Proc. of International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA 2026)
5-7 February 2026, Boracay Island-Philippines

An integrated framework for pose extraction and tracking in parent-child interaction research

Huu-Thiet Nguyen
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
huuthiet.nguyen@ntu.edu.sg

Isabella-Sole Bisio
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
bisioisabellasole@gmail.com

Angshuk Dutta
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
angshuk.dutta@ntu.edu.sg

Amritha Varshini Devarajan
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
amritha.vd@ntu.edu.sg

Juan-Pablo Ortega
School of Physical and Mathematical Sciences, NTU
Singapore 637371, Singapore
Juan-Pablo.Ortega@ntu.edu.sg

Domenico Campolo
School of Mechanical and Aerospace Engineering, NTU
Singapore 639798, Singapore
d.campolo@ntu.edu.sg

Su Zhang
College of Computing and Data Science, NTU
Singapore 639798, Singapore
su.zhang@ntu.edu.sg

Vishal Ramanathan
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
VISHALPA001@e.ntu.edu.sg

Thomas Allen
Department of Applied Mathematics and Theoretical Physics
University of Cambridge
Cambridge CB3 0WA, UK
taallen47@gmail.com

Cheryl Jing Yi Tan
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU Singapore 308232, Singapore
CS-CHERYL.TAN@ntu.edu.sg

Victoria Leong
Early Mental Potential and Wellbeing Research (EMPOWER) Centre, NTU
Singapore 308232, Singapore
Department of Pediatrics, University of Cambridge
Cambridge CB2 1TN, UK
victorialeong@ntu.edu.sg

***Abstract*—The study of parent-child social interactions has proven to be valuable in enhancing understanding of early child development and identifying avenues for intervention. Although rarely studied, analyses of the body pose patterns of parent-child dyads during naturalistic interactions (e.g. in videos) can yield valuable insights into their social dynamics and level of engagement. Resulting metrics such as interpersonal pose synchrony may inform our understanding of developing cognition and its disorders. However, existing methods for pose extraction and tracking lack user-friendly pipelines or fail to cater specifically to mother-child interactions, resulting in reduced effectiveness and accuracy of data extraction. This paper aims to address these limitations, by proposing an integrated methodology that offers a stable and reliable system for extracting and tracking body pose, specifically tailored for mother-child interaction videos. The proposed framework combines pose estimation and object detection techniques to extract skeletal representations of participants and track their identities throughout the video. This integration enables accurate assignment of individual identities to the corresponding skeletal data. To facilitate the adoption of this pipeline for automatic batch processing of videos, we provide the complete framework and detailed instructions in a publicly available GitHub repository. Researchers can leverage this resource to streamline their video analysis processes and extract valuable insights regarding parent-child interaction from their video data.**

***Keywords*—parent-child interaction, cognitive development, pose estimation, pose tracking.**

This research is supported by the RIE2025 Human Potential Programme Prenatal/Early Childhood Grants (H22P0M0002, H24P2M0008), administered by A*STAR.

979-8-3315-7191-7/26/$31.00 ©2026 IEEE

I. Introduction

The significant role of parent-child interactions (PCIs) in children's early developmental processes has been widely recognized [1]–[3]. Specifically, mother-child interactions (MCIs) have been shown to have a strong influence on the development of social and intellectual abilities in children [2]. Parent-child interaction is a core component in the context of a behavioral parent training program known as parent-child interaction therapy [4], [5]. Within PCIs, the examination of synchrony between parents and children contributes to understanding of cognitive development and its disorders [6]. Moreover, PCIs provide insights into the study of autism in children [7]. Through MCIs with instrumented toys [8], infant cognitive flexibility can also be assessed.

Monitoring parent-child daily interactions at home can provide valuable insights into children's development and help identify potential developmental abnormalities early on, enabling timely interventions. However, the number of available PCI datasets is limited, and most datasets involving children's interactive behaviors include activities that are unnatural or differ from typical home interactions [9], [10]. To effectively apply research techniques to large-scale environments, including home settings, it is crucial to work with datasets that capture naturalistic behaviors of children during their interactions with their mothers. Moreover, while studies on the verbal aspects of PCI are more prevalent [11], [12], multimodal datasets are able to capture richer patterns of PCI. Specifically, the dataset should include video, audio, and other modalities for a comprehensive analysis of these interactions.

In order to understand the context and evaluate various facets of the parent-child interactions, direct *manual annotation* of behaviors and events occurring in PCI videos is generally conducted [13]. However, direct annotation by human coding is typically labor-intensive and time-consuming, particularly when dealing with a large number of videos that need to be coded. Furthermore, it is known for its subjective nature, where different coders may provide varying assessments for the same behavior. Hence, automating the coding process would be extremely helpful for monitoring children's development through recorded videos, especially at home. *Quantifying movements* in videos serves as an indirect approach to aid video coding, complimenting direct observations of interacting individuals in videos, helping to reduce the time and effort especially when coding a substantial volume of videos. Motion energy analysis (MEA), also known as frame-differencing method, is a simple yet common method used in psychology for quantification of movements [14]. However, the method is sensitive to noise like changing lighting conditions, and it normally requires the interacting people's positions to be fixed during the interactions, which is not usually possible for parent-child interactions where the movements are mostly unstructured.

Pose estimation has been recently used as an alternative to MEA [15]. The method can detect the coordinates of the joints (or keypoints) of human body and create a skeleton. This method is believed to be more robust to noise. When it comes to synchrony analysis, it is also very useful for detecting which part of body is important for generating movement synchrony. Using pose estimation, however, requires correctly identifying and tracking the detected skeletons over the span of video frames to gain meaningful insights into the movements and interactions of the people over time. However, the unstructured natures of these interactions, including persistent occlusions and frequent exits or entrances, pose challenges for general pose tracking algorithms which often generate numerous IDs for each individual or exhibit frequent ID switching within short periods. To address this, we aim to develop a specialized detection model tailored for mother-child interaction scenarios.

Understanding the importance of research in parent-child interactions, especially those that are naturalistic, and the substantial potential benefits of a unified framework for pose estimation and tracking in videos capturing mother-child interactions, this paper makes the following notable contributions:

- The collection of a parent-child interaction dataset containing videos of naturalistic interactions between mothers and their children. The interactions resemble those typically performed by dyads at home, aiming to assess the daily interactive behaviors of children.
- The development of an integrated framework designed to automate the analysis of mother-child interaction videos, incorporating both pose estimation and head identification for the establishment of a robust and reliable pose tracking system. The framework is equipped with detailed instructions that aim to facilitate psychologists, social scientists and researchers without much technical expertise.
- The development of a head detector capable of discerning between the mother and child's heads. Our experimental findings suggest that fine-tuning this detector using a small selection of video frames yields promising results. This serves as a guideline for others who may wish to fine-tune the model with their own data.

II. Related Studies

A. Parent-child interaction

Research on children's interactive behaviors, especially those with autism, has been extensively studied over the years [10], [16]. This includes robot-assisted therapy, where children's interactions with robots are considered [10]. Other child datasets includes ChildPlay [17], which focuses on gaze behaviors, and Emoreact [18], which focuses on emotional responses. Rehg et al. have discussed children's social and communicative behaviors during interactions, introducing a dataset containing interactions between children aged 1-2 years old and an adult [9].

The number of datasets targeting parent-child social interaction has been limited compared to other types of human-human interactions. A recent study by Doyran et al. [19] focused on parent-infant interactions, specifically detecting physical contact between the two. Additionally, a new multimodal

dataset for dyadic parent-child interactions, named DAMI-P2C, and its preliminary analysis have been introduced [20], [21]. However, it's worth noting that the study primarily focuses on story-reading activities, which differs from our focus on daily interactive actions such as playing together with or without toys.

B. Multi-person pose estimation and tracking

Since the late 2010s and particularly in recent years, there has been the emergence of different techniques for estimating poses in scenes with multiple individuals. Many studies in the field of *multi-person pose estimation* leverage the advancements in CNN-like structures to construct their models. Some notable models for multi-person pose estimation include DeeperCut [22], SimpleBaseline [23], OpenPose [24], AlphaPose [25], MMPose [26]. The models are commonly grouped into 2 different approaches: top-down and bottom-up (or part-based). The former approach involves detecting every person in the scene and subsequently performing pose estimation for each individual. Representative models of this method include those discussed in [27], [23] and AlphaPose [25]. Conversely, the latter approach firstly detects body parts and then links them together to obtain a complete skeleton. Examples of this bottom-up approach are DeeperCut [22] and OpenPose [24]. The bottom-up approach is generally believed to be faster compared to the top-down method, particularly in scenarios involving numerous individuals within the scene.

Pose identification/tracking is another important component within video analysis. However, numerous pose estimation models, such as in [22], [24], solely provide the coordinates of skeletons without associating identities, presenting challenges in determining ownership of each skeleton across frames. In such cases, employing a tracking algorithm becomes necessary to aid in the identification process. Studies on general object tracking like Deep SORT [28] could provide a viable solution, where the Kalman filter is employed to forecast the detected object movements. Another noteworthy attempt in the domain of pose tracking is Pose Flow [29]. This approach adopts a top-down strategy, where pose estimation is conducted first, followed by the construction of a pose flow builder which facilitates the association of poses across frames. Moreover, in addition to independent tracking methodologies, some pose estimation models also attempt to incorporate pose tracking into their algorithms [23], [25].

C. Object detection and face detection

Object detection is a widely-used technology in the field of computer vision (CV), with various object detectors having been developed over the years. The detection methods fall into two primary categories: one-stage and two-stage approaches. The one-stage approach prioritizes the algorithm's speed, aiming to enable models to work efficiently in real-time situations. Examples of this approach are YOLO [30] and SSD [31]. In contrast, the two-stage approach, such as R-CNN [32], Faster R-CNN [33], Mask R-CNN [34], emphasizes detection accuracy. Among these techniques, YOLO has garnered significant interest and is extensively applied across various domains.

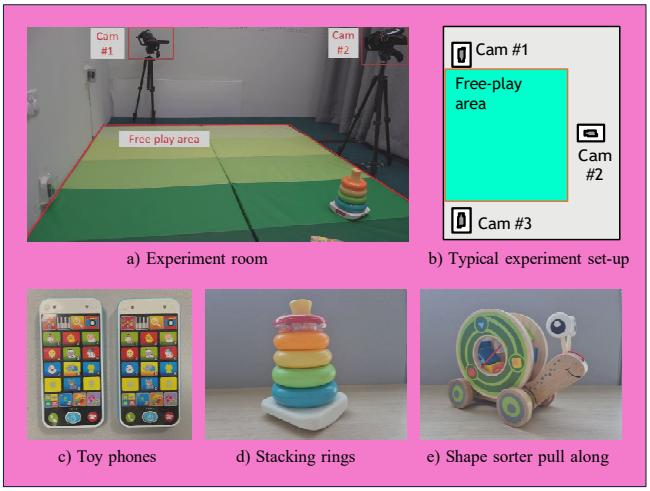


Fig. 1. Experimental setup and some materials used for parent-child interaction task. The upper row: a) experiment room with arranged floor mat and cameras (as seen from a third camera which hence does not appear in this image), b) diagram of a typical experiment set-up. The lower row (c, d, e): Some toys used in the experiment.

Face detection is commonly viewed as a particular case of object detection, enabling the identification and differentiation of individuals within each video frame. Numerous face detection techniques have emerged over several decades of research. These include methods based on active shape model [35], snakes [36], deformable templates [37], edge [38], local binary pattern [39], Gabor features [40], and neural networks [41]. However, when the interactions are unstructured, these face detectors [35]–[41] could fall short due to various factors. First, there exists a large degree of face occlusion during naturalistic mother-child interaction. The subjects often present their face with a side or back view for a considerable amount of time. The non-frontal issue is partly due to the spontaneous experiment setting which does not restrict the subject from head or body movement. A second potential issue pertains to wearable devices or clothing (e.g. EEG caps, glasses, facial masks, face coverings) which may occlude or alter some key face features, resulting in false negatives.

III. The Singapore PCI Dataset

The broad goal of the PCI experiments is to examine parent-child interactions in a naturalistic and cross-culturally appropriate manner while also supporting later data harmonization and validation. By achieving this, the methods employed for the dataset can be applied to natural settings (e.g., home-based).

A. Experiment set-up

Data collection was conducted at Nanyang Technological University in Singapore. The participants (each dyad consisting of a mother and a child) were invited to a designated

Fig. 2. Some example video frames, sampled in segments without toys (left images) and with toys (right images).(Images used with specific parental consent.)

experiment room. The room was equipped with experimental materials as shown in Figure 1, including:

- Toys: including picture books, play phone, a stacking ring (rock-a-stack), and a wooden toy (shape sorter pull along toy),
- Camera (x3): ideally, one facing the mother, one facing the child, and one for side view or top-down view,
- Soft mat: to protect the mother and child during interaction, and encourage more movement,
- Stopwatch: to monitor the time of segments and the entire session, and
- Session sheet: covering basic information of the PCI session.

The parent was instructed to play with her child as they would at home for about 10 minutes. The toys were introduced after 5 minutes of playtime.

Each PCI session consists of three segments, one of which is optional:

- Segment 1 (without toys): Parent plays and talks with their child as they would at home.
- Segment 2 (with toys): Parent and their child play together with the toys.
- Segment 3 (toy passing): Parent requests a toy from their child (optional).

Typically, three cameras recorded the mother-child interactions from different angles, resulting in three videos for each experiment. It is important to note that these videos capture the same interactions but from different viewpoints. The videos were then synchronized so that all videos of the same dyad start and end at the same time. Figure 2 shows example footage from two participant dyads for reference.

B. Dataset properties

a) Technical information

The dataset consists of multiple modalities, including images, videos, audio, and physiological signals such as EEG and ECG (not analyzed here).

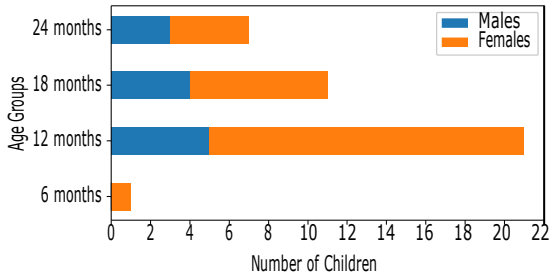


Fig. 3. Age distribution of the cohort, broken down by sex

There were 40 participant dyads. Most participants have three synchronized videos capturing the same interactions from different angles. One participant, on the other hand, has four videos. Due to technical issues, three participants only have two videos each. In total, there are 118 parent-child interaction (PCI) videos. The frame rate is typically 25 FPS, 30 FPS, or 100 FPS. Each video was approximately 10 minutes long.

b) Video contents

The videos contained mostly unstructured interactions between the mother and her child, as they normally occur at home. Throughout the videos, various actions were observed from the mothers such as cradling, holding, hugging, carrying, leading, or following their child. Mothers also elicited their child's attention through vocal cues (singing), gesture or movement.. The children could be seen moving in and out of the scene, moving towards or away from the cameras, sitting, standing, crawling, rolling over, walking, and lying down on their belly or back.

The analysis of these videos presents several practical challenges, including people entering and leaving the scene, frequent occlusions, partial body visibility or occlusion, experimenters (besides the mother and child) entering the scene, and instances where mothers and children wear EEG caps throughout the interaction. Additionally, it is common for participants to face away from or sideways to the camera, resulting in only the back or side of their heads being visible. These challenging characteristics of the dataset render existing general tracking methods and standard face detection models ineffective. This underscores the necessity of our framework as a practical tool for researchers in similar fields.

c) Demographics

The age range of the participants, along with their respective numbers, are shown in Figure 3. The mean age of the included infants was 15.55 months (SD = 4.75 months). The gender distribution within each age group is also provided in Figure 3.

IV. Methodology

The overall diagram of our proposed framework is demonstrated in Figure 4, which consists of two stages: the per-frame analysis and multi-frame processing. In the first stage of per-frame analysis, one frame of a PCI video is considered at a time, aiming to output the skeletons (if any) with their

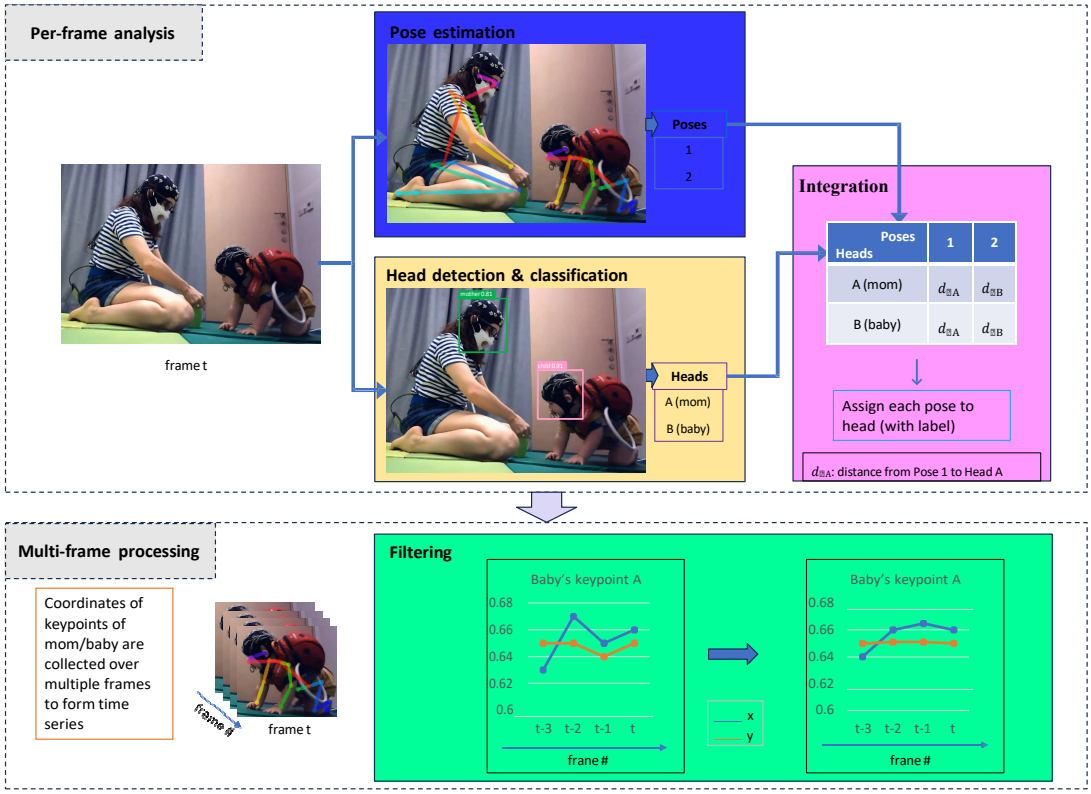


Fig. 4. The overall structure of our proposed framework. (Images used with specific parental consent.)

corresponding labels for the frame. To do that, each frame is independently fed through a pose estimation model and a head detection & classification model. The outputs are then combined in an integration mechanism to assign specific labels (mother or child) to the detected skeletons. After the per-frame analysis is completed for a batch of frames or the entire video, the second stage, multi-frame processing, begins. This stage aims to create smooth and clean time series data that is useful for further research in parent-child interaction. To that end, the keypoints of the same person's skeleton are first stacked together over consecutive frames of the video to form multiple time series, each corresponding to a keypoint. The filters are then applied to the time series to remove noise and smooth them out. For training (fine-tuning) and evaluating the models in our proposed framework, the Singapore PCI dataset is employed.

A. Pose estimation

We employ OpenPose [24], a well-known multi-person pose estimation model, to extract poses from our video data. OpenPose stands out as one of the prevalent tools for human 2D pose extraction (readers may refer to Table I which provides statistics on the popularity of various GitHub repositories for human pose estimation). Its advantages include real-time performance, high quality results and a user-friendly application programming interface (API). Most commonly-used feature of OpenPose is the estimation of the keypoints (or joints) in the main human body. Besides body keypoints, OpenPose also offers models to extract additional detailed keypoints of the foot, hand, and face. Notably, the model has also been added to OpenCV, a highly popular open-source computer vision and machine learning software library.

TABLE I. Some GitHub repositories for human pose estima-tion (as of August 2025, number of stars and forks are rounded to 0.1k).

Methods	First Commit	Stars	Forks
OpenPose[1] [24]	Apr 24, 2017	32.9k	8.0k
HRNet[2] [27]	Feb 25, 2019	4.4k	0.9k
AlphaPose[3] [25]	Dec 22, 2019	8.4k	2.0k
MMPose[4] [26]	Jul 10, 2020	6.8k	1.4k
SimpleBaseline[5] [23]	Aug 1, 2018	3.0k	0.6k

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[2] https://github.com/leoxiaobin/deep-high-resolution-net.pytorch
[3] https://github.com/MVIG-SJTU/AlphaPose
[4] https://github.com/open-mmlab/mmpose
[5] https://github.com/microsoft/human-pose-estimation.pytorch.

OpenPose offers several approaches to use its models. Users can rebuild the models or use the pretrained ones. Different configurations of the body skeleton are also provided, such as 16 points and 25 points.

B. Head detection & classification

The goal of head detection is to obtain the bounding boxes for the dyadic subjects for each frame. Employing a face detector is the natural way to this end. Indeed, face detection is the cornerstone for any face analysis systems. However, as

discussed earlier, the regular face detectors would fall short due to the unstructured nature of parent-child interactions.

We resort to a deep learning-based object detection method for head detection. Specifically, the YOLO (You Only Look Once) algorithm [30] is used for detection of the head of the dyads in each video frame. YOLO is one of the most common tools in computer vision for object detection. It has gained widespread acclaim for its exceptional speed and accuracy. Unlike traditional object detection methods, which involve multiple stages and complex post-processing, YOLO simplifies the process by treating object detection as a unified task. It essentially divides an image into a grid and simultaneously predicts bounding boxes, class probabilities, and confidence scores for each grid cell. This unique approach enables YOLO to detect multiple objects in a single pass through the neural network, resulting in real-time performance even on resource-constrained devices. We employ YOLOv7 [42] since it is the latest version as of the writing of this paper.

To proceed, the pretrained YOLOv7 model[1] is downloaded and finetuned using samples from the video recordings. Specifically, for each dyad, approximately 100 frames are obtained by sampling across all their videos at a fixed time interval based on the total duration. Each frame is then manually annotated with bounding boxes around the heads of the child and the mother. Finally, YOLOv7 is finetuned using the training script provided in the repository. Details of these steps are provided below.

1) Head annotation: This annotation task is performed for all 118 videos of 40 participants. There are approximately 100 images sampled from each video. The annotation task includes manual drawing the head bounding boxes for the mother and the child for each image, and assign to the bounding boxes either one of two labels: ”mother” or ”child”. To speed up the labeling process, we employ semi-automatic annotation. Initially, a model is trained using a small set of hand-labeled images. Subsequently, this trained model is used to make predictions on the unlabeled images. The labels generated for images that has not been manually labeled are reviewed by annotators to ensure accuracy.

2) Model training using cross-validation: After obtaining labeled data for all the sampled images, the 40 participants are randomly divided into 5 groups labeled as A, B, C, D, and E, with each group comprising 8 participants. These 5 groups then undergo a k-fold cross-validation process (k=5), where the test and validation sets are systematically rotated.

While the participants within each group and the group composition within each iteration remain consistent, there is variability in the number of images per video that are used for fine-tuning YOLOv7. From approximately 100 labeled images for each video, we use 1/5, 1/4, 1/3, 1/2, 1/1 of those images (which corresponds to approximately 20, 25, 33, 50, 100 images per video, respectively) to fine-tune our YOLOv7 head detector. This approach allows us to analyze the performance trend as we increase the number of data samples for training and serves as an indicator for determining the number of frames that need to be labeled to achieve satisfactory performance.

3) Head bounding box generation: After training, the best model obtained in each iteration, based on evaluation on the validation set, is used to generate head bounding boxes for the videos. The data from head bounding boxes is stored in a CSV file per video, with each line corresponding to a frame. Each line of the CSV file only contains at most 1 bounding box for the mother and at most 1 bounding box for the child. These bounding boxes are those with the highest likelihood of being the mother and the child, respectively.

C. Integration of the pose estimation and head detection for tracking purposes

The head keypoints (including the eyes, ears and nose) obtained from pose estimation (OpenPose) and the head bounding boxes resulted from head detection (YOLO) are used as the inputs of the integration step, as shown in Figure 5.

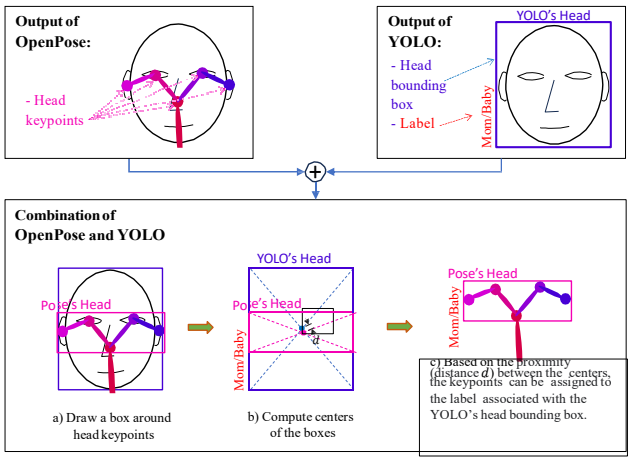


Fig. 5. Integration of the outputs of OpenPose and YOLO, to assign the skeleton to a label.

To associate the OpenPose-detected head keypoints with YOLO-detected head bounding boxes, we create bounding boxes around the OpenPose head keypoints, as illustrated in Figure 5a. Then, we compute the coordinates of the centers for both the OpenPose-detected and YOLO-detected bounding boxes (called Pose's head and YOLO's head respectively), as shown in Figure 5b. Generally in a frame, there are multiple Pose's heads and YOLO's heads. To match their centers, we use the Euclidean distance d (measured in pixels) between them (as shown in Figure 5c) and utilize the linear sum assignment problem, as depicted in Table II.

The linear sum assignment problem seeks to match the Pose's heads to the corresponding YOLO's heads by minimizing the total distance (sum of d_{ij}) between their centers. This matching allows the skeletal data to be accurately linked to the corresponding individual identities.

[1]https://github.com/WongKinYiu/yolov7

TABLE II. OpenPose–YOLO heads assignment task. Suppose that there are 3 head bounding boxes (A, B, C) detected by YOLO and 2 skeletons (1, 2) with available head keypoints detected by OpenPose in a frame.

Pose's Head / YOLO's Head	A	B	C
1	d_{1A}	d_{1B}	d_{1C}
2	d_{2A}	d_{2B}	d_{2C}

D. Filtering for movement quantification

After the integration of the OpenPose and YOLO for each video frame, the keypoints for the mother and the keypoints for the child are stored independently from one another. Collecting the data accross the frames of the video generate the time series data for each keypoint of each person (mother and child). In the following, we present some additional techniques aimed at reducing noise within the acquired time series data. Although these methods are implemented in our codebase, they are not applied in the analyses reported in Section V and therefore do not influence the presented results.

1) Raw filtering: For the elimination of extreme values in each time series, such as false detections, we implement a moving average (MA) technique. The MA serves as a simple low-pass filter, and is normally used to smooth time series data. In our framework, the raw filter is exclusively employed to eliminate abnormal data points. By utilizing the moving average, outliers that significantly deviate from the current MA value by a specified threshold are effectively removed.

2) Fine filtering: In our effort to further reduce noise in the time series, the data obtained from the raw filter is subjected to an additional filtering process aimed at fine-tuning and smoothing the series. For this purpose, we employ a digital filter known as the SavGol (Savitzky–Golay) filter [43].

Using the fine filter yields time series data that is notably less noisy and smoother. This refined data is then better suited for a variety of research purposes, including analyses related to synchrony and other research applications.

V. Tracking Results

This section presents the results of applying our proposed framework to the Singapore PCI dataset. It includes details on the evaluation metrics used and the performance of the models employed.

The source code is open and available on GitHub at https://github.com/thiethnguyen/MCI-pose-tracking.

A. Evaluation metrics

In our research, we employed OpenPose to generate the pose and then integrated it with our fine-tuned head detection model to determine whether the generated pose corresponds to the mother, the child, or neither. Consequently, the evaluation metrics should access how effectively the model classifies the pose into the 3 classes (mother, child, other). Therefore, we employed metrics commonly associated with classification tasks, such as accuracy and F1-score, to evaluate the performance of our proposed framework.

B. Performance of the proposed framework

As mentioned in Section IV-B2, we trained five different models to analyze the performance trend as we increase the number of data samples. The five models are named according to the ratio of labeled images in each video used for fine-tuning: 1/1, 1/2, 1/3, 1/4, and 1/5. For each model, we collected results from each data fold, and subsequently computed the averages to obtain the final figures for training, validation, and test sets.

Table III summarizes the average accuracies for different scenarios involving the utilization of varying data ratios for fine-tuning YOLOv7, while figures 6 and 7 present the F1-scores specifically for the mother and the child, respectively.

TABLE III. Average results for different models fine-tuned by different ratios of data.

Ratio of data *	Accuracy		
	Training	Validation	Test
1/5	0.9 ± 0.034	0.867 ± 0.037	0.867 ± 0.062
1/4	0.936 ± 0.02	0.908 ± 0.027	0.903 ± 0.034
1/3	0.937 ± 0.012	0.915 ± 0.022	0.913 ± 0.025
1/2	0.953 ± 0.012	0.936 ± 0.021	0.929 ± 0.016
1/1	0.965 ± 0.005	0.941 ± 0.014	0.944 ± 0.015

* The ratio over the number of images that have been labeled.

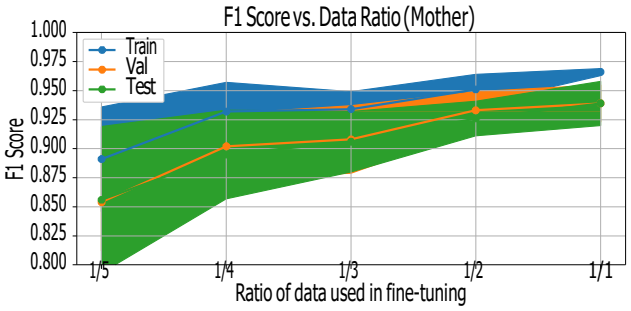


Fig. 6. Average results for different models fine-tuned by different ratios of data - Mother.

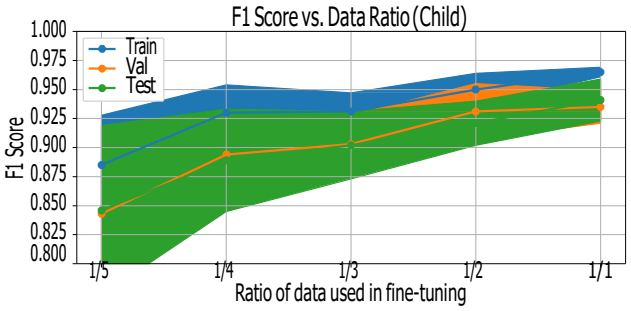


Fig. 7. Average results for different models fine-tuned by different ratios of data - Child.

Figures 6 and 7 show the trend of the mean F1-score as different portions of the labeled images are used for fine-tuning YOLOv7. Notably, the results show a general increase

in the averaged F1-score as the number of labeled images used for head detection model fine-tuning increases—a predictable trend. Nevertheless, the scores remain high across all data ratios. With only one-fifth of the labeled data (20 images per video), the F1-score on the test sets reaches approximately 0.85, while using the full dataset (100 images per video) raises it by only about 0.09, to around 0.94.

VI. Using the Proposed Framework

In this section, we apply the proposed framework to the task of calculating the total movement of the child in each video. This serves as a tutorial for anyone who wants to use our framework for their purposes. The application also demonstrates that the proposed framework effectively distinguishes between the skeletons of the mother, child, and others, and generates time series data for the keypoints.

A. Defining the task

Using the time series of a person's keypoint that has been filtered to remove noise, (i.e., an output of our proposed framework) the total movement of that keypoint throughout the video is calculated based on the summation of all the Euclidean distances the keypoint moves from a previous frame *(t − 1)* to the next frame *(t)*.

$$\mathrm{TM} = \sum_{t=2}^{N} \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2} \tag{1}$$

where TM stands for total movement, t denotes frame index, N denotes the total number of frames of the video. As the lengths of the videos can be different, a normalized total movement (NTM) is introduced as

$$\mathrm{NTM} = \frac{\mathrm{TM}}{N} \times N_0 \tag{2}$$

where the factor N_0 is the standardized number of frames of a video.

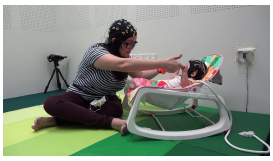
B. Adoption of the framework

The fine-tuned YOLO model was used to classify the extracted skeletons in each frame of the PCI video. After that, the filtering techniques mentioned in Section IV-D were applied to preprocess the time series for each keypoint.

With our current task of movement calculation, we obtained time series data of children's neck movements from all the videos. Figure 9 presents selected time series data from two children in different videos (with corresponding snapshots shown in Figure 8), following completion of the preprocessing process.

The total movements were then calculated using equations 1 and 2 with $N_0 = 15000$ being used, as the standard video is defined as 10 minutes in length and at a frame rate of 25 FPS.

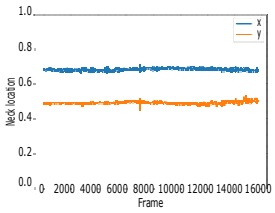
As each participant has several videos, we picked the highest value of NTM among those videos as the number for that participant. We then calculated the mean and SD for the participants in the same age group. The results are shown in Table IV.



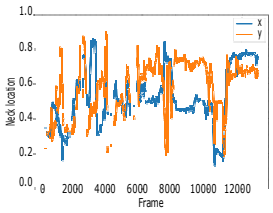
(a) Interaction of a 6 month-old child (left)

(b) Interaction of a 12 month-old child (right)

Fig. 8. Interaction of a 6 month-old child (left) and a 12 month-old child (right). (Images used with specific parental consent.)



(a) Neck location of a 6 month-old child (left)



(b) Neck location of a 12 month-old child (right)

Fig. 9. Neck location of a 6 month-old child (left) and a 12 month-old child (right).

C. Discussion on the result of the task

As can be seen from Table IV, there is general increasing trend in the movement of the children when their ages are older. A distinct difference between the child of age 6 months with other children of older age groups can also be seen, where the normalized total movement of the 6 month-old child is much less than averages of other age groups. This is reasonable as the posture of the 6-month old infant is generally supine throughout the experiment, as compared with other older children who are able to stand, sit, run, walk during the experiment. It is also worth noting that even though the child is generally supine, the interactions with the mother have contributed to the value of the total movement, also not to mention the error of the OpenPose or the incapability of the filters to completely remove the noise of the time series data.

Therefore, the results in Table IV, while not perfectly capturing the total neck movement of the child, can still indicate the efficacy of our proposed framework. The steps outlined in this Section VI also provide a solid starting point for users interested in applying this method.

VII. Further Discussions and Limitations

When using general tracking algorithms, it is common to obtain a substantial number of tracking IDs for each individual, particularly in scenarios where human subjects exhibit extensive movement, interaction, and frequent entries and exits from the scene. An additional challenge arises from the fact that dyad members may share the same set of IDs. This can pose significant difficulties for coders and researchers, as it requires

TABLE IV. Average normalized total movement (NTM) of the neck for different infant age groups.

Age group (months)[1]	Average NTM of the neck
6	14.50 ± 0
12	29.06 ± 8.56
18	31.24 ± 10.48
24	31.82 ± 6.39

[1] The infants are categorized to suitable age group based on the closest to the actual age.

careful observation of the output videos to accurately attribute the correct ID to the right individual at various points in time. This discussion highlights the complexities of tracking in scenarios with dynamic human interactions, emphasizing the need for more specific tracking methods to improve precision and reduce the burden on researchers.

The results presented in Section V-B indicate that it may not be necessary to use a large amount of sampled data for fine-tuning the head detection model to attain satisfactory results. In practical scenarios where time constraints exist, the option of reducing the number of sampled images requiring labeling becomes a viable consideration. This approach can help streamline the data labeling process without significantly compromising the quality of the head detection model.

Our work presents a straightforward unified framework for dyadic pose extraction, suitable for researchers in psychology and child development studies who wish to use machine learning methods to analyze pose patterns in PCI videos. It also serves as a tutorial for non-experts outside the machine learning community, guiding them in selecting, implementing, and integrating various ML tools suited to their research needs. Our contribution primarily benefits a specific group of researchers in these fields, rather than making significant technical contributions to the broader machine learning or video analysis communities. It is also important to note that the dataset used in our study is modest in size and is still under development.

VIII. Conclusion and Future Work

In this paper, we have introduced an integrated methodology that provides a robust and reliable system for extracting and tracking body poses, with a specific focus on mother-child interaction videos. Through the integration of OpenPose as a pose estimator and YOLOv7 as a head detector, we have been able to assign identifications to the skeletons detected. Our experimental findings have shown the efficacy of this framework, even when fine-tuning the head detection model with a relatively small subset of video frames. This suggests that the proposed methodology offers a practical and efficient solution for researchers working with mother-child interaction videos. The code and models shared within this research serve as a valuable contribution to the field, offering a useful tool that can be applied to various studies in the domain of child development and health.

Our current and future directions include making the entire framework even easier for non-expert users, increasing the accuracy of the models for different cross-cultural interaction videos, and investigating the postural synchrony and coordination between dyad members.

Acknowledgment

This research is supported by the RIE2025 Human Potential Programme Prenatal/Early Childhood Grants (H22P0M0002, H24P2M0008), administered by A*STAR. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the A*STAR.

ChatGPT was used in parts of the manuscript for language editing and grammar refinement.

References

[1] M. L. Hoffman, "Moral internalization, parental power, and the nature of parent-child interaction." *Developmental psychology*, vol. 11, no. 2, p. 228, 1975.

[2] A. P. STREISSGUTH and H. L. Bee, "Mother-child interactions and cognitive development in children," *Young Children*, pp. 154–173, 1972.

[3] C. Konijnenberg, M. Sarfi, and A. Melinder, "Mother-child interaction and cognitive development in children prenatally exposed to methadone or buprenorphine," *Early human development*, vol. 101, pp. 91–97, 2016.

[4] C. B. McNeil, T. L. Hembree-Kigin, and K. Anhalt, "Parent-child interaction therapy," 2010.

[5] R. Thomas, B. Abell, H. J. Webb, E. Avdagic, and M. J. Zimmer-Gembeck, "Parent-child interaction therapy: A meta-analysis," *Pediatrics*, vol. 140, no. 3, 2017.

[6] C. Leclère, S. Viaux, M. Avril, C. Achard, M. Chetouani, S. Missonnier, and D. Cohen, "Why synchrony matters during mother-child interactions: a systematic review," *PloS one*, vol. 9, no. 12, p. e113571, 2014.

[7] A. Loncarevic, M. T. Maybery, J. Barbaro, C. Dissanayake, J. Green, K. Hudry, T. Iacono, V. Slonims, K. J. Varcin, M. W. Wan *et al.*, "Parent-child interactions may help to explain relations between parent characteristics and clinically observed child autistic behaviours," *Journal of Autism and Developmental Disorders*, pp. 1–15, 2023.

[8] V. Ramanathan, M. Z. Ariffin, G. D. Goh, G. L. Goh, M. A. Rikat, X. X. Tan, W. Y. Yeong, J.-P. Ortega, V. Leong, and D. Campolo, "The design and development of instrumented toys for the assessment of infant cognitive flexibility," *Sensors*, vol. 23, no. 5, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2709

[9] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3414–3421.

[10] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, "3d human sensing, action and emotion recognition in robot assisted therapy of children with autism," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2158–2167.

[11] B. Huber, R. F. Davis III, A. Cotter, E. Junkin, M. Yard, S. Shieber, E. Brestan-Knight, and K. Z. Gajos, "Specialtime: Automatically detecting dialogue acts from speech to support parent-child interaction therapy," in *Proceedings of the 13th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2019, pp. 139–148.

[12] A. Bird, E. Reese, K. Salmon, K. Waldie, E. Peterson, P. Atatoa-Carr, and S. Morton, "Maternal depressive symptoms and child language development: Exploring potential pathways through observed and self-reported mother-child verbal interactions," *Development and psychopathology*, vol. 36, no. 4, pp. 1959–1972, 2024.

[13] H. Aspland and F. Gardner, "Observational measures of parent-child interaction: an introductory review." *Child and Adolescent Mental Health*, 2003.

[14] F. T. Ramseyer, "Motion energy analysis (mea): A primer on the assessment of motion from video." *Journal of counseling psychology*, vol. 67, no. 4, p. 536, 2020.

[15] K. Fujiwara and K. Yokomitsu, "Video-based tracking approach for nonverbal synchrony: a comparison of motion energy analysis and openpose," *Behavior Research Methods*, vol. 53, pp. 2700–2711, 2021.

[16] A. Baird, S. Amiriparian, N. Cummins, A. M. Alcorn, A. Batliner, S. Pugachevskiy, M. Freitag, M. Gerczuk, and B. Schuller, "Automatic classification of autistic child vocalisations: A novel database and results," 2017.

[17] S. Tafasca, A. Gupta, and J.-M. Odobez, "Childplay: A new benchmark for understanding children's gaze behaviour," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 935–20 946.

[18] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "Emoreact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*, 2016, pp. 137–144.

[19] M. Doyran, R. Poppe, and A. A. Salah, "Embracing contact: Detecting parent-infant interactions," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 198–206.

[20] S. Alghowinem, H. Chen, C. Breazeal, and H. W. Park, "Body gesture and head movement analyses in dyadic parent-child interaction as indicators of relationship," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021, pp. 01–05.

[21] H. Chen, S. Alghowinem, S. J. Jang, C. Breazeal, and H. W. Park, "Dyadic affect in parent-child multimodal interaction: Introducing the dami-p2c dataset and its preliminary analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3345–3361, 2022.

[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14*. Springer, 2016, pp. 34–50.

[23] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466–481.

[24] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[25] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[26] M. Contributors, "Openmmlab pose estimation toolbox and benchmark," https://github.com/open-mmlab/mmpose, 2020.

[27] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[28] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.

[29] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu, "Pose flow: Efficient online pose tracking," *arXiv preprint arXiv:1802.00977*, 2018.

[30] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[31] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 21–37.

[32] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[33] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[35] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[36] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[37] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *International journal of computer vision*, vol. 8, pp. 99–111, 1992.

[38] S. Anila and N. Devarajan, "Simple and fast face detection system based on edges," *International Journal of Universal Computer Sciences*, vol. 1, no. 2, pp. 54–58, 2010.

[39] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[40] M. Sharif, A. Khalid, M. Raza, and S. Mohsin, "Face recognition using gabor filters." *Journal of Applied Computer Science & Mathematics*, no. 11, 2011.

[41] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE signal processing letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[42] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. 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*, 2023, pp. 7464–7475.

[43] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.