Challenges in Video-Based Infant Action Recognition: A Critical Examination of the State of the Art

Elaheh Hatamimajoumerd [†,1,2], Pooria Daneshvar Kakhaki [†,1], Xiaofei Huang[1], Lingfei Luan[3], Somaieh Amraee[1,2], Sarah Ostadabbas[1*]
[1]Department of Electrical & Computer Engineering,Northeastern University, MA, USA
[2] Roux Institute, Northeastern University, ME, USA
[3] University of Minnesota, MN, USA
[*]Corresponding author's email: Ostadabbas@ece.neu.edu

Abstract

Automated human action recognition, a burgeoning field within computer vision, boasts diverse applications spanning surveillance, security, human-computer interaction, tele-health, and sports analysis. Precise action recognition in infants serves a multitude of pivotal purposes, encompassing safety monitoring, developmental milestone tracking, early intervention for developmental delays, fostering parent-infant bonds, advancing computer-aided diagnostics, and contributing to the scientific comprehension of child development. This paper delves into the intricacies of infant action recognition, a domain that has remained relatively uncharted despite the accomplishments in adult action recognition. In this study, we introduce a groundbreaking dataset called "InfActPrimitive", encompassing five significant infant milestone action categories, and we incorporate specialized preprocessing for infant data. We conducted an extensive comparative analysis employing cutting-edge skeleton-based action recognition models using this dataset. Our findings reveal that, although the PoseC3D model achieves the highest accuracy at approximately 71%, the remaining models struggle to accurately capture the dynamics of infant actions. This highlights a substantial knowledge gap between infant and adult action recognition domains and the urgent need for data-efficient pipeline models[†].

1. Introduction

Automated human action recognition is a rapidly evolving field within computer vision, finding wide-ranging applications in areas such as surveillance , security [23], human-computer interaction [11], tele-health [21], and sports analysis [28]. In healthcare, especially concerning infants and young children, the capability to automatically detect and interpret their actions holds paramount importance. Precise action recognition in infants serves multiple vital purposes, including ensuring their safety, tracking developmental milestones, facilitating early intervention for developmental delays, enhancing parent-infant bonding, advancing computer-aided diagnostic technologies, and contributing to the scientific understanding of child development.

The notion of action in the research literature exhibits significant variability and remains a subject of ongoing investigation [19]. In this paper, we focus on recognizing infants' fundamental motor primitive actions, encompassing five posture-based actions (sitting, standing, supine, prone, and all-fours) as defined by the Alberta infant motor scale (AIMS) [5]. These actions correspond to significant developmental milestones achieved by infants in their first year of life.

To facilitate the accurate recognition of these actions, we employ skeleton-based models, which are notable for their resilience against external factors like background or lighting variations. In comparison to RGB-based models, these skeleton-based models offer superior efficiency. Given their ability to compactly represent video data using skeletal information, these models prove to be especially useful in situations where labeled data is scarce. Therefore, their employment enables a more efficient recognition of the aforementioned hierarchy of infant actions, even with "small data" [15].

While state-of-the-art skeleton-based human action recognition and graphical convolution network (GCN) models [12,30] have achieved impressive performance, they are primarily focused on the adult domain and relied heavily on large, high-quality labeled datasets. However, there

[†]These authors contributed equally to this work.
[†]The code and our data are publicly available at https://github.com/ostadabbas/Video-Based-Infant-Action-Recognition.

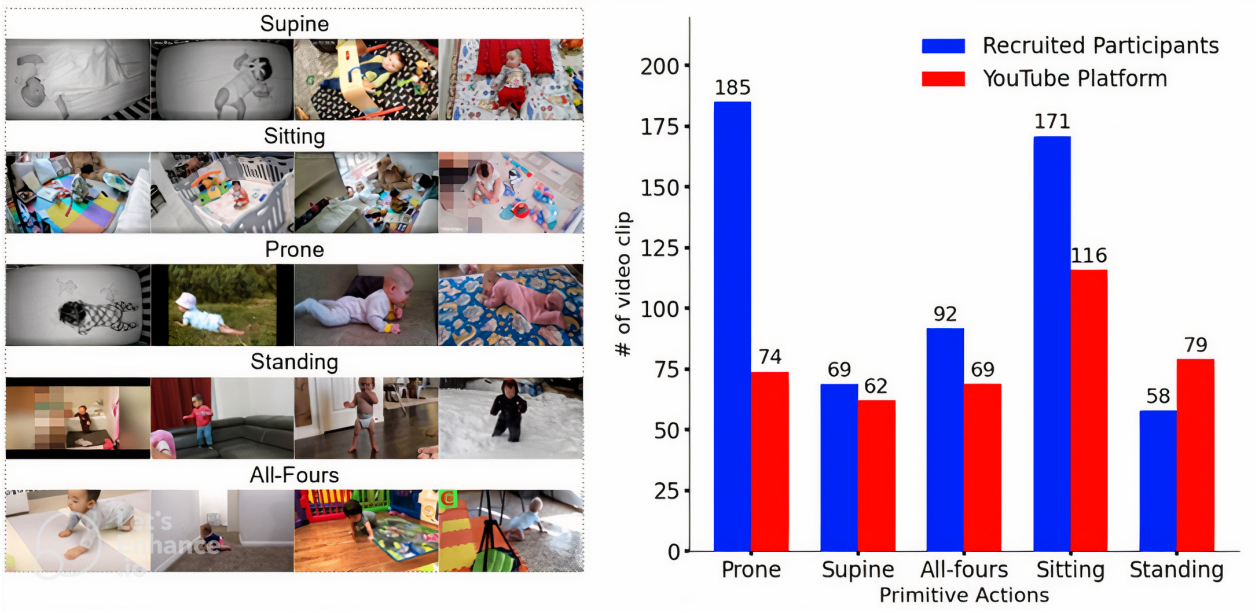

Figure 1. Some snapshots from the InfActPrimitive dataset are displayed on the left side. Each row corresponds to one of the five infant primitive action classes of the dataset. On the right side, the frequency of each action class is depicted, collected from both the YouTube platform and our recruited participants through an IRB-approved experiment.

exists a significant domain gap between the adult and infant action data due to differences in body shape, poses, range of actions, and motor primitives. Additionally, even for the same action, there are discernible differences in how it is performed between infants and adults. For example, sitting for adults often involves the use of chairs or elevated surfaces, providing stability and support, while infants typically sit on the floor, relying on their developing core strength and balance, resulting in different skeleton representations. Furthermore, adult action datasets like "NTU RGB+D" [22]' and "N-UCLA" [26] primarily include actions such as walking, drinking, and waving, which do not involve significant changes in posture. In contrast, infant actions like rolling, crawling, and transitioning between sitting and standing require distinct postural transitions. This domain gap poses significant challenges and hampers the current models' ability to accurately capture the complex dynamics of infant actions.

This paper contributes to the field of infant action recognition by highlighting the challenges specific to this domain, which has been largely unexplored despite the successes in adult action recognition. The limitations in available infant data necessitate the identification of new action categories that cannot be learned from existing datasets. To address this issue, the paper's focus is on adapting action recognition models trained on adult data for use on infant action data, considering the adult-to-infant shift, and employing data-efficient methods.

In summary, this paper introduces several significant contributions:

- A novel dataset called infant action (InfActPrimitive) specifically designed for studying infant action recognition. Figure 1 shows some snapshots of InfActPrimitive. This dataset includes five motor primitive infant milestones as basic actions.
- Baseline experiments conducted on the InfActPrimitive dataset using state-of-the-art skeleton-based action recognition models. These experiments provide a benchmark for evaluating the performance of infant action recognition algorithms.
- Insight into the challenges of adapting action recognition models from adult data to infant data. The paper discusses the domain adaptation challenges and their practical implications for infant motor developmental monitoring, as well as general infant health and safety.

Overall, these contributions enhance our understanding of infant action recognition and provide valuable resources for further research in this domain.

2. Related Work

The existing literature on vision-based human action recognition can be classified into different categories based on the type of input data, applications, model architecture, and techniques employed. This paper focuses on reviewing studies conducted specifically on skeleton data (i.e. 2D or

3D body poses) in human action recognition. Additionally, it discusses the vision-based approaches that have been applied to the limited available infant data.

Recurrent neural network structures methods, such as long short term memory (LSTM) and gated recurrent unit (GRU), treat the skeleton sequences as sequential vectors, focusing primarily on capturing temporal information. However, they often overlook the spatial information present in the skeletons [14]. Shahroody et al. [22] introduced a part-aware LSTM model that utilizes separate stacked LSTMs for processing different groups of body joints, with the final output obtained through a dense layer combination, enhancing action recognition by capturing spatiotemporal patterns. [16] proposed the global context-aware attention LSTM (GCA-LSTM) that incorporates a recurrent attention mechanism that selectively emphasizes the informative joints within each frame.

Graph convolutional network (GCN) has emerged as a prominent method for skeleton-based action recognition. It enables the efficient representation of spatiotemporal skeleton data by encapsulating the intricate nature of an action into a sequence of interconnected graphs. Spatial temporal graph convolution network (ST-GCN) introduced inter-inframe edges, connecting corresponding joints across consecutive frames. This approach enhances the modeling of inter-frame relationships and improves the understanding of temporal dynamics within the skeletal data. InfoGCN [2] combines a learning objective and an encoding method using attention-based graph convolution that captures discriminative information of human actions.

3D convolutional networks capture the spatio-temporal information in skeleton sequences using image-based representations. Wang et al. [27] encoded joint trajectories into texture images using HSV space, but the model performance suffered from trajectory overlapping and the loss of past temporal information. Li et al. [13] addressed this issue by encoding pair-wise distances of skeleton joints into texture images and representing temporal information through color variations. However, their model encountered difficulties in distinguishing actions with similar distances.

Available datasets for human action recognition are mainly incorporate RGB videos with 2D/3D skeletal pose annotations. The majority of the aforementioned studies employed large labeled skeleton-based datasets, such as NTU RGB+D [22], which consisted of over 56 thousand sequences and 4 million frames, encompassing 60 different action classes. The Northwestern-UCLA (N-UCLA) [26] is another widely used skeleton based dataset consists of 1494 video clips featuring 10 volunteers, captured using 3 Kinect cameras from multiple angles to obtain 3D skeletons with 20 joints, encompassing a total of 10 action categories.

Infant-specific computer vision studies have been relatively scarce while there have been notable advancements in computer vision within the adult domain. The majority of these studies have been primarily focused on infant images for tasks such as pose estimation [7, 31], facial landmarks detection [24, 32], posture classification [8, 10], and 3D synthetic data generation [18]. [20] finetuned VGG-16 pretrained with adult faces for infant facial action unit recognition. They applied their methods to the CLOCK [6] and MIAMI [1] datasets, which were specifically designed to investigate neurodevelopmental and phenotypic outcomes in infants with craniofacial microsomia and assess the facial actions of 4-month-old infants in response to their parents, respectively. Zhu et al. [32] proposed a CNN-based pipeline to detect and temporally segment the non-nutritive sucking pattern using nighttime in-crib baby monitor footage. [3] introduced BabyNet that uses a ResNet model followed by an LSTM to capture the spatial and temporal connection of annotated bounding boxes to interpret the onset and offset of reaching and to detect a complete reaching action. However, the focus of these studies has predominantly been on a limited set of facial actions or the detection of specific actions, thereby neglecting actions that involve diverse poses and postures. Huang et al. [9] addressed this issue by creating a small dataset containing a diverse range of infant actions and few samples for each action. The authors developed a posture classification model that was applied on every frame of an input video to extract the posture probability signal. Subsequently, a bi-directional LSTM is employed to segment the signal and estimate posture transitions and the action associated with that transition. Despite presenting a challenging dataset, their action recognition pipeline is not an end-to-end approach.

In this paper, we enhance the existing dataset initially employed in Huang et al.'s study [9] to create a more robust dataset. This expansion involves classifying actions into specific simple primitive motor actions, including "sitting," "standing," "prone," "supine," and "all-fours." Additionally, we collected additional video clips of infants in their natural environment, encompassing both daytime play and nighttime rest, in various settings such as playtime and crib environments. Finally, we tackle the intricate task of infant action recognition through a comprehensive end-to-end approach, with a specific focus on the challenges associated with adapting action recognition models from the adult domain to the unique infant domain.

3. Methods

The goal of a human action recognition framework is to assign labels to the actions present in a given video. In the infant domain, our focus is the most common actions, related to infant motor development milestones. This section introduces our dataset and pipeline for modeling infant skeleton sequences, aiming to create distinct representations for infant action recognition. We begin by introduc-

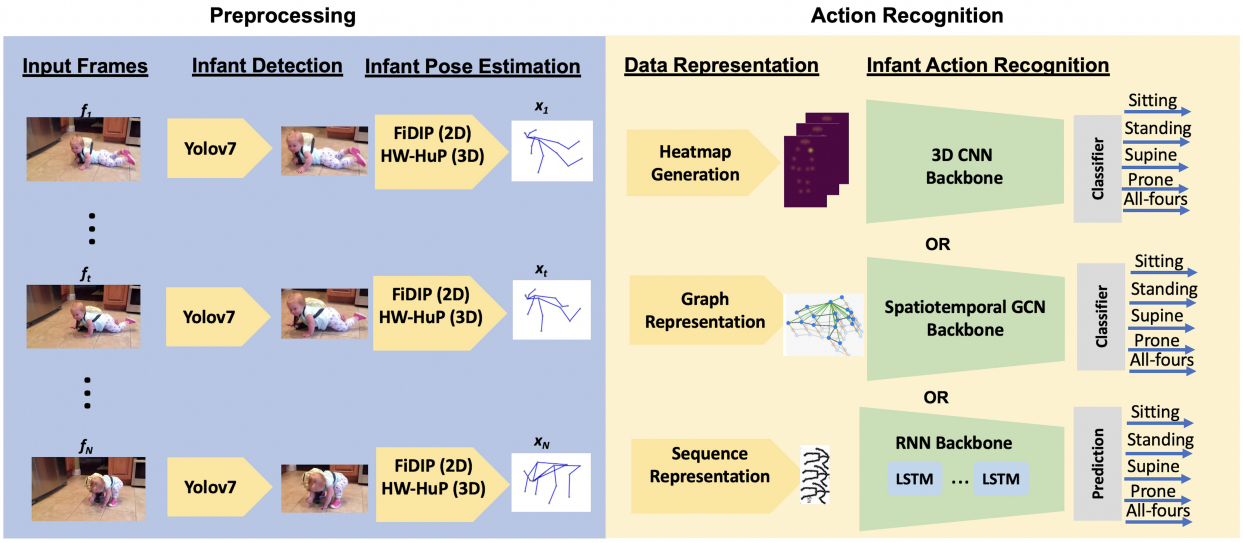

Figure 2. Schematic of the overall infant action recognition pipeline, encompassing infant-specific preprocessing and the action recognition phase. The infant is initially detected in raw frames using YOLOv7 [25] and subsequently serves as input for both 2D and 3D pose estimation facilitated by FiDIP [7] and HW-HuP-Infant [17] algorithms, respectively. The resulting pose information can be further processed into heatmaps, serving as input for CNN-based models, or represented as graphs or sequences for graph- and RNN-based models to predict infant actions.

ing the InfActPrimitive dataset, which serves as the foundation for training and evaluating our pipeline. Subsequently, we delve into the details of the pipeline, which encompasses the entire process from receiving video frames as input to predicting infant action.

3.1. InfActPrimitive Dataset

We present a new dataset called InfActPrimitive as a benchmark to evaluate infant action recognition models. Videos in InfActPrimitive are provided from two sources. (1) Videos submitted by recruited participants: We collected infant videos using a baby monitor from their home and in an unscripted manner. The experiment was approved by the Committee on the Use of Humans as Experimental Subjects of Northeastern university (IRB number:22-11-32). Participants provided informed written consent before the experiment and were compensated for their time. (2) Videos gathered from public video-sharing platforms. This portion of video clips in our dataset has been adapted from [9], which was acquired by performing searches for public videos on the YouTube platform. InfActPrimitive contains 814 infant action videos of five basic motor primitives representing specific postures such as sitting, standing, prone, supine, and all four. The start and end time of every motor primitive is meticulously annotated in this dataset. The InfActPrimitive, with its motor primitives defined by the Alberta Infant Motor Scale (AIMS) as significant milestones, is ideal for developing and testing models for infant action recognition, milestone tracking, and detection of complex actions. Figure 1 shows the screenshots from various videos within the InfActPrimitive dataset, illustrating the diversity of pose, posture, and action among the samples. The diverse range of infant ages and a wide variety of movements and postures within the InfActPrimitive dataset pose significant challenges for action recognition tasks. The right side of the panel in Figure 1 shows the statistical analysis of InfActPrimitive for each sources of data separately.

3.2. Infant Action Recognition Pipeline

Infant specific prepossessing, skeleton data prediction, and action recognition are the key components of our pipeline, as shown in Figure 2. To achieve this, input frames are processed through the pipeline's components, enabling infant-specific skeleton data generation and alignment as input to the different state-of-the-art action recognition models.

Preprocessing– Input video V is represented as sequence of T frames, $V = (f^1, \ldots, f^t, \ldots, f^T)$. We customized the YOLOv7 [25] to locate the bounding box around the infants at every frame as a region of interest. We then extracted either a 2D or 3D infant skeleton pose prediction $x^t \in \mathbb{R}^{J \times D}$, where $J = 17$ 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 [7] for 2D and the heuristic weakly supervised 3D human pose estimation infant (HW-HuP-Infant) model [17] for 3D were specifically adapted for the infant domain.

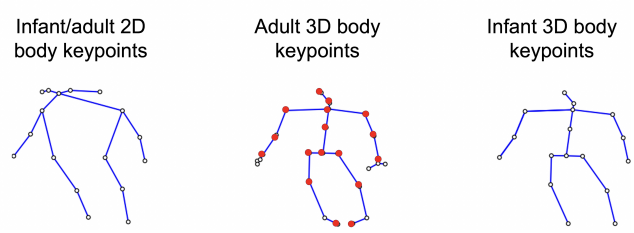

Figure 3. Visualization of three distinct skeleton layouts employed in skeleton-based action recognition datasets. The adult skeleton data adheres to the NTU RGB+D layout, while the 3D version of InfActPrimitive adopts the Human3.6M layout. Action recognition models utilize the common keypoints shared between these layouts, highlighted in red. Additionally, both the 2D versions of adult and infant skeleton data conform to the COCO layout.

Infant-adult skeleton alignment– One of the major challenges in the domain of skeleton-based action recognition lies in the significant variability of skeleton layouts across different datasets and scenarios. The diversity in joint definitions, proportions, scaling, and pose configurations across these layouts introduces complexity that directly impacts the efficacy of action recognition algorithms and makes transferring knowledge between two different datasets inefficient. The challenge of reconciling these layout differences and enabling robust recognition of actions regardless of skeletal variations is a critical concern in our studies.

As shown in Figure 3, NTU RGB+D indicates the location of 25 joints in a 3D space. The layout of the infants 3D skeletons in the InfActPrimitive on the other hand, is based on the Human3.6M skeleton structure, which supports a total of 17 joints. To match the number of keypoints and align the skeleton data in these two datasets, We only select a subset of joins of NTU RGB+D skeleton that are common with the Human3.6M layout. We also reordered these joints, so the structures became as similar as possible. For the 2D skeletons, layouts of both NTU RGB+D and InfActPrimitive are based on the COCO structure.

Action recognition– After preprocessing, we fed the extracted sequence of body keypoints from the input video into various state-of-the-art skeleton-based action recognition models leveraging different aspects of infant-specific pose representations. We categorize these skeleton-based models into three groups: CNN-based, graph-based, and RNN-based models to fully exploit the information encoded in the pose data and perform a comprehensive comparative analysis of the results.

- **Recurrent neural network structures** capture the long-term temporal correlation of spatial features in the skeleton. We applied the part-aware LSTM (P-LSTM) [22] to segment body joints into five part groups and used independent streams of LSTMs to handle each part. At each timeframe t, the input x^t is broken into $(x_1^t, \ldots, x_P^t)$ parts, corresponding to P parts of the body. These inputs are fed into P streams of LSTM modules, where each LSTM has its own individual input, forget, and modulation gates. However, the output gate of these streams will be concatenated and will be shared among the body parts and their corresponding LSTM streams.

- **Graph convolutional networks (GCNs)** represent skeletal data as a graph structure, with joints as nodes and connections as edges. To capture temporal relationships, we applied ST-GCN, which considers inter-frame connections between the same joints in consecutive frames. Furthermore, we employed InfoGCN [2], which integrates a spatial attention mechanism to understand context-dependent joint topology, enhancing the existing skeleton structure. InfoGCN utilizes an encoder with graph convolutions and attention mechanisms to infer class-specific characteristics. μ_c and diagonal covariance matrix of a multivariate Gaussian distribution σ_c. With an auxiliary independent random noise $\epsilon \sim N(0, I)$, Z is sampled as $Z = \mu_c + \Sigma_c \epsilon$. The decoder block of the model, composed of a single linear layer and a softmax function, converts the latent vector Z to the categorical distribution.

- **3D convolutional networks** are mainly employed in RGB-based action recognition tasks to capture both spatial and temporal features across consecutive frames. To utilize the capabilities of a CNN-based framework, We first convert keypoints in each frame into heatmaps. These heatmaps were generated by creating Gaussian maps centered at each joint within the frame. Subsequently, we applied the PoseC3D [4] method, which involved stacking these heatmaps along the temporal dimension, enabling 3D-CNNs to effectively handle skeleton-based action detection. Lastly, the representations extracted from each input sequence using the 3D convolutional layer were fed into a classifier. This classifier consists of a single linear layer followed by a softmax function, ultimately yielding the final class distribution.

4. Experimental Results

In this section, we assess the performance of the models presented in our pipeline. We begin by providing an overview of our experimental setup and the datasets employed. Subsequently, we present the outcomes of various experiments. Finally, we conduct ablation studies and delve into potential avenues for future enhancements.

Table 1. Results of 2D/3D skeleton-based action recognition models using our proposed pipeline on both adult (NTU RGB+D) and infant (InfActPrimitive) dataset. FT denotes that the model was pre-trained on NTU RGB+D during the transfer learning experiments. PoseC3D achieves the best performance on 2D data in both adult and infant datasets. PoseC3D only supports 2D data, and the results in 3D space are marked with ✗.The DeepLSTM model also resulted in very unsatisfactory performance when applied to 3D skeleton data, which we denoted with ✗

	Based on 2D Pose			Based on 3D Pose		
Action Model	NTU RGB+D	InfActPrimitive	InfActPrimitive (+FT)	NTU RGB+D	InfActPrimitive	InfActPrimitive (+FT)
DeepLSTM [22]	87.0	24.3	17.2	✗	✗	✗
ST-GCN [29]	81.5	64.0	66.9	82.5	**67.1**	**69.7**
InfoGCN [2]	91.0	29.7	29.7	**85.0**	29.7	29.7
PoseC3D [4]	**94.1**	**66.9**	**69.7**	✗	✗	✗

4.1. Evaluation Datasets

NTU RGB+D [22] is a large-scale action recognition dataset with both RGB frames and 3D skeletons. This dataset contains 56,000 samples across 60 action classes. Video samples have been captured by three Microsoft Kinect V2 camera sensors concurrently. 3D skeletal data contains the 3D locations of 25 major body joints at each frame. HRNet is used to estimate the 2D pose, which results in the coordination of 17 joints in the 2D space. Given that each video in this dataset features a minimum of two subjects, our approach involves evaluating the models within a cross-subject setting. In this particular setup, the models are trained using samples drawn from a designated subset of actors, while the subsequent evaluation is carried out on samples featuring actors who were not part of the training process. We have employed a train-test split paradigm that mirrors the methodology outlined in [22]. Specifically, we partition the initial cohort of 40 subjects into distinct training and testing groups, with each group composed of 20 subjects. In the context of this evaluative exercise, both the training and testing sets encompass a substantial number of samples, totaling 40, 320, and 16,560, respectively. It is noteworthy to mention that the training subjects for this particular evaluation bear the following identification numbers: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38. The remaining subjects have been thoughtfully reserved for the purpose of conducting rigorous testing.

InfActPrimitive, as detailed in subsection 3.1, combines video clips from two primary sources: data collected from the YouTube platform and data acquired through our independent data collection efforts. To evaluate our pipeline's performance on this dataset, the training set comprises all videos collected from YouTube, totaling 116 (sitting), 79 (standing), 62 (supine), 74 (prone), and 69 (all-fours) actions. Similarly, the test set consists exclusively of videos from our independently collected data, including 171 clips for sitting, 58 clips for standing, 62 clips for supine, 185 clips for prone, and 92 clips for all fours. This partitioning strategy enables us to assess the pipeline's ability to generalize across previously unobserved data and diverse sources, ensuring a comprehensive representation of various actions in both the training and test sets. This approach enhances the robustness of our evaluation by encompassing a wide range of settings and conditions found in YouTube videos and our collected data.

4.2. Experimental Setup

In this section, we detail the series of experiments conducted using our infant action recognition pipeline. We will also provide a comparative analysis, examining the outcomes in relation to the adult skeleton data.

Baseline experiment– In our baseline experiment, we trained various action recognition models, as detailed in subsection 3.2, separately on both the NTU RGB+D and InfActPrimitive datasets from scratch. With the exception of PoseC3D, all these models established baseline performance levels for both 2D and 3D-based action recognition tasks across both adult and infant domains. This baseline performance provides a starting point against which the performance of future experiments, such as fine-tuning or incorporating domain-specific knowledge, can be compared. We set the hyperparameter for ST-GCN, InfoGCN, deepLSTM and PoseC3D models a exactly as they were specified in [29], [2], and [4]. In Table 1, the first pair of columns illustrate the experimental findings with 2D skeleton sequences from both the NTU RGB+D and InfActPrimitive datasets, respectively. Simultaneously, the fourth and fifth columns present the results in the context of 3D data. As demonstrated, PoseC3D consistently outperforms other models in both adult and infant action recognition domains. Nevertheless, a significant performance gap persists between infant and adult action recognition, which can be attributed to disparities in sample size and class distribution. The adult model benefits from a more abundant dataset, enabling it to effectively capture the spatiotemporal nuances of various actions, a characteristic that the InfActPrimitive dataset lacks.

Figure 4 displays the confusion matrices for PoseC3D, InfoGCN, and ST-GCN methods. As illustrated, the se-

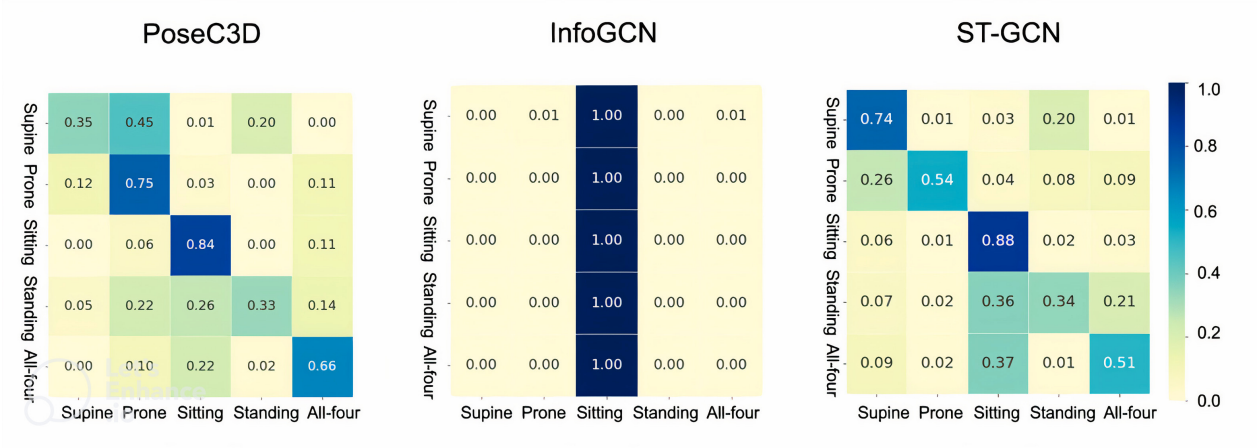

Figure 4. The classification results of three models, along with their respective confusion matrices, are displayed. As shown, InfoGCN faces challenges in achieving clear distinctions between classes, whereas the other models demonstrate varying degrees of proficiency in classifying different primitive categories.

quences associated with the "Sitting" action class exhibit superior separability compared to other classes. However, it is evident that the InfoGCN model miserably fails in the infant action recognition

Transfer learning experiment– To utilize the knowledge embedded in the adult action recognition, we initialized the model weights using the learned parameters obtained from prior training on the NTU RGB+D dataset. To address the substantial class disparities between the two datasets, we excluded the classifier weights, and for this experiment, initialized them randomly.

Given the significant disparity in the number of classes between the two datasets and the substantial impact of training set size on model performance, we chose to delve deeper into the implications of this experimental parameter. Notably, limited data availability posed challenges to achieving high accuracy in models trained on InfActPrimitive. To determine whether this issue extended beyond the domain of infant action recognition, we made modifications to the training subset of NTU RGB+D. Specifically, we curated a subset comprising only five action classes, namely, 'sit down,' 'stand up,' 'falling down,' 'jump on,' and 'drop,' which closely matched those in InfActPrimitive. We then restricted the number of samples per class in this subset to align with the size of the InfActPrimitive training subset. The validation samples for these selected classes remained unchanged.

As shown in Figure 5, the latent variables demonstrate a significantly greater degree of separability within the adult domain compared to the infant domain. This finding highlights the potential limitations of models pretrained on infants in capturing the underlying patterns specific to the infant domain. The disparity can be attributed to the substantial differences between the adult and infant domains, emphasizing the necessity for domain-specific model adaptations or training approaches.

Intra-class data diversity experiment– In our final experiment, we investigate the impact of intra-class diversity on action recognition model performance. We hypothesize that the absence of structural coherence and the inherent variations among samples from the same class can significantly reduce validation accuracy. While traditional action recognition datasets like NTU RGB+D are known for rigid action instructions and minimal intra-class variation, our InfActPrimitive dataset, derived from in-the-wild videos, exhibits a higher level of variability in performed actions. To test this hypothesis, we conducted cross-validation training, dividing our training dataset into five subsets and training on four while validating on the fifth. The original validation set of InfActPrimitive was used for testing. Given the superior results achieved with the PoseC3D model using 2D skeleton data, we considered this model as an infant action recognition model. Our findings, presented in Table 2, shed light on the influence of intra-class diversity on action recognition model performance.

As shown in Table 2, although each experiment yields high training accuracy, there are substantial variations in validation and testing accuracies across experiments. These outcomes reveal discrepancies in the training datasets, leading to inconsistent learning, and underscore distinctions between videos collected from diverse sources.

5. Conclusion

Our work has introduced a unique dataset for infant action recognition, which we believe will serve as an invaluable benchmark for the field of infant action recognition and milestone tracking. Through our research,

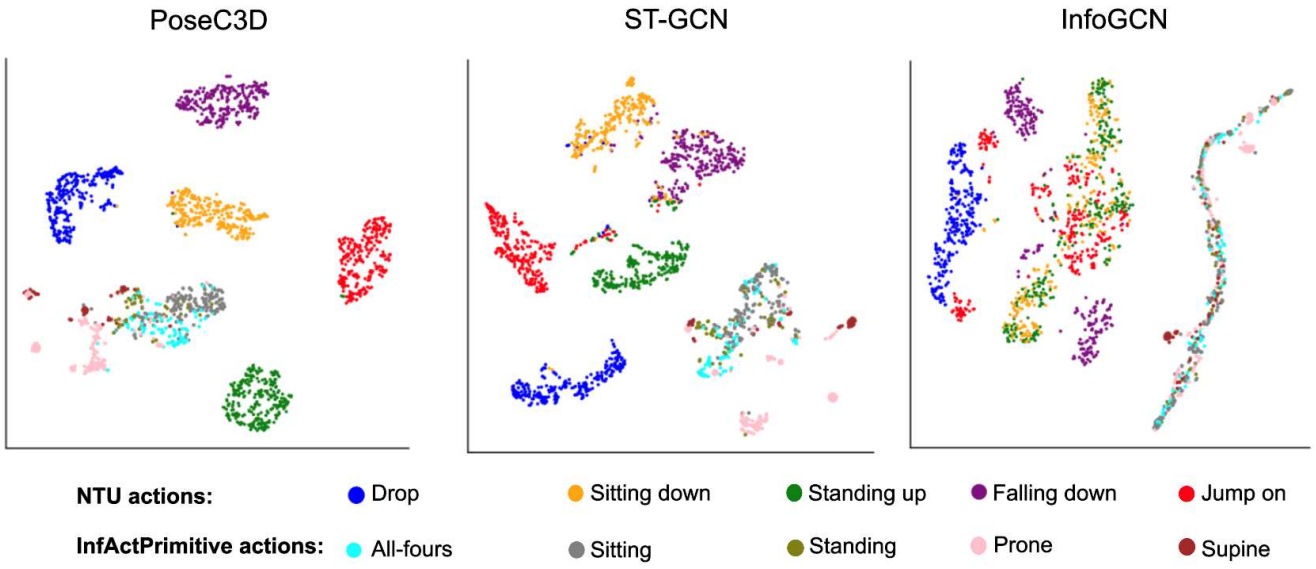

Figure 5. 2D latent projections generated through t-SNE for validation samples from both the NTU RGB+D and InfActPrimitive datasets. The results, presented from left to right, demonstrate the projection of the latent variables produced by PoseC3D, InfoGCN, and ST-GCN. While these methods effectively capture patterns in adult actions within the NTU RGB+D dataset, they struggle to distinguish between infant actions in the InfActPrimitive dataset.

Table 2. Infant action recognition results with inter-class data diversity using PoseC3D [4]. InfActPrimitive training set is partitioned into five folds, with one fold reserved for validation while the remaining folds were used to train the model. The last row of the table presents the mean and variance computed across all folds.

Held-out fold	Train	Validation	Test
Fold 1	**93.7**	83.7	**64.3**
Fold 2	87.5	**91.2**	61.2
Fold 3	**93.7**	83.0	56.3
Fold 4	**93.7**	78.7	60.8
Fold 5	**93.7**	85.0	50.6
Average	92.50± 6.2	84.3± 16.3	58.6±22.7

we applied state-of-the-art skeleton-based action recognition techniques, with Pose3D achieving reasonable performance. However, it is important to note that most other successful state-of-the-art action recognition methods failed miserably when it came to categorizing infant actions. This stark contrast underscores a significant knowledge gap between infant and adult action recognition modeling. This divergence arises from the distinct dynamics inherent in infant movements compared to those of adults, emphasizing the need for specialized, data-efficient models tailored explicitly for infant video datasets. Addressing this challenge is crucial to advancing the field of infant action recognition and ensuring that the developmental milestones of our youngest subjects are accurately tracked and understood. Our findings shed light on the unique intricacies of infant actions and pave the way for future research to bridge the gap in modeling techniques and foster a deeper understanding of infant development.

References

[1] Meng Chen, Sy-Miin Chow, Zakia Hammal, Daniel S Messinger, and Jeffrey F Cohn. A person-and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics. *Multivariate behavioral research*, 56(5):739–767, 2021. 3

[2] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 3, 5, 6

[3] 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, 2021. 3

[4] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 5, 6, 8

[5] Rubia do N Fuentefria, Rita C Silveira, and Renato S Procianoy. Motor development of preterm infants assessed by the alberta infant motor scale: systematic review article. *Jornal de pediatria*, 93:328–342, 2017. 1

[6] Zakia Hammal, Wen-Sheng Chu, Jeffrey F Cohn, Carrie Heike, and Matthew L Speltz. Automatic action unit detection in infants using convolutional neural network. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 216–221. IEEE, 2017. 3

[7] 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. 3, 4

[8] Xiaofei Huang, Shuangjun Liu, Michael Wan, Nihang Fu, Bharath Modayur, David Li Pino, and Sarah Ostadabbas. Appearance-independent pose-based posture classification in infants. In *Workshop at the International Conference on Pattern Recognition (ICPRW)*, 8 2022. 3

[9] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2023. 3, 4

[10] Xiaofei Huang, Michael Wan, Lingfei Luan, Bethany Tunik, and Sarah Ostadabbas. Computer vision to the rescue: Infant postural symmetry estimation from incongruent annotations. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1 2023. 3

[11] Alejandro Jaimes and Nicu Sebe. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134, 2007. 1

[12] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 1

[13] Chuankun Li, Yonghong Hou, Pichao Wang, and Wanqing Li. Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628, 2017. 3

[14] Chuankun Li, Pichao Wang, Shuang Wang, Yonghong Hou, and Wanqing Li. Skeleton-based action recognition using lstm and cnn. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 585–590. IEEE, 2017. 3

[15] Guiyu Liu, Jiuchao Qian, Fei Wen, Xiaoguang Zhu, Rendong Ying, and Peilin Liu. Action recognition based on 3d skeleton and rgb frame fusion. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 258–264. IEEE, 2019. 1

[16] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, 27(4):1586–1599, 2017. 3

[17] 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

[18] Shuangjun Liu, Michael Wan, Xiaofei Huang, and Sarah Ostadabbas. Heuristic weakly supervised 3d human pose estimation in novel contexts without any 3d pose ground truth. *arXiv*, 2023. 3

[19] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006. 1

[20] Itir Onal Ertugrul, Yeojin Amy Ahn, Maneesh Bilalpur, Daniel S Messinger, Matthew L Speltz, and Jeffrey F Cohn. Infant afar: Automated facial action recognition in infants. *Behavior research methods*, 55(3):1024–1035, 2023. 3

[21] Behnaz Rezaei, Yiorgos Christakis, Bryan Ho, Kevin Thomas, Kelley Erb, Sarah Ostadabbas, and Shyamal Patel. Target-specific action classification for automated assessment of human motor behavior from video. *Sensors*, 19(19):4266, 2019. 1

[22] 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, 3, 5, 6

[23] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. Suspicious human activity recognition: a review. *Artificial Intelligence Review*, 50:283–339, 2018. 1

[24] Michael Wan, Shaotong Zhu, Lingfei Luan, Gulati Prateek, 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 26th International Conference on Pattern Recognition (ICPR)*, pages 4486–4492. IEEE, 2022. 3

[25] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7464–7475, June 2023. 4

[26] 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, 3

[27] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. Action recognition based on joint trajectory maps using convolutional neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 102–106, 2016. 3

[28] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. A survey on video action recognition in sports: Datasets, methods and applications. *IEEE Transactions on Multimedia*, 2022. 1

[29] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 6

[30] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019. 1

[31] 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, 2021. 3

[32] Shaotong Zhu, Michael Wan, Elaheh Hatamimajoumerd, Cholpady Vikram Kamath, Kashish Jain, Samuel Zlota, Emma Grace, Cassandra Rowan, Matthew Goodwin, Rebecca Schwartz-Mette, Emily Zimmerman, Marie Hayes, and Sarah Ostadabbas. A video-based end-to-end pipeline for non-nutritive sucking action recognition and segmentation in young infants. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 10 2023. 3