Decoding Contact: Automatic Estimation of Contact Signatures in Parent-Infant Free Play Interactions

Metehan Doyran
m.doyran@uu.nl
Utrecht University
Utrecht, NL

Albert Ali Salah
a.a.salah@uu.nl
Utrecht University
Utrecht, NL

Ronald Poppe
r.w.poppe@uu.nl
Utrecht University
Utrecht, NL

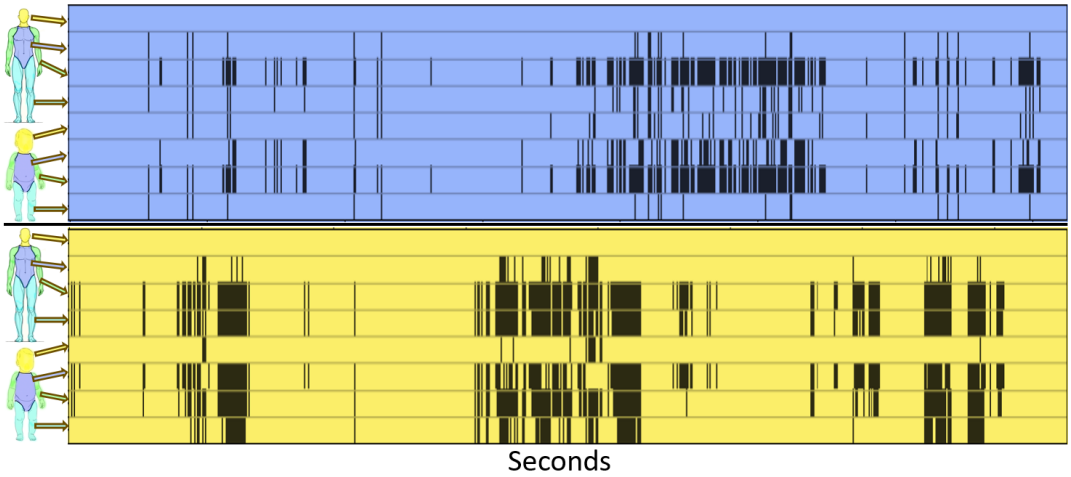

Figure 1: Visualization of our course contact segmentation model's predictions on two test videos (~13 mins). Top (blue) depicts an interaction characterized by brief moments of hand touch and a longer period of the child sitting on the parent's lap. Bottom (yellow) has segments of parent's arm touch to the child's body, suggesting a restrictive dynamic between parent and infant.

ABSTRACT

In parent-child interactions (PCIs), there is frequent physical contact between the two actors. Quantifying this contact provides valuable input to assess the nature of the interaction or the relation between parent and child. Here, we explore the application of vision-based techniques to automatically detect contact signatures at each frame of video recordings of playful parent-infant interactions. We employ two separate models: (i) a multimodal convolutional neural network (CNN) that integrates 2D pose and body part information, and (ii) a unimodal graph convolutional neural network (GCN) that utilizes only 2D pose. We showcase the potential and limitations of automatic contact signature estimation through quantitative and qualitative assessments using a parent-infant free play interaction dataset consisting of 100 parent-child dyadic interactions, covering 20 hours. Additionally, our experiments provide insights into various design choices through systematic experimentation. By releasing our annotations and code, we aim to enable further research in the automatic contact signature estimation during free play interactions between parents and infants[1].

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**; • **Information systems** → *Video search*; • **Applied computing** → Psychology.

KEYWORDS

Parent-Child Interaction; Interaction Analysis; Contact Detection; Pose Estimation; Free Play; Graph Convolutional Neural Network

ACM Reference Format:
Metehan Doyran, Albert Ali Salah, and Ronald Poppe. 2024. Decoding Contact: Automatic Estimation of Contact Signatures in Parent-Infant Free Play Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3678957.3685719.

1 INTRODUCTION

Unraveling the dynamics in parent-child interactions (PCI) gives insight into the processes that affect a child's development. Interactions with a parent reveal characteristics of early child development such as language acquisition, cognitive growth, and socio-emotional development [30, 31]. Videos of PCIs are used to assess and track development of children [32, 36]. However, manually annotating

ICMI '24, November 4–8, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0462-8/24/11
https://doi.org/10.1145/3678957.3685719.

[1]This is the uncorrected Author Proof including the supplementary.

these videos, as done traditionally, is both labor-intensive and requires trained professionals [13, 29]. Computational advancements enable automated approaches to extract significant features of interaction from much larger amounts of data and to analyze them more objectively [18].

Physical contact plays a crucial role in the communication of emotion as an essential component of nonverbal communication [17]. In PCIs, physical contact facilitates emotional bonding and regulation [1, 2]. Traditional assessments of contact through self-report questionnaires are good indicators but they may include bias toward representing specific activities and particular types of contact [3]. Moreover, these instruments do not provide spatially detailed insights, nor do they allow for temporal analysis as an interaction unfolds. To overcome these limitations, we propose an automated system for estimating *contact signatures*, encoding which body parts are in contact, during parent-infant free play.

Contact detection, an under-researched topic, is challenging. Interactions with contact include body parts of the participants occluding each other. Fieraru *et al.* [11] introduced two contact detection sub-tasks to improve 3D reconstruction for humans in contact: *contact segmentation*, and *contact signature estimation*. Contact segmentation provides a set of body part regions that are in contact, coded at the individual. Contact signature estimation increases the granularity of contact segmentation by specifying which regions of one person are in contact with which regions of the other. Different from [11], we focus exclusively on parent-infant playful interactions in this paper, with a larger variety of body parts in contact on average compared to adult-adult daily interactions.

During free play, a parent and a child can freely select which toys to use, and engage in a much broader variety of interactions compared to a structured setting, where a specific task is being performed. Subsequently, different behaviors may be elicited that can provide cues about the quality of interaction, prompting and directive behaviors of the parents, attention seeking and sharing, affective displays, and other behaviors that can ultimately provide information about the child's psycho-social outcomes [21].

In this paper, we investigate the feasibility of a vision-based method to automatically provide frame-level contact signatures in parent-child interactions. We compare and combine two models: (i) a multimodal convolutional neural network (CNN), adapted from [9], with two input modalities, namely 2D pose and body parts; and (ii) a unimodal graph convolutional neural network (GCN), with only 2D pose as its input. Our main contributions are as follows:

- We analyze physical contact in parent-infant interactions with contact signatures for the first time.
- We show the potential and limitations of automatic contact signature estimation, using systematic experiments to understand the contribution of the implementation choices.
- Our quantitative and qualitative results emphasize the importance of analyzing physical contact between parent and child to provide a good understanding of their interaction.
- We provide contact signature annotations for the publicly available YOUth PCI dataset [25] to encourage research on automatic contact assessment in parent-infant interaction during free play. We release our code base and models[2][3] as well as the annotation tool[4].

The remainder of this paper is structured as follows. We first discuss related work on parent-child interaction analysis, with a focus on physical contact. In Section 3, we summarize the YOUth PCI dataset and detail our contact signature annotations. Our methodology, including the implementation details of the computational models, appears in Section 4. We quantitatively and qualitatively evaluate our approach and discuss the results in Section 5, and conclude in Section 6.

2 RELATED WORK

Our work bridges several research areas. In this section we categorize and discuss the relevant literature in these areas.

Contact in Parent-Child Interactions. The study of parent-infant interaction is widely explored within the realms of psychology, child development, and attachment theory [1, 2]. The prominence of inter-personal touch is widely acknowledged as essential for a child's developmental stages [28], providing a foundational basis for other forms of communication that develop later [17]. Touch not only affects physiological conditions, but also promotes healthy biological growth and plays a crucial role in social development [24].

Social Touch in Parent-Infant Interactions. Touch plays a significant role in establishing and reinforcing social bonds, especially in the context of parent-infant interactions. In a notable study, researchers mapped relationship-specific social touch allowance [35]. Participants were asked to mark the body areas where different individuals could touch them. These maps demonstrated that children were mostly allowed to touch the head, upper back, shoulders, and arms of their parents.

Contact Classification and Signature Estimation. The role of touch in parent-child interactions is critical, yet research on physical contact detection using computer vision is limited, with notable exceptions such as the works of Chen et al. [6, 7]. In both studies, they utilized computer vision methods to detect interactions involving touch between a parent's hand and a child's body. The data are collected in a controlled environment where both participants are seated opposite each other. This means that the hands are the primary means to contact, whereas in a free play setting, touch can happen in much greater variety. In [6], the proposed approach achieved a precision rate of approximately 48%, with a high recall rate (99%) in such settings. They also classified the touch locations on the infant's body. Our research extends this approach by analyzing touch events in more complex free play scenarios, without restricting it to hand-to-body contacts and, thus, we do not rely exclusively on hand segmentation.

Fieraru *et al.* [11] introduced an innovative method for detecting and analyzing physical contact from visual data, filling a significant void in existing literature. Although their main goal was 3D reconstruction, their auxiliary tasks to estimate physical contact regions between adults across various scenarios laid the foundations of

[2]https://github.com/dmetehan/Image2Contact
[3]https://github.com/dmetehan/Pose2Contact
[4]https://github.com/dmetehan/HumanContactAnnotator

contact signature estimation, which concerns the identification of specific regions of contact on the bodies involved in the interaction.

2D Pose Estimation. In our study, the advancement of pose detection technologies plays a crucial role in enhancing contact signature estimation. Various approaches like OpenPose [4], AlphaPose [10], PoseNet [19], DarkPose [39], and HRNet [34] have been used in applications ranging from human activity recognition to human-object interaction analysis. Since infant bodies have different shapes and proportions, models trained with adult bodies may not perform well on infants. DarkPose in particular is known to perform better with infant poses [9]. In this work, we use HRNet integrated with DarkPose to extract pose-related features for contact signature estimation.

Body Part Segmentation. DeepLab [5] and U-Net [27] are among the leading models for body part segmentation, the task of extracting the areas of each body part in an image. Lin *et al.* [23] introduced a cross-domain adaptation approach, offering potential improvements in segmentation tasks. More recently, the introduction of the Segment Anything Model (SAM) [20], trained on an extensive dataset with 11 million images and a billion segmentation masks, highlighted the ongoing improvements in visual segmentation. In our study, segmentation is the key technology for performing a more detailed analysis by distinguishing various body parts of the interacting people.

Human-Human Interactions. Several studies have focused on analyzing and predicting human-human interactions [14, 15, 26, 37]. These works explored different aspects of pose forecasting, pose refinement, and instance segmentation in physically close human-human interactions. In addition to strong deep neural network backbones, they used attention mechanisms, different initialization procedures, and network architectures to improve the performance and stability of interaction analysis (see [33] for a review). Research on contact detection might benefit from human-human interaction classification, and *vice versa.*

3 PARENT-INFANT INTERACTION DATASET

The YOUth Cohort [25] is a large-scale, longitudinal cohort following nearly 4,000 Dutch children from pregnancy until early adulthood. The study focuses on neurocognitive development involved in two core characteristics of behavioural development: social competence and behavioural control, respectively. In our experiments, we used parent-child interaction (PCI) videos[5] for 10 to 12 month-old children (average 11.4 months old with a standard deviation of 1.2 months) where they play freely with one of their parents. We use the terms *child* and *infant* interchangeably.

Videos are captured in a room with a play area of roughly five square meters. In each recording, a parent and child play freely on the floor. A variety of toys (car, doll, switch box, flower, book, baby bottle, shape box with four different colored shapes) are available to play with. These play sessions include a range of (joint) behaviors, including parent and infant playing together, infant playing and parent observing, infant crawling away from the play area and parent restricting or bringing them back to the play area, as well as parent guiding the infant to play with a specific toy.

[5]Access to these videos is granted to researchers following an ethics screening process.

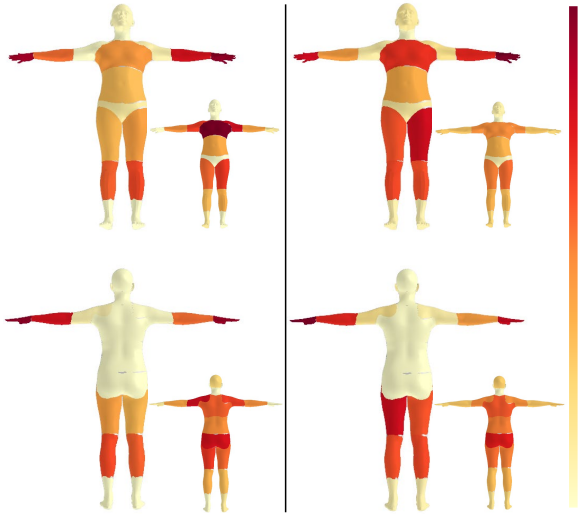

Figure 2: Contact segmentation heatmaps for the predictions of the model (left) vs. the actual annotations (right). For each visualization the parent is on the left and infant is on the right, showing both front (top) and back (bottom). Colors indicate the amount of time in which a body part was involved in contact, aggregated over all predicted and annotated interactions.

Videos are captured with four uncalibrated dynamic cameras, positioned close to ground level, as both the parent and the child are predominantly seated on the ground. We used all views to annotate the data, but limit our automatic analysis to a single view to address a more practical setting. Since manual contact signature annotation is time-consuming, we randomly selected 100 parent-child interaction videos from the available pool of more than 1,500 videos. These videos were trimmed to focus on portions of the interaction, where only the parent and the child are in the frame. The trimmed videos have an average duration of 12 : 34 minutes (standard deviation of 32 seconds). This selection covers around 20 hours of free play between parents and their infants.

Each 5 seconds, frames from the selected videos were extracted and annotated for contact/no contact using a single view, as in [9]. In specific cases, the distinction could not be made confidently, for example because the infant was not visible due to occlusion by the parent or left the recording area. Resulting ambiguous frames were discarded. Annotating from a single view serves a specific purpose when training contact detection models. Using multiple views may provide a more accurate label for the actual contact, but if the contact is invisible (fully occluded) from a particular viewpoint, marking it as a contact will confuse the model during learning. We would be forcing the model to predict something that has no visible indication in the acquired image.

3.1 Contact Segmentation Annotations

The word *segmentation* here does not refer to the classical image segmentation task and instead we follow the definition of Fieraru *et al.* [11] where Contact segmentation refers to segmenting the body into a number of regions, and indicating the contact between

two people with binary annotations per individual per body region. Diverging from their work, instead of 75 regions per person, we combine some neighboring regions, resulting in 21 regions per person. We argue that a more fine-grained analysis is unlikely to significantly improve the value of the annotations. We refer to our contact segmentation annotations with 21 regions as "21 + 21". We also derive a lower-resolution version of these annotations by combining neighboring regions into 6 regions per person, referred to as "6 + 6". We report our findings of 21 + 21 for fine-grained, and 6 + 6 for coarse-grained estimation.

Figure 2 illustrates the 21+21 contact segmentation heatmaps for parents and infants for both the best model predictions (left) and for the manual annotations (right). Comparing our manual annotation heatmaps with those from Fieraru *et al.* (obtained predominantly from adult-adult interactions), we observe significant differences. In their work, the hands, arms, shoulders and upper back are the regions with the highest contact frequency. In contrast, our data clearly show that, while parents most often use their hands during interactions, their shoulders and back rarely come into contact with the child. Parents' legs and chest are among the regions with the highest contact frequency. This aligns with the intuitive observation that infants' heatmaps show the highest contact occurrence at the buttocks, as they often sit on their parents' laps.

Previous research by Suvilehto *et al.* [35] highlighted the unrestricted nature of parent-child touches, and showed that it allowed contact in areas that would be considered off-limits in adult-adult interactions across relationships. Our study corroborates these findings and, additionally, quantifies the frequency of such touch interactions during free play. Contrary to Fieraru *et al.*, who reported minimal contact in the lower regions between adults, our contact segmentation heatmaps show the highest contact frequency in these areas for infants. This increased range of physical contact highlights the unique nature of parent-infant touch, which is not restricted by the social constraints typical for adult interactions.

The parent contact segmentation heatmaps in Figure 2 show asymmetry, with the more contact at the left side. It can be a bias in the dataset. Parents might have been predominantly right-handed and might have interacted with toys more with the right hand, leaving the left hand to provide support for their children. It can also be the case that the room setup makes it easier to reach the toys from the right side with respect to the parents and keep the children on the left side. Further analysis should look into this bias.

3.2 Contact Signature Annotations

Similar to contact segmentation annotations, contact signature annotations also encode the contact between interacting people. A contact signature represents the regions in contact with a set of tuples, where each tuple denotes the contact between a body part for person 1 and a body part for person 2. Our annotators solely annotated contact signatures for 21 regions per interactant. We denote these annotations as "21 × 21", since for each of the 21 regions in person 1 (*i.e.*, the parent), a contact/no contact value is chosen for each of the 21 regions in person 2 (*i.e.*, the child). A lower resolution version of these annotations (6 × 6) was derived using the same procedure as the contact segmentation annotations. It is important to note that it is trivial to derive contact segmentation annotations from the contact signature annotations.

Since the contact signature annotations require a higher level of understanding about the three dimensional nature of the scene, annotators used all four views for the annotations to get a better ground truth annotation quality. All the frames were annotated by a single annotator and a second annotator annotated a subset of the frames to determine the quality of the annotations.

Let A and B be annotations of two different annotators for a single frame. Then, the Jaccard score J is calculated as the number of elements in the intersection of the positive labels from both A and B divided by the number of elements in the union of all the positive annotations. The Jaccard score is also known as intersection over union, and is given in Eq. 1:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

A Jaccard score of 42.46% was calculated for the inter-annotator agreement for the 21 × 21 contact signature annotations. We have also calculated the Cohen's Kappa, which was 0.47.

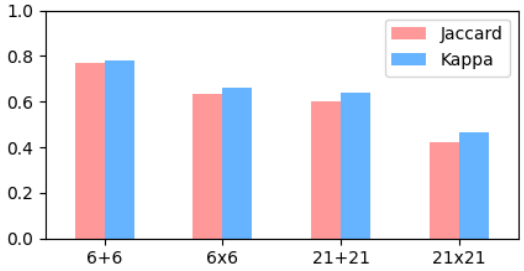

Figure 3: Inter-annotator agreement for contact signature (6× 6, 21×21) and contact segmentation (6+6, 21+21) annotations.

The inter-annotator agreement for the derived contact segmentation annotations, 21 + 21 and 6 + 6, as well as the derived lower resolution contact signature annotations, 6 × 6, are shown in Figure 3. Naturally, the easier the task, the higher the inter-annotator agreement. To give an illustrative example, suppose the first annotator annotates a contact signature as [parent left fore-arm - infant left upper leg], and the second annotator annotates it as [parent left upper-arm - infant left upper leg]. For the 21 × 21 contact signature annotations, this counts as a disagreement, even though the difference between the fore-arm and upper-arm is small. However, if we consider the (derived) lower resolution 6 × 6 contact signature annotations, where the fore-arm and the upper arm are combined into one region (called 'arm'), these annotations count as an agreement (*i.e.*, [parent left arm - infant left leg]).

Additional analysis and visualizations of the contact signature annotations appear in the supplementary material.

4 METHODOLOGY

To explore the potential of automatic contact signature estimation, we explore two different architectures: a CNN-based contact signature estimator called Image2Contact and a GCN-based estimator with 2D joint locations as input, called Pose2Contact. We introduce both architectures in this section.

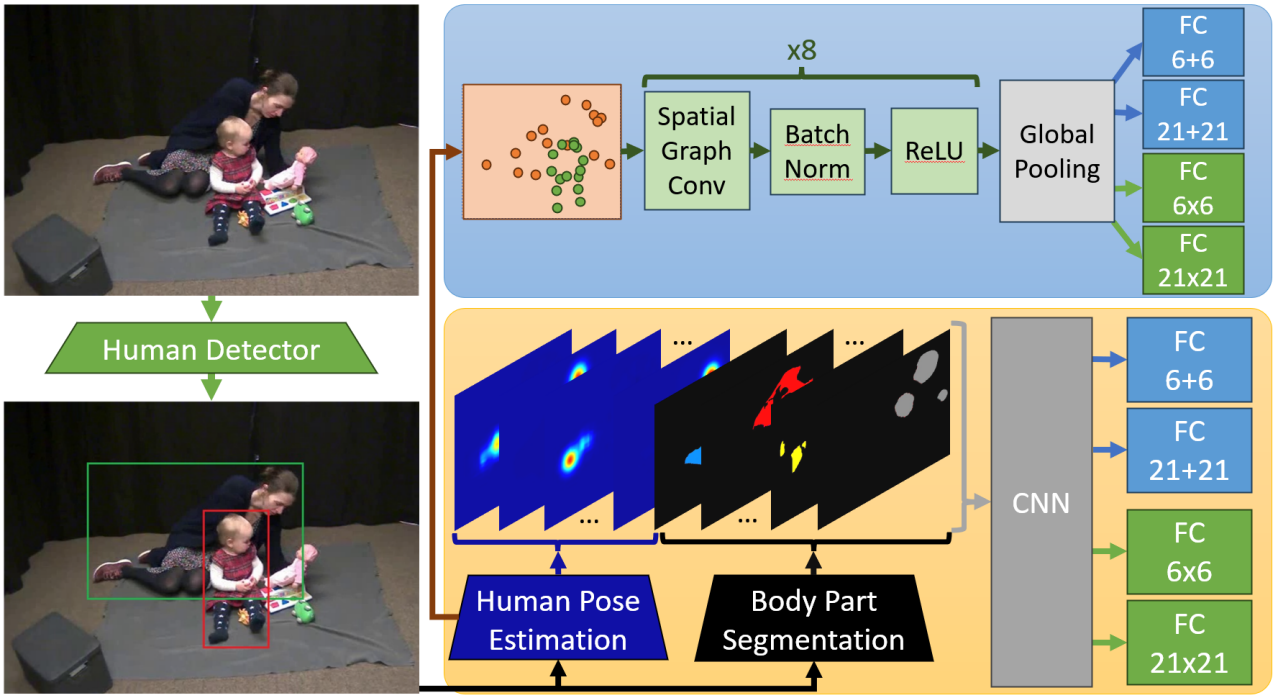

Figure 4: Pipelines for our two different methods: Pose2Contact (top) and Image2Contact (bottom). They both have four output heads for different resolutions for the contact segmentation and contact signature annotations. Frame used with permission.

4.1 Image2Contact

We use a convolutional neural network (CNN), with architecture and inputs largely following the implementation of a binary contact classifier introduced in [9]. The pipeline of this method appears in Fig. 4. We consider two modalities extracted from a 960×540 frame: (1) a set of 2D pose heatmaps, and (2) a set of body part segmentation maps. We start by detecting the two interacting people using the human detector YOLOx [12]. We select the two bounding boxes with the highest confidence and crop the tightest region that includes both bounding boxes with a margin of 11% in each direction. These cropped images are resized and padded to $N \times N$ ($N = 112$). Other modalities are mapped onto this cropped image space.

The two bounding boxes, detected by the human detector with the highest confidence scores, are processed through the DarkPose model [38] with an HRNet (W48) backbone [34] to obtain 17 body landmark heatmaps per person (*i.e.*, 34 maps for two people). For the second input modality, we use a state-of-the-art body part labeling model [23], which outputs 14 body part heatmaps and a single background heatmap. Diverging from Doyran *et al.* [9], we encode these 15 heatmaps with integer values to preserve the confidence results, as opposed to taking the most likely body part per pixel.

A modified ResNet-18 [16] is used as the backbone. The first layer is adjusted to take the 34 body landmark and 15 body part heatmaps. The last layer is replaced by four parallel fully connected (FC) layers with four sets of outputs, for different resolutions (6 and 21, respectively) and different annotations (contact segmentation and contact signature). The input of these FC layers has a dimension of 512 and the output dimension depends on the prediction (12 for $6 + 6$, 42 for $21 + 21$, 36 for 6×6, and 441 for 21×21).

4.2 Pose2Contact

Our second contact signature estimator, the graph convolutional method, is inspired by the 2P-GCN model [22]. Instead of using 2D pose heatmaps, we directly use the 2D poses of both people as inputs together with their confidences ($2 \times 17 \times 3$) for the spatial graph convolutional network. The dimensionality of the input is much lower compared to the CNN-based contact signature estimator (see Section 4.1). This is beneficial, because it reduces the risk of model overfitting during training. The drawback, of course, is that less information is available. For example, depth ordering is less apparent from 2D poses. While joints of two people might be close in an image, the 2D pose representation provides few cues whether they are in contact, or whether one is well behind the other.

The Pose2Image model is used to explore the benefits of using a graph convolutional network for representing both the intra- and inter-person connections with an adjacency matrix. The model has eight layers of spatial graph convolutions followed by batch normalization and an activation function (ReLU). After the eighth layer, a global pooling layer is applied. At the output stage, we use the same output strategy as the Image2Contact model with four different fully connected layers for different outputs (see Figure 4).

5 EXPERIMENTS AND RESULTS

In this section, we describe the experimental analysis of our automated methods. We first introduce the experimental setting, before outlining the model selection experiments. We present our main results in Section 5.2 and then explore different interaction types. We then present a qualitative analysis. We discuss our main findings and limitations in Section 5.5.

Table 1: Cross validation results for different modalities using Image2Contact.

2D pose heatmaps	Cropped image	Body part maps	6 + 6	6 × 6	21 + 21	21 × 21
✓			45.80 (1.19)	**24.79** (2.40)	31.48 (2.60)	11.16 (1.40)
	✓		39.72 (2.35)	18.46 (2.65)	22.05 (2.59)	7.44 (1.43)
		✓	44.96 (1.56)	22.70 (1.84)	29.72 (1.37)	10.10 (1.05)
✓		✓	**46.31** (2.05)	24.43 (2.22)	**31.67** (3.06)	**11.77** (1.11)
✓	✓	✓	44.65 (2.03)	23.48 (2.36)	30.53 (1.69)	10.95 (1.13)

Dataset. We use the YOUth Parent-Infant Interaction dataset. After processing each frame with the human detector and human pose estimation network, we discarded 5 parent-infant pairs which had missing detections for at least one subject, usually the infant being extremely occluded by the parent. The remaining 95 unique parent-infant pairs in the dataset are divided into five folds for our experiments. The first fold is only used for testing. The other four folds are used for model and hyperparamter selection. Each fold consists of around 400 contact frames and the frames from any parent-infant video can only be in one of the folds.

Implementation and Training. Our main annotations are contact signatures for 21×21, hence that is also the main prediction task and resolution for our models. To utilize the power of multi-task learning, we added three additional auxiliary tasks: 6×6 for lower resolution contact signature prediction, and contact segmentation with two different resolutions (6 + 6, 21 + 21). All of our results are reported using the Jaccard score (Eq. 1) in percentages as the evaluation measure. All of our model selection tables report the mean performance across four folds of cross validation using the training folds and the standard deviations are given in parentheses.

5.1 Model Selection Experiments

We perform the first model selection study for the modalities of the CNN-based contact signature estimator. In Table 1, using the processed input types (2D pose heatmaps and body part maps) performed better than using the raw input (*i.e.*, the cropped image), alone or in conjunction with the other modalities. Even when the backbone was pre-trained on the ImageNet [8] dataset and image data augmentation techniques were applied, using only the cropped image as the input performed significantly worse than any combination of the other input modalities. Adding the cropped image as the third modality also decreases the performance.

Table 2: Cross validation results with different backbones for the Image2Contact model.

	ResNet-18	ResNet-34	ResNet-50
6+6	**46.31** (2.05)	44.59 (1.85)	45.34 (1.79)
6x6	**24.43** (2.22)	23.58 (1.62)	23.11 (1.46)
21+21	**31.67** (3.06)	30.43 (1.79)	30.15 (1.23)
21x21	**11.77** (1.11)	11.42 (0.81)	10.97 (0.94)

Our second model selection experiment is to verify the suitability of the backbone for our setup. The input modalities are chosen as the 2D pose heatmaps and the body part maps. We compared two alternative ResNet models with the layer sizes of 34, and 50, in addition to the ResNet-18 model we have used. ResNet-18, which is the smallest of these models, performs the best in Table 2. The performance drops gradually when more layers are used.

Fixing ResNet-18 as the backbone and the processed input modalities (2D pose heatmaps and body part maps) as our CNN-based model's inputs, our third experiment checks whether the additional annotations (6 + 6, 21 + 21, and 6×6) are actually beneficial for the model if used in a multi-task learning setting (see Table 3).

Comparing the performance of predicting our main annotations (*i.e.*, 21×21), predicting only these annotations without any multi-task learning performs the worst (Table 3, row 2). Introducing only the smaller resolution version of these annotations (*i.e.*, 6×6) as the second learning task performs better (row 3), while adding only the same resolution contact segmentation annotations (*i.e.*, 21 + 21), has even higher results (row 5). The final row of the table shows that predicting all annotations together leverages multitask learning fully and yields the best results for our CNN-based model, Image2Contact. In Table 3 we have dashed out the annotations that are not predicted by a particular setting (as indicated by a zero loss weight in the corresponding column on the left).

We perform the last model selection experiment by testing the loss weight distribution for predicting the four annotations for our GCN-based contact signature estimator. Similar to the CNN-based contact signature estimator, Table 3 shows that Pose2Contact model also benefits from multi-task learning. The worst performing version for predicting the main annotations (21×21) is achieved by only learning to predict 21×21 (row 2). Introducing lower resolution 6×6 annotations or 21 + 21 contact segmentation annotations perform the best. Diverging from the CNN-based model, predicting all four annotations does not perform the best, signaling that for this specific model, focusing on either the contact signature annotations or the same resolution contact segmentation annotations are better than diversifying the output types. Side by side comparison of the contact segmentation heatmaps in Figure 2 reveals that the model predicts contact regions in a similar fashion to the human annotators. However, the best performing model predicts fewer regions for the parents and more for the children compared to the human annotations.

5.2 Contact Signature Prediction

We now address the task of automated contact signature prediction on the test set of the YOUth PCI dataset.

Selected Models. For the remaining experiments, we report only on the contact signature annotations (6×6 and 21×21) since the Pose2Contact model with the best settings would not be trained on the contact segmentation annotations (6 + 6 and 21 + 21 having loss weights of 0). For the Image2Contact model we use ResNet-18 with 2D pose heatmaps and body part maps as the input modalities.

Table 3: Cross validation results with different loss weights - Image2Contact and Pose2Contact.

Model	6+6	6x6	21+21	21x21	6+6	6x6	21+21	21x21
Image2Contact	0	1	0	0	-	**24.54** (2.09)	-	-
	0	0	0	1	-	-	-	9.90 (0.85)
	0	0.5	0	0.5	-	23.29 (4.00)	-	10.43 (1.65)
	0.5	0.5	0	0	45.69 (1.87)	23.58 (2.18)	-	-
	0	0	0.5	0.5	-	-	31.42 (2.61)	10.86 (0.81)
	0.25	0.25	0.25	0.25	**46.31** (2.05)	24.43 (2.22)	**31.67** (3.06)	**11.77** (1.11)
Pose2Contact	0	1	0	0	-	24.46 (2.31)	-	-
	0	0	0	1	-	-	-	7.90 (0.83)
	0	0.5	0	0.5	-	24.68 (2.36)	-	**10.65** (1.20)
	0.5	0.5	0	0	**49.06** (2.73)	**25.22** (1.65)	-	-
	0	0	0.5	0.5	-	-	**30.61** (2.45)	10.64 (1.32)
	0.25	0.25	0.25	0.25	48.71 (2.46)	24.35 (1.81)	30.49 (2.35)	9.50 (0.84)

Baseline. We considered different baseline methods such as uniform, majority, stratified, and constant. On the test set, the best performing baseline is constant, predicting all contact regions.

Fusion. As an ensemble method we considered three different summarizing functions for decision level fusion. Since both Image2Contact and Pose2Contact models output confidence scores per contact region, we perform fusion by passing these scores through minimum, average, and maximum functions.

With the two selected models, we report the results in Table 4, along with the fusion and baseline results. Image2Contact model performs better than Pose2Contact model. Both models separately perform significantly better than the baseline of predicting all contact. As expected, a decision level fusion of the two networks performed better than both individually. Average and minimum fusion functions performed better than the standalone results of each of the networks, with the minimum function performing the best. This might be caused by the selected loss function, evaluation metric and thresholds for sigmoid scores causing networks to predict contact regions with overconfidence. Using minimum as our fusion method corrects these cases. Taking the maximum score as the decision results in worse performance than both of the models.

Table 4: Test results on the test fold using the two models separately and with decision level fusion.

Model	6x6	21x21
Image2Contact	25.83	11.21
Pose2Contact	25.68	10.56
Fusion - Maximum	24.70	9.65
Fusion - Average	27.02	11.78
Fusion - Minimum	**27.52**	**13.59**
Baseline	9.9	1.1

5.3 Exploration of Interaction Types

To give more meaning to the test results, we annotated the test set with frequently occurring interaction type labels. The selected types are "picking up child" (4%), "supporting" (27%), and "on the lap" (24%). The rest of the frames are annotated as "other" (45%), and omitted from analysis. "Picking up child" occurs when the parent picks the infant up from the ground, usually grabbing below the shoulders and around the chest using both hands. First chart in Figure 6 shows that the parents' most common body parts in contact when picking the infant up are the arms, whereas the infants' core and arms are mostly in contact.

"Supporting" is annotated if the parent actively supports the infant to stand up, not to fall or balance. During this interaction type, the parents touch minimally to allow the child to move freely.

"On the lap" interaction type happens most frequently, and with the highest number of contact regions per interaction (Figure 6). The most common scenario is when they are sitting on the legs of their parents' with core-to-core contact and arms-to-arms contact.

We categorize our automated contact signature estimator's performance into three annotated interaction types. Figure 6 shows the performance results per interaction type on the test set. As the graph shows both "on the lap" and "picking up child" interactions are predicted with much better scores compared to the "supporting" class. We can see the contact distributions of the better predicted interaction types to be more symmetrical compared to the "supporting" interaction type (Figure 5).

5.4 Qualitative Analysis

The aim of the qualitative analysis and the visualizations are to provide some insights in how our work, while being not perfect, can still be used in a practical scenario. To gain more understanding in the potential to automatically assess PCIs, we discuss a visualization of our model's 6 + 6 contact segmentation predictions for two videos in Figure 1. For viewing purposes, we have combined right and left limbs resulting in 4 + 4 contact segmentation. The black lines indicate contact for that region. On the y axis, shows the body parts of the parent and child. The x axis represent the time. If we compare the predictions for the two videos it is clear that the first video includes much shorter interaction segments with contact whereas the second video includes multiple longer duration interaction segments with contact. Observing directly the videos also reveal that the parent in the first video plays with the child in a much relaxed way as the child is usually stationary or stays nearby the parent. Watching the second video it is seen that the child tries many times to walk outside the play area and the parent

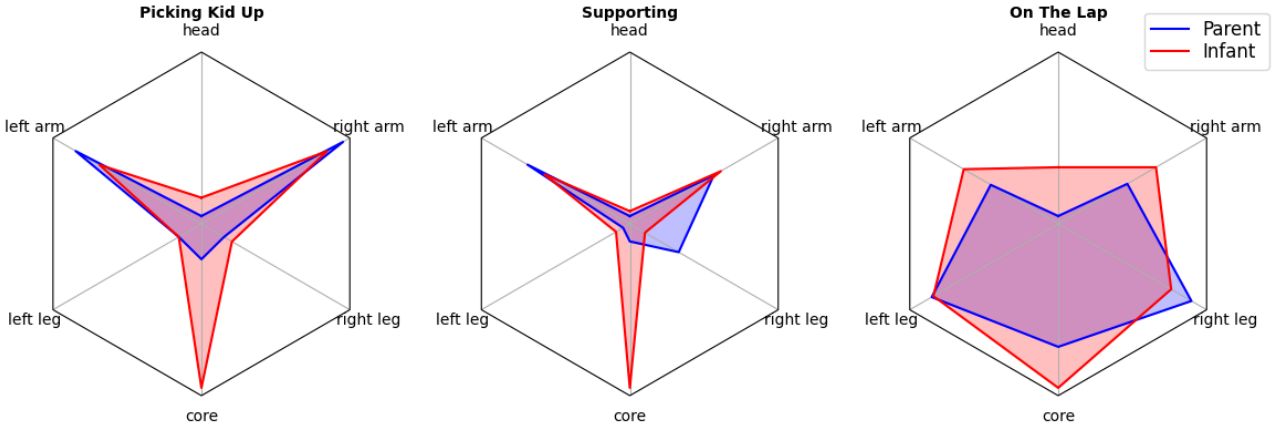

Figure 5: Contact distribution per interaction type using $6 + 6$ contact segmentation annotations.

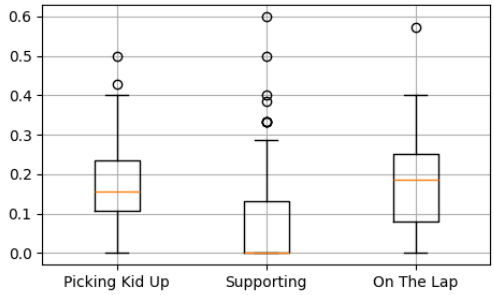

Figure 6: Jaccard scores per interaction type calculated with the best performing fusion model.

had to restrict him multiple times which the predictions show as blocks of contact regions. These blocks are not visible in the first video because the contact duration of the interactions are rather short compared to the second video except on region in the middle where the child sits on the laps of the parent. Here the model fails to capture the legs in contact for the majority of the frames while predicting correctly the arms being in contact.

5.5 Discussion

At a first glance, our selected models perform poorly on the fine-grained (21×21) contact signature estimation problem. But considering that this is a 441-class problem, it is clear that informative patterns are being learned. From the inter-annotator agreement calculations (Figure 3), we also observed that the problem is challenging. However, when contact regions are combined into 6 regions per person (6×6), both the inter-annotator agreement and model performances increase significantly. Our qualitative analysis reveal informative patterns, both between videos and within an interaction over time. While there is room for improvement, we argue that our models show potential to be used for real-life applications.

When comparing the two models, the Image2Contact model has more prediction power since its input is more extensive. Image2Contact utilizes the 2D pose heatmaps instead of pure maximum likelihood coordinates which allows it to have flexibility locally. Additionally, it uses body part maps which encode both the outer edges of the body parts, and depth ordering. These advantages of Image2Contact also brings more complexity to the model. We expect that Image2Contact has the potential to perform better when given more training data, whereas the performance of Pose2Contact might hit a ceiling much earlier.

By analyzing our annotations, we observed parent-child contact patterns that differ significantly from those observed in adult-adult interactions. The less constrained nature of touches between parents and their children causes a wide variation in contact signatures. Our work has demonstrated the feasibility of automatically assessing contact in PCIs but we argue that a fine-grained, temporal analysis can reveal more relations between characteristics of contact, and developmental outcomes.

Limitations. Our analyses provide a solid basis for automated analysis of parent-child contact, but our work is not without limitations. First, dataset-specific issues such as data quality (dark frames and relatively low resolution), uncalibrated dynamic cameras, realistic doll (false negatives with pose/human detectors), and most importantly the dataset size can be improved. Especially the limited number of training samples is likely to be insufficient to capture the wide variety of different touches and to train larger networks.

Second, the inputs of the Image2Contact model could be improved. In our implementation, the 2D body part heatmaps do not encode the identity of the person, unlike the 2D pose heatmaps. We expect that this causes the lower performance when the 2D body part heatmaps are used, because a model is unable to differentiate between which body part belongs to which person.

Third, our loss function and evaluation metric force models to learn more general predictions, making it harder to learn specific single-contact region interactions. Jaccard scores penalize the errors more if there are fewer positive labels. Future work can explore ways to categorize contact based on the region count and penalize the models equally to improve the training.

Finally, we have addressed frame-level contact signature prediction but, as revealed in the qualitative analysis, the analysis of contact over time provides meaningful indications of the quality of the interaction. Based on our work, such analyses are relatively straightforward, and we welcome future research in this direction.

6 CONCLUSION

In this paper we have, for the first time, addressed automated contact signature estimation in video recordings of free play parent-child interactions. We have shown that, even with a limited dataset, the current state-of-the-art allows us to analyze contact in detail. While there is room for improvement in terms of the accuracy of our models, our automated analyses can reveal patterns over time. As such, they provide opportunities to assess interaction quality at a larger scale and with far more detail than was previously possible. These estimators can be used to extract explainable contact features which can be further used to predict parenting styles, interaction quality, and other higher level annotations. They can also be used to automatically annotate large datasets for analyzing both session level and dataset level physical intimacy. Our findings, publicly available annotations, and models may open a doorway to further research to the largely unexplored field of parent-infant automatic contact signature estimation.

REFERENCES

[1] Mary D Salter Ainsworth, Mary C Blehar, Everett Waters, and Sally N Wall. 2015. *Patterns of attachment: A psychological study of the strange situation.* Psychology press, New York, NY.

[2] John Bowlby. 1982. Attachment and loss: retrospect and prospect. *American journal of Orthopsychiatry* 52, 4 (1982), 664.

[3] Alicja Brzozowska, Matthew R. Longo, Denis Mareschal, Frank Wiesemann, and Teodora Gliga. 2021. Capturing touch in parent–infant interaction: A comparison of methods. *Infancy* 26, 3 (2021), 494–514.

[4] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser A. Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (2019), 172–186. Issue 1.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. Springer, Munich, Germany, 801–818.

[6] Qingshuang Chen, Rana Abu-Zhaya, Amanda Seidl, and Fengqing Zhu. 2019. CNN Based Touch Interaction Detection for Infant Speech Development. In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, San Jose, CA, 20–25.

[7] Qingshuang Chen, He Li, Rana Abu-Zhaya, Amanda Seidl, Fengqing Zhu, and Edward J Delp. 2016. Touch event recognition for human interaction. *Electronic Imaging* 2016, 11 (2016), 1–6.

[8] Jia. Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*. IEEE, Miami, FL, USA, 248–255.

[9] Metehan Doyran, Ronald Poppe, and Albert Ali Salah. 2023. Embracing Contact: Detecting Parent-Infant Interactions. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 198–206. https://doi.org/10.1145/3577190.3614147

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Venice, Italy, 2334–2343.

[11] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2020. Three-Dimensional Reconstruction of Human Interactions. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, Seattle, WA, USA, 2596–2605.

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430* (2021).

[13] Tsfira Grebelsky-Lichtman. 2014. Children's Verbal and Nonverbal Congruent and Incongruent Communication During Parent–Child Interactions. *Human Communication Research* 40 (07 2014). https://doi.org/10.1111/hcre.12035

[14] Wen Guo. 2020. Multi-person pose estimation in complex physical interactions. In *Proceedings of the 28th ACM International Conference on Multimedia*. IEEE/CVF, Seattle, WA, USA, 4752–4755.

[15] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2022. Multi-person extreme motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13053–13064.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[17] Matthew J Hertenstein, Julie M Verkamp, Alyssa M Kerestes, and Rachel M Holmes. 2006. The communicative functions of touch in humans, nonhuman primates, and rats: a review and synthesis of the empirical research. *Genetic, social, and general psychology monographs* 132, 1 (2006), 5–94.

[18] Berfu Karaca, Albert Ali Salah, Jaap Denissen, Ronald Poppe, and Sonja M.C. de Zwarte. to appear. Survey of Automated Methods for Nonverbal Behavior Analysis in Parent-Child Interactions. In *Proceedings of the International Conference on Face and Gesture Recognition (FG)*.

[19] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *Proceedings of the IEEE international conference on computer vision*. IEEE, Santiago, Chile, 2938–2946.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.

[21] Lisa J. G. Krijnen, Marjolein Verhoeven, and Anneloes L. van Baar. 2023. Observing mother-child interaction in a free-play vs. a structured task context and its relationship with preterm and term born toddlers' psychosocial outcomes. *Frontiers in Child and Adolescent Psychiatry* 2 (2023), 1176560.

[22] Zhengcen Li, Yueran Li, Linlin Tang, Tong Zhang, and Jingyong Su. 2023. Two-Person Graph Convolutional Network for Skeleton-Based Human Interaction Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 33, 7 (2023), 3333–3342. https://doi.org/10.1109/TCSVT.2022.3232373

[23] Kevin Lin, Lijuan Wang, Kun Luo, Yinpeng Chen, Zicheng Liu, and Ming-Ting Sun. 2020. Cross-Domain Complementary Learning Using Pose for Multi-Person Part Segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* 31 (2020), 7277–7286. Issue 3.

[24] Ashley Montague. 1986. *Touching: The human significance of the skin* (3rd ed.). Harper & Row, New York, NY, USA.

[25] N Charlotte Onland-Moret, Jacobine E Buizer-Voskamp, Maria EWA Albers, Rachel M Brouwer, Elizabeth EL Buimer, Roy S Hessels, Roel de Heus, Jorg Huijding, Caroline MM Junge, René CW Mandl, et al. 2020. The YOUth study: Rationale, design, and study procedures. *Developmental cognitive neuroscience* 46 (2020), 100868.

[26] Muhammad Rameez Ur Rahman, Luca Scofano, Edoardo De Matteis, Alessandro Flaborea, Alessio Sampieri, and Fabio Galasso. 2023. Best Practices for 2-Body Pose Forecasting.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, Springer, Munich, Germany, 234–241.

[28] Reva Rubin. 1963. Maternal touch. *Nursing outlook* 11 (1963), 828–829.

[29] Sara E Schroer and Chen Yu. 2022. The Real-Time Effects of Parent Speech on Infants' Multimodal Attention and Dyadic Coordination. *Infancy : the official journal of the International Society on Infant Studies* 27, 6 (2022), 1154–1178. https://doi.org/10.1111/infa.12500

[30] Jack P Shonkoff and P Hauser-Cram. 1987. Early intervention for disabled infants and their families: a quantitative analysis. *Pediatrics* 80, 5 (1987), 650–658.

[31] Jack P Shonkoff and Samuel J Meisels. 2000. *Handbook of Early Childhood Intervention* (2 ed.). Cambridge University Press.

[32] Anja Sommer, Claudia Hachul, and Hans-Günther Roßbach. 2016. *Video-Based Assessment and Rating of Parent-Child Interaction Within the National Educational Panel Study.* Springer Fachmedien Wiesbaden, Wiesbaden, 151–167. https://doi.org/10.1007/978-3-658-11994-2_9

[33] Alexandros Stergiou and Ronald Poppe. 2019. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding* 188 (2019), 102799.

[34] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, Long Beach, CA, USA, 5693–5703.

[35] Juulia T. Suvilehto, Enrico Glerean, Robin I. M. Dunbar, Riitta Hari, and Lauri Nummenmaa. 2015. Topography of social touching depends on emotional bonds between humans. *Proceedings of the National Academy of Sciences* 112, 45 (2015), 13811–13816. https://doi.org/10.1073/pnas.1519231112

[36] Ines Van Keer, Eva Ceulemans, Nadja Bodner, Sier Vandesande, Karla Van Leeuwen, and Bea Maes. 2019. Parent-child interaction: A micro-level sequential approach in children with a significant cognitive and motor developmental delay. *Research in Developmental Disabilities* 85 (2019), 172–186. https://doi.org/10.1016/j.ridd.2018.11.008

[37] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. 2023. Hi4D: 4D instance segmentation of close human interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17016–17027.

[38] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. 2020. Distribution-Aware Coordinate Representation for Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE/CVF, Seattle,

WA, USA, 7093–7102.
[39] Feng Zhang, Xiatian Zhu, and Mao Ye. 2019. Fast human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE/CVF, Long Beach, CA, USA, 3517–3526.

7 SUPPLEMENTARY MATERIAL

7.1 Distribution of Contact Signatures

Here, we analyze how the contact signatures of our annotated dataset are distributed. To this end, we employ t-SNE (t-distributed Stochastic Neighbor Embedding to visualize our fine-grained, 21×21 contact signatures in a latent 2D space. The top part of Figure 7 shows this distribution. Dots with a blue color represent the training samples and the remaining red, yellow, and green colors show the performance results per test sample using our best performing model.

We manually inspected what different regions in the 2D space corresponded to, indicated by the ellipses and textual labels. clearly, similar signatures correspond to similar interactions. It can be seen that the dataset has many samples of child sitting on the laps of their parents whereas hand-to-hand contact occurs significantly less than the others. There are also some small clusters that shows very similar frames with parent and child being relatively still and having almost the same contact signatures between different frames. Still, there is significant variation within the clusters, which reflects the challenging nature of the classification problem.

The bottom plot in Figure 7 only shows the test samples with corresponding performance-related colors. Red color shows samples predicted with a Jaccard score lower than 0.25, yellow color represents the samples predicted with a Jaccard score between 0.25 and 0.5, and the green color indicates the best predicted samples with a Jaccard score of higher than 0.5.

Overall, the distribution of the predictions nicely follows that of the full dataset. This is important because it means that there are no obvious systematic biases in the predictions.

Still, some regions stand out. For example, the region between the "parent legs to child bottom/back" and "child sitting on the lap" contains several predictions but does not have any corresponding training data. Interestingly, the Jaccard scores of the predictions are reasonable.

The visualization also reflects the analysis of interaction type in Section 5.3 of the main paper. We already observed better performance for the "on the lap" type. From the figure, we can appreciate that the Jaccard scores of the predictions in this area are indeed often higher.

7.2 Average Contact Pair Counts

To understand the contact signature distribution better in our dataset, we count the contact pairs for the 21×21 contact signature annotations of each frame. Then we calculate the average for each video in the dataset, with results sorted and visualized in Figure 8. As can be seen in the graph, the majority of the videos have less than 10 contact region pairs per frame on average. This highlights the sparseness of the contact signature problem (441 contact region pair possibility). Six videos only have frames with only one contact region pair per frame.

Still, the average number of contact pairs is significant, indicating that our classification task is much more difficult than a simple binary decision.

During manual annotations, our annotators also noted that annotating the videos with less than two contact region pairs per frame took much less time compared to annotating the videos with more than 10 contact region pairs per frame. Aside from annotating more regions, it was significantly more challenging to differentiate between some of the contact regions such as upper arm and lower arm of the infant or chest and stomach of the parent. Even though using four cameras helped the annotators to make better decisions, the highly occluded nature of the parent-infant interactions made it difficult to use more than two cameras at a time since the other two cameras only showed the back of the parent occluding the infant extremely or often completely.

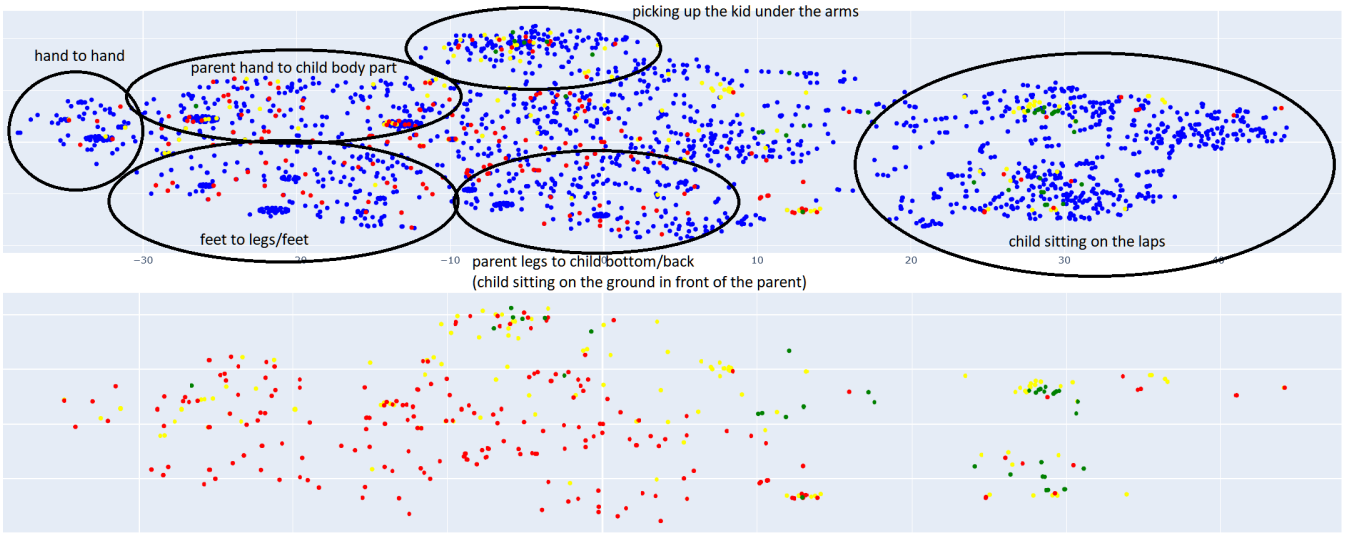

Figure 7: t-SNE visualization of the 21×21 contact signature annotations. (blue: training set; red: test set, jaccard score < 0.25; yellow: test set, 0.25 < jaccard score < 0.5; green: test set, 0.5 < jaccard score)

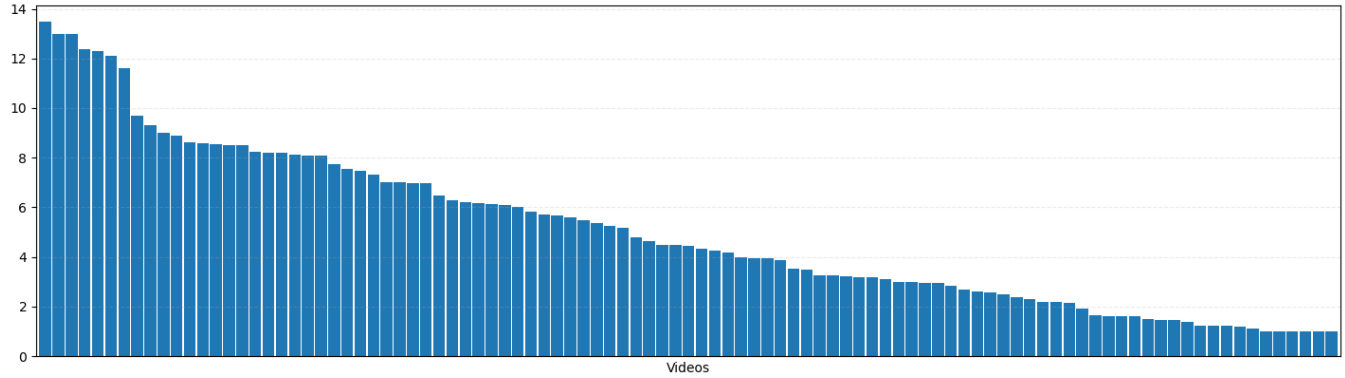

Figure 8: Average contact pair counts per frame for each of the videos in the dataset using 21×21 contact signature annotations.