segments, and (3) Action Recognition: identifying action label for the entire video clip based on the start and end posture labels of the corresponding segments after refinement and majority voting on the posture probability signals. Contents in the dotted boxes indicate intermediate outputs.

[14], from our lab, is an infant pose image dataset including 700 real infant images from YouTube/Google Images and 1000 synthetic infant images generated by rendering skinned multi-infant linear (SMIL) body model with augmented variations in viewpoints, poses, backgrounds, and appearances. 17 joints were annotated for all infant images, and posture labels are also given in four categories (i.e. supine, prone, sitting , and standing) for each real image. Even though this dataset covers various infant poses in the wild, it can only be used to train image-wise models not for dynamic movement learning, such as action or activity recognition. (3) MINI-RGBD [12] was proposed as a benchmark for a standardized evaluation of pose estimation algorithms in infants. It contains RGB and depth images of infants up to the age of 7 months lying in supine position. These images are created by applying SMIL model to build realistic infant body movement sequences with precise 2D and 3D 24 joint positions. (4) AggPose [4] was proposed to train a deep aggregation transformer for human/infant pose detection. They adopted general movements assessment (GMA) devices to record infant movement videos in supine position. More than 216 hours of videos and 15 million frames were extracted. They randomly sampled 20,748 frames from the videos and let professional clinicians annotate infant 21 keypoints locations. Both MINI-RGBD and AggPose have considerable amounts of data. However, they only include infants performing very simple poses in supine positions, and they can only be employed in newborn pose estimation or behavior analysis. The models trained on these dataset do not have ability to handle more complicated poses or movements performed by infants, who are learning to roll over, sit down, or stand up. Therefore, the need for a more general infant action dataset is unmet.

3. Methodology

In general, human action recognition aims to understand human behavior by assigning labels to the actions present in a given video. In the infant behavior domain, we focus on the most common actions, which are related to infant motor development milestones, such as rolling, sitting down, standing up, etc. Here, we present our data-efficient infant recognition model, alongside our novel infant in-the-wild action dataset, consisting of annotated video of infant actions each clipped to feature a single transition between initial and final periods of stable postures (e.g., sitting $\rightarrow$ sit-to-stand transition $\rightarrow$ standing). Our three-part pipeline, illustrated in Figure 1, has the following components: (1) a pose-based infant posture classification model which produces frame-wise posture predictions (and associated probabilities), (2) a transition segmentation model which is trained to predict the start and end times of periods of posture transition (between periods of stable posture), and (3) an action recognition model, which classifies postures in each of the stable posture periods before and after the transition, by smoothing posture prediction probability signals, and then produces a final action label based on those predicted postures.

Problem Formulation– We conceptualize an infant action as a change from one stable posture to another one with a transition period in between, with stable postures defined as those lasting at least one second. Some samples following on this schema are shown in Figure 2. Formally, we represent a video X as sequence of T image frames, $X = (x^1, \ldots, x^T)$. The infant action label of the video takes the form of $A = (p^s, p^e)$, where $p^s, p^e \in \{\text{Supine}, \text{Prone}, \text{Sitting}, \text{Standing}, \text{All-fours}\}$ are the stable start and end postures, respectively. These five critical atomic posture classes are taken from the Alberta infant motor scale (AIMS) guideline [29]. We also assume $p^s \neq p^e$, so there are 20 possible action classes based on the posture combinations. For given action A, the transition period between stable postures is given by $Y = (y^s, y^e)$, with y^s the index of the last frame of the start posture p^s, and $y^e > y^s$ the index of the first frame of the end posture p^e.

3.1. Infant Action Recognition Pipeline

As outlined above, our three-part pipeline, depicted in Figure 1, consists of a posture predictor, a transition segmenter, and an action recognizer.

3.1.1 Posture Prediction

We modify the appearance independent posture classification method from [15] to each frame x^t of the action video sequence X to obtain a posture prediction p^t, for $t \in \{1, \ldots, T\}$. This method in [15] works by first extracting either a 2D or 3D human skeleton pose prediction $J^t \in \mathbb{R}^{N \times D}$, where $N = 12$ is the number of skeleton joints (corresponding to the shoulders, elbows, wrists, hips, knees, and ankles), and $D \in \{2, 3\}$ is spatial dimension of the coordinates. The underlying pose estimators—the fine-tuned domain-adapted infant pose (FiDIP) model [14] for 2D and the heuristic weakly supervised 3D human pose estimation infant (HW-HuP-Infant) model [25] for 3D—were specifically adapted for the infant domain. Then the pose J^t is fed into a 2D or 3D pose-based posture classifier, resulting in the posture prediction p^t. However, the original posture classification model in [15] produces one of four posture classes, so we retrain their network using images representing our five classes extracted from the synthetic and real infant pose (SyRIP) dataset [14]. See Section 4.2 for training details.

3.1.2 Transition Segmentation

To predict the frame indices of the transition period, $Y = (y^s, y^e)$, we adapt a speech sequence segmentation model from [1]. As input, we take the underlying feature vectors $\bar{p} = (\bar{p}^1, \ldots, \bar{p}^T)$ from the last layer of the posture estimation model. The datapoint $\bar{p}$ is used to train the speech sequence segmentation model, a bi-directional recurrent neural network (Bi-RNN), supervised by the ground truth label $Y = (y^s, y^e)$. During training, the model searches through possible start and end transition timings to minimize the loss function measuring a distance between the predicted transition state $\tilde{Y}$ and ground truth label Y.

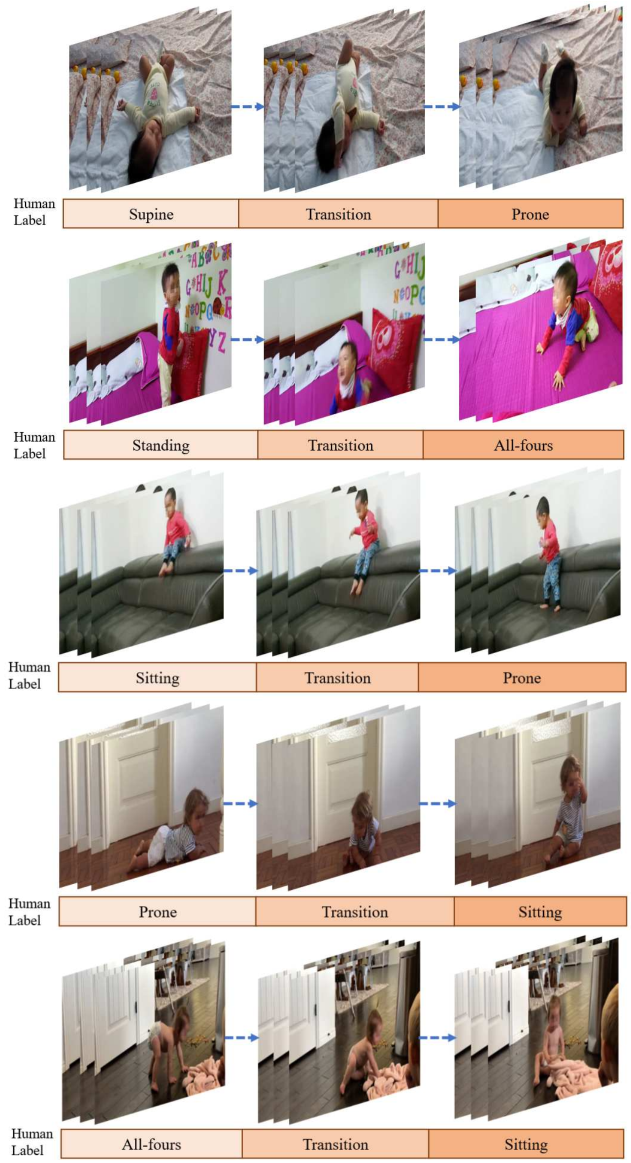


Figure 2. Video examples of InfAct dataset. Here, we exhibit several frames of start posture segment, transition segment, and end posture segment of each video clip. The color bar indicates the ground truth of posture labels and transition segment. All five posture classes are shown here.

3.1.3 Action Recognition

With the initial, transition, and final segments identified, the task that remains is to predict the posture classes of the stable start and end segments, which together entail the overall predicted action class. Our transition segmentation prediction yields frame indices $Y = (y^s, y^e)$, and from these we can derive the sub-sequences of posture predictions for to the start and end stable posture periods, $(p^1, ..., p^{y^s})$ and $(p^{y_e}, ..., p^T)$, respectively. We apply different moving average techniques to smooth out short-term fluctuations and highlight longer-term trends [46], and obtain smoothed posture sequences $(\hat{p}^1, ..., \hat{p}^{y^s})$ and $(\hat{p}^{y_e}, ..., \hat{p}^T)$. Then we aggregate these sequences with majority voting to produce final class estimations $A = (\tilde{p}^s, \tilde{p}^e)$. See Section 4.4 for details on the smoothing methods.

3.2. InfAct: An Infant Action Dataset

In order to enable research in computer vision infant action comprehension, and to provide a testbed for infant action recognition algorithms like ours, we produced a specialized infant action dataset, which we call InfAct, consisting of 200 video clips of infant activities and 400 images of infant postures, with structured action and transition segmentation labels. Figure 2 illustrates the form of the video data, which comprises transitions from a stable starting posture to a stable ending posture.

Our video sourcing and selection procedure was developed by our Nth author, an experienced psychologist. The methodology featured a comprehensive search of public videos from YouTube to obtain a representative cross-section of infant postures and actions, and to ensure the inclusion of a wide range of both infant-specific and general characteristics, including apparent race and ethnicity, stable and transitional postures, and environmental settings. Stringent selection criteria were applied to ensure that postures and transitions were represented consistently and with sufficient duration. After selection, videos were pre-processed and clipped to yield a final set of short videos depicting a transition between a stable strating posture and a stable ending posture, with broad representation of postures on both ends, and movements in the transition stage. Finally, the resulting action clips were annotated with start and end timestamps for the transition period, and labels for the posture classes in the initial and final stable posture periods.

Based on visual inspection and evidence from the source video titles, we estimate that infants in the InfAct dataset range in age from 3 to 18 months. Clip resolutions vary from 720×576 to 1280×720 pixels. Recording environments also vary, with 104 videos from living rooms, 71 videos from bedrooms, 20 from outdoors, 3 videos from bathrooms, 1 recorded from the kitchen, and 1 recorded from the playground. Figure 3 shows the statistical analysis of InfAct.

4. Experimental Results

We use data from InfAct to evaluate the performance of three model components, including posture classification, transition segmentation, and action recognition (described in Section 3.1 and illustrated in Figure 1).

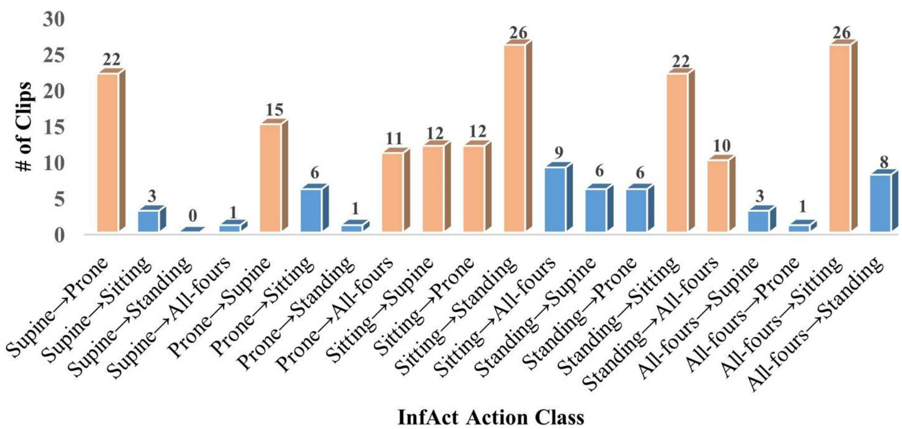
4.1. Datasets

To evaluate our action recognition model, we excluded videos samples from our InfAct dataset having improbable actions, resulting in 156 videos across the 9 action classes highlighted in Figure 3. We used 50 videos for final action recognition test set and the remaining 106 along with the rest of InfAct videos in the InfAct (totally 150 clips) have been used to train the segmentation model. We also created a posture dataset of 400 images by extracting one frame at the beginning and end of each video in InfAct, and defined a 300-100 train-test split. Furthermore, we re-annotated 700 real infant images from SyRIP dataset with our five posture classes (modified from the existing four), and defined a 600-100 train-test split.

4.2. Pose-based Posture Classification

We first trained both the 2D and 3D pose-based posture classification networks on the SyRIP dataset (400 epochs, Adam optimizer, learning rate of 0.00006) using a network with four fully connected layers [15]. We then fine-tuned the trained network with additional InfAct training image (10 epochs, learning rate of 0.001, batch size of 50). We report the posture prediction accuracy scores of both the initial model trained on SyRIP and the fine-tuned model trained further on InfAct in Table 2. These results show that fine-tuning on InfAct notably improves performance, as does adopting the 3D posture model. The fine-tuned 3D pose-based posture model reaches a high overall accuracy of 91.0%. The corresponding prediction confusion matrices are shown in Figure 5 also attest to strong performance. They also reveal a higher-than-typical confusion between the prone and all-fours postures, possibly due to the similarity of these poses, or simply the limited availability of training data.

Visualizations of pose and posture predictions are shown in Figure 4. The first row shows examples in which the posture is correctly predicted with both 2D and 3D pose information as inputs. In examples in the second row, the 3D pose-based posture prediction model succeeds while the 2D pose-based model fails, and in the third row, 3D models fail because predicted 3D poses are wrong. The better performance of the 3D pose-based model could be due to the underlying 3D pose estimations being more robust across a variety of camera angles, resulting in more reliable posture estimations.



Sequences	Duration (s) (min, max)	mean±*std*.
Start Posture	(1.0, 19.1)	3.1±2.1
End Posture	(1.0, 12.6)	3.3±2.0
Overall	(2.7, 20.8)	7.6±3.6

Figure 3. Descriptive Statistics of InfAct dataset. (Left) action classes distribution. The nine common action classes used for our infant action recognition task are highlighted in orange. (Right) duration analysis of video sequences.

Table 2. Performance of our five-posture classification models trained on SyRIP and fine-tuned posture models on InfAct test set in accuracy.

Model	Posture Model	Posture Accuracy (%) Average	Supine	Prone	Sitting	Standing	All-fours
2D	Trained on SyRIP	77.0	**93.8**	**75.0**	67.7	77.8	80.0
	Fine-tuned on InfAct	83.0	87.5	**75.0**	83.9	83.3	86.7
3D	Trained on SyRIP	79.0	81.3	60.0	**93.6**	66.7	86.7
	Fine-tuned on InfAct	**91.0**	**93.8**	75.0	**93.6**	**100.0**	**93.3**

Figure 4. Visualization of our pose-based posture classification performance. For each example, the predicted 2D pose is overlaid on original image, while the predicted 3D body mesh is overlaid on the other cropped one. Ground truth label is given in blue box, 2D pose-based posture prediction is in orange box, and 3D pose-based result is in green box. Wrong predictions are written in red.

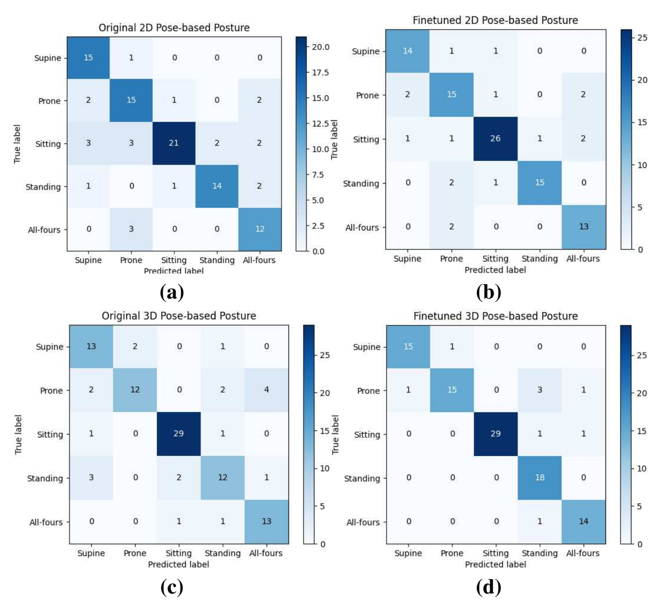


Figure 5. The confusion matrices of infant 2D pose-base and 3D pose-based posture classification models before and after fine-tuning on InfAct training images.

4.3. Posture-based Transition Segmentation

The transition segmenter consists of two bidirectional LSTM layers, each followed by a dropout layer, and is wellsuited to handle variable-length sequences. We trained this network on InfAct data with the following configurations of input derived from the preceding posture estimation model.

Posture Probabilities: For each frame, a vector of five probabilities from the posture estimation model corresponding to each of the five posture classes. The input dimension is $L \times C$ for a sequence of length L and $C = 5$ classes.

Joint Locations: For each frame, a residual vector obtained by applying principle components analysis (PCA) [31] to the sequence of keypoint coordinates for each body joint. The PCA reduction converts coordinate vectors of 17×2 or 17×3 dimensions, depending on the spatial dimension, down to $K = 10$ dimensions, for an overall input dimension of $L \times K$ for a sequence of length L.

Posture Features: For each frame, a residual vector obtained by applying PCA to the feature vector representation of the image in the penultimate layer of the posture estimation model. The PCA reduction converts coordinate vectors down from 16 to $K = 10$ dimensions, for an overall input dimension of $L \times K$ for a sequence of length L.

We train the model with the Adam optimizer at a learning rate of 0.01, with batch size 10. Following the original speech segmentation model [1], the loss for a prediction $\tilde{Y} = (\tilde{y}^s, \tilde{y}^e)$ relative to the ground truth $Y = (y^s, y^e)$ is given by the *structured loss*:

$$\ell(Y, \tilde{Y}) = \sum_{i=s,e} \max\left(0, \left\|y^i - \tilde{y}^i\right\| - \tau\right),$$

with units in frames, and $\tau = 5$ frames is a tolerance factor to allow for natural variations in human annotation. The video framerate is 25 Hz, and each frame is used in the input to the segmentation model. Test results for the transition segmentation model based on structured loss are shown on the left side of Table 3. The results show that, under both the 2D and 3D paradigms, transition segmentation estimation performance is stronger when posture estimation model features (such as classification probabilities or last layer features) are used as input, compared to the raw joint locations. This is to be expected, as in our conceptual framework and in the InfAct dataset, the notion of the transition period is heavily tied to the notion of posture, which the posture estimation model is of course trained to reason about. This is very clear in the visualizations presented in Figure 6, where posture probabilities, transition segments, and video frames are aligned: transitions are strongly correlated with periods of posture prediction uncertainty.

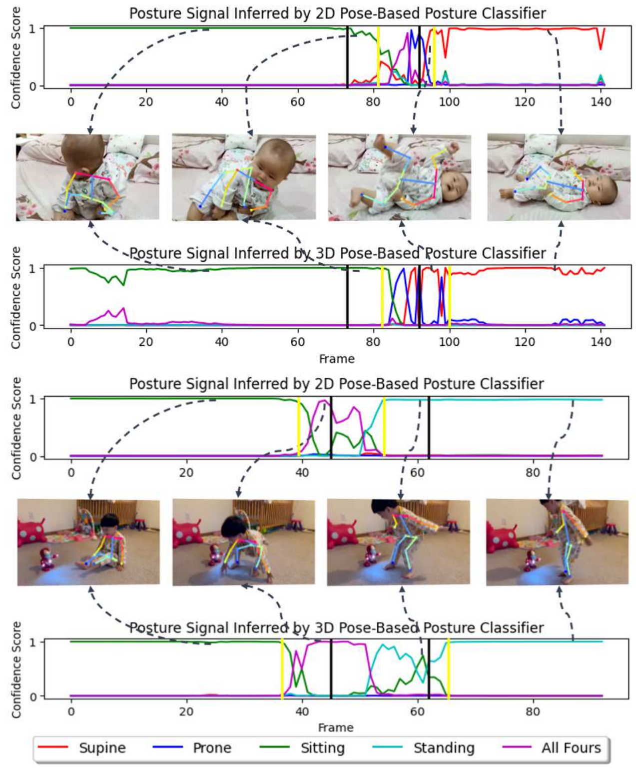


Figure 6. Visualized examples of predicted infant posture probability signals and corresponding estimated transition segmentation results. The vertical yellow lines indicate the predicted index of the last frame of the start posture and the predicted index of the first frame of the end posture respectively, while the black lines indicate the ground truth.

Table 3. Performance of our transition segmentation models on InfAct test videos.

Posture Estimation	Input Sequence	Structured Loss Frames	Structured Loss s
2D Pose-based	Posture Probabilities	39.0	1.6
	Joint Locations	58.5	2.3
	Posture Features	37.7	**1.5**
3D Pose-based	Posture Probabilities	39.2	1.6
	Joint Locations	52.8	2.1
	Posture Features	**36.5**	**1.5**

The results also show that using 3D pose-based posture model features (either model probabilities or last layer features) as input boosts performance over 2D pose-based posture features, but interestingly this advantage is erased when joint locations alone are used as input. The strongest model, which uses 3D pose-based posture model features as input, has an average structured loss of 36.5 frames or $\sim$1.5 s, which is reasonable relative to human perception.

Table 4. Performance of our action recognition method on InfAct test set with different kinds of input sequences by applying different refinement methods.

Posture Estimation	Posture Feature	Transition Segment		Raw Acc. (%)	MA Acc. (%)	EWMA Acc. (%)
2D Pose-based	Posture Preds.	Pred. from	Posture Probs.	64.0	66.0	66.0
			Joint Locs.	54.0	54.0	54.0
			Posture Feats.	62.0	66.0	66.0
		Ground Truth		70.0	72.0	74.0
3D Pose-based	Posture Preds.	Pred. from	Posture Probs.	72.0	72.0	72.0
			Joint Locs.	60.0	62.0	62.0
			Posture Feats.	**78.0**	**78.0**	**80.0**
		Ground Truth		86.0	86.0	86.0

4.4. Posture-based Action Recognition

The final step in our pipeline is to predict posture classes in the starting and ending stable posture periods, and thus infer the final action class label, as detailed in Section 3.1.3. The posture prediction is based on majority voting of the predicted posture class over the two stable posture periods, with start and end timestamps for those stable periods determined by the preceding temporal segmentation model. For our test results, we vary the posture estimation model (2D or 3D), the transition segmentation input format (posture model probabilities, joint coordinate locations, or last-layer posture model features), and also test with the transition segment determined by the ground truth, for reference. Furthermore, while the majority voting is always based on the sequence of predicted posture classes (regardless of which sequence of posture features is fed into the transition segmentation model), we do experiment with two methods of smoothing this sequence to stabilize the raw signal. In particular, we apply a moving average (MA) and an exponentially weighted moving average (EWMA) with a fixed window size of five frames. Taken together, the smoothing and subsequent majority voting produce a single class label for each of the starting and ending stable postures, from which a single overall action class can be inferred for each video clip. The classification accuracy of this final action class label against the ground truth label, for each of the methodological variations we have discussed, is tabulated in Table 4. It should be emphasized that a correct prediction requires that *both* the starting and ending stable class posture be correctly identified, highlighting the roughly "squared" difficulty of the prediction task.

On the whole, the results track and are largely determined by the performance of the underlying transition segmentation model, with segmentation based on 3D pose-based posture estimation coming out on top. 3D-based transition segmentation yields much better results than 2D, as does posture model-based sequential input for transition segmentation compared to joint coordinate location sequential input. Indeed, the extent to which improvements in segmentation results lead to improvements in action recognition is remarkable—a structured loss delta of $\sim$0.8 s between the best and worst segmentation performances yields up to a 24 percentage point gain in action recognition, to 78.0%. Using the ground truth segmentation labels bumps performance further to 86.0%. This may be explained in part by the statistical effect alluded to earlier, wherein the action recognition accuracy is roughly the square of the stable posture estimation accuracy, so improvements in segmentation leading to improvements in stable posture estimation are magnified for final action recognition. We note that the performances of the different smoothing methods are much more balanced, with results slightly favouring the EWMA. The small training and especially test set sizes make it difficult to draw definitive conclusions about the comparative differences among these methods.

The task of infant action recognition from videos is at present characterized by the extreme limitation on available video data, let alone high-quality videos with reliable annotations. In this context, we have proposed a conception, dataset, and action recognition model based around the template of infant actions as transitions between stable postures. The results from our pilot study, based on testing on 9 simple action categories in which we have sufficient data, show the feasibility of our basic approach.

5. Conclusion

Alongside creating InfAct—a pioneering dataset featuring a range of diverse infant actions and equipped with posture and action labels—we developed a data-efficient pipeline for infant action recognition that robustly detects actions given very limited number of samples for each category of common action. The InfAct dataset advances the field of video-based infant action recognition by offering a more accurate and objective mechanism to assess infant motor development. 200 thoroughly annotated home videos show the potential of video-based infant action recognition for motor development monitoring. Our proposed pipeline and dataset can serve as a starting point for future research in this area, which has been under-explored due to the lack of appropriate datasets.

References

[1] Yossi Adi, Joseph Keshet, Emily Cibelli, and Matthew Goldrick. Sequence segmentation using joint rnn and structured prediction models. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2422–2426. IEEE, 2017. 4, 7

[2] Abid Ali, Farhood F Negin, Francois F Bremond, and Susanne Thümmler. Video-based behavior understanding of children for objective diagnosis of autism. In *VISAPP 2022-17th International Conference on Computer Vision Theory and Applications*, 2022. 1

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 2

[4] Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. *In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22) Special Track on AI for Good*, 2022. 2, 3

[5] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1932–1939. IEEE, 2009. 1

[6] Chen Chen, Kui Liu, and Nasser Kehtarnavaz. Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, 12:155–163, 2016. 1

[7] Marco Cristani, Ramachandra Raghavendra, Alessio Del Bue, and Vittorio Murino. Human behavior analysis in video surveillance: A social signal processing perspective. *Neurocomputing*, 100:86–97, 2013. 1

[8] Ryan Anthony J de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):333, 2020. 1

[9] Amel Dechemi, Vikarn Bhakri, Ipsita Sahin, Arjun Modi, Julya Mestas, Pamodya Peiris, Dannya Enriquez Barrundia, Elena Kokkoni, and Konstantinos Karydis. Babynet: A lightweight network for infant reaching action recognition in unconstrained environments to support future pediatric rehabilitation applications. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pages 461–467. IEEE, 2021. 2

[10] Kimberly A Fournier, Chris J Hass, Sagar K Naik, Neha Lodha, and James H Cauraugh. Motor coordination in autism spectrum disorders: a synthesis and meta-analysis. *Journal of autism and developmental disorders*, 40:1227–1240, 2010. 1

[11] David Gerónimo and Hedvig Kjellström. Unsupervised surveillance video retrieval based on human action and appearance. In *2014 22nd International Conference on Pattern Recognition*, pages 4630–4635. IEEE, 2014. 1

[12] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 3

[13] Cheng-Ming Huang, Yi-Ru Chen, and Li-Chen Fu. Real-time object detection and tracking on a moving camera platform. In *2009 ICCAS-SICE*, pages 717–722. IEEE, 2009. 1

[14] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 2, 3, 4

[15] Xiaofei Huang, Shuangjun Liu, Michael Wan, Nihang Fu, David Pino, Bharath Modayur, and Sarah Ostadabbas. Appearance-independent pose-based posture classification in infants. In *ICPR T-CAP Workshops*, 2022. 4, 5

[16] Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, and Sarah Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1909–1917, 2023. 1

[17] Jana M Iverson and Mary K Fagan. Infant vocal–motor coordination: precursor to the gesture–speech system? *Child development*, 75(4):1053–1066, 2004. 1

[18] Jana M Iverson and Esther Thelen. Hand, mouth and brain. the dynamic emergence of speech and gesture. *Journal of Consciousness studies*, 6(11-12):19–40, 1999. 1

[19] Jungseock Joo, Erik P Bucy, and Claudia Seidel. Automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication (19328036)*, 2019. 1

[20] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia tools and applications*, pages 1–27, 2020. 1

[21] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[22] Hayley C Leonard and Elisabeth L Hill. The impact of motor development on typical and atypical social cognition and language: A systematic review. *Child and Adolescent Mental Health*, 19(3):163–170, 2014. 1

[23] Chang Li, Qian Huang, Xing Li, and Qianhan Wu. Human action recognition based on multi-scale feature maps from depth video sequences. *Multimedia Tools and Applications*, 80:32111–32130, 2021. 1

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[25] Shuangjun Liu, Xiaofei Huang, Nihang Fu, and Sarah Ostadabbas. Heuristic weakly supervised 3d human pose estimation in novel contexts without any 3d pose ground truth. *arXiv preprint arXiv:2105.10996*, 2021. 4

[26] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in brief*, 33:106329, 2020. 2

[27] Fabian Nater, Helmut Grabner, and Luc Van Gool. Exploiting simple hierarchies for unsupervised human behavior analysis. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2014–2021. IEEE, 2010. 1

[28] Juan Carlos Niebles and Li Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *2007 IEEE Conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 1

[29] Martha Piper and Johanna Darrah. *Motor Assessment of the Developing Infant-E-Book: Alberta Infant Motor Scale (AIMS)*. Elsevier Health Sciences, 2021. 2, 4

[30] Hossein Rahmani and Mohammed Bennamoun. Learning action recognition model from depth and skeleton videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5832–5841, 2017. 1

[31] Sam Roweis. Em algorithms for pca and spca. *Advances in neural information processing systems*, 10, 1997. 7

[32] Leah Sack, Christine Dollaghan, and Lisa Goffman. Contributions of early motor deficits in predicting language outcomes among preschoolers with developmental language disorder. *International journal of speech-language pathology*, 24(4):362–374, 2022. 1

[33] Adrian Sanchez-Caballero, David Fuentes-Jimenez, and Cristina Losada-Gutiérrez. Exploiting the convlstm: Human action recognition using raw depth video-based recurrent neural networks. *arXiv preprint arXiv:2006.07744*, 2020. 1

[34] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[35] Ahmed Snoun, Nozha Jlidi, Tahani Bouchrika, Olfa Jemai, and Mourad Zaied. Towards a deep human activity recognition approach based on video to image transformation with skeleton data. *Multimedia Tools and Applications*, 80:29675–29698, 2021. 1

[36] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 2

[37] Pieter Vanneste, José Oramas, Thomas Verelst, Tinne Tuytelaars, Annelies Raes, Fien Depaepe, and Wim Van den Noortgate. Computer vision and human behaviour, emotion and cognition detection: A use case on student engagement. *Mathematics*, 9(3):287, 2021. 1

[38] Kathan Vyas, Rui Ma, Behnaz Rezaei, Shuangjun Liu, Michael Neubauer, Thomas Ploetz, Ronald Oberleitner, and Sarah Ostadabbas. Recognition Of Atypical Behavior in Autism Diagnosis from Video using Pose Estimation over Time. In *IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2019. 1

[39] Michael Wan, Xiaofei Huang, Bethany Tunik, and Sarah Ostadabbas. Automatic assessment of infant face and upper-body symmetry as early signs of torticollis. *arXiv preprint arXiv:2210.15022*, 2022. 1

[40] Michael Wan, Shaotong Zhu, Lingfei Luan, Prateek Gulati, Xiaofei Huang, Rebecca Schwartz-Mette, Marie Hayes, Emily Zimmerman, and Sarah Ostadabbas. InfAnFace: Bridging the infant-adult domain gap in facial landmark estimation in the wild. In *2022 International Conference on Pattern Recognition (ICPR)*, 2022. 1

[41] Haoran Wang, Baosheng Yu, Kun Xia, Jiaqi Li, and Xin Zuo. Skeleton edge motion networks for human action recognition. *Neurocomputing*, 423:1–12, 2021. 1

[42] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014. 2

[43] Di Wu, Nabin Sharma, and Michael Blumenstein. Recent advances in video-based human action recognition using deep learning: A review. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2865–2872. IEEE, 2017. 1

[44] Qingqiang Wu, Guanghua Xu, Fan Wei, Jiachen Kuang, Penglin Qin, Zejiang Li, and Sicong Zhang. Supine infant pose estimation via single depth image. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022. 2

[45] Jie Xu, Rui Song, Haoliang Wei, Jinhong Guo, Yifei Zhou, and Xiwei Huang. A fast human action recognition network based on spatio-temporal features. *Neurocomputing*, 441:350–358, 2021. 1

[46] G Udny Yule. The applications of the method of correlation to social and economic statistics. *Journal of the Royal Statistical Society*, 72(4):721–730, 1909. 5

[47] M Zhdanova, V Voronin, E Semenishchev, Yu Ilyukhin, and A Zelensky. Human activity recognition for efficient human-robot collaboration. In *Artificial Intelligence and Machine Learning in Defense Applications II*, volume 11543, pages 94–104. SPIE, 2020. 1

[48] Jianxiong Zhou, Zhongyu Jiang, Jang-Hee Yoo, and Jenq-Neng Hwang. Hierarchical pose classification for infant action analysis and mental development assessment. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1340–1344. IEEE, 2021. 2

[49] Shaotong Zhu, Michael Wan, Elaheh Hatamimajoumerd, Kashish Jain, Samuel Zlota, Cholpady Vikram Kamath, Cassandra B. Rowan, Emma C. Grace, Matthew S. Goodwin, Marie J. Hayes, Rebecca A. Schwartz-Mette, Emily Zimmerman, and Sarah Ostadabbas. A Video-based End-to-end Pipeline for Non-nutritive Sucking Action Recognition and Segmentation in Young Infants, Mar. 2023. arXiv:2303.16867 [cs]. 1