Embracing Contact: Detecting Parent-Infant Interactions

METEHAN DOYRAN, Utrecht University, NL
RONALD POPPE, Utrecht University, NL
ALBERT ALI SALAH, Utrecht University, NL and Boğaziçi University, TR

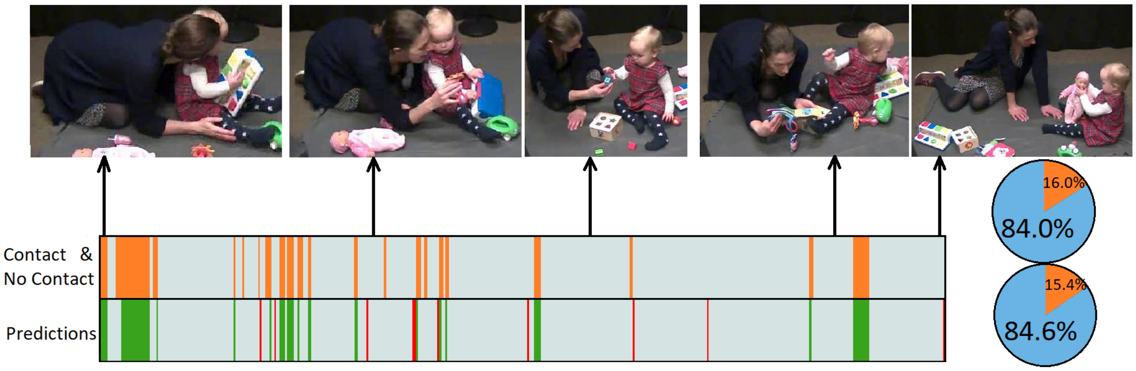

Fig. 1. Conceptual overview of the contact detection system.

We focus on a largely overlooked but crucial modality for parent-child interaction analysis: physical contact. In this paper, we provide a feasibility study to automatically detect contact between a parent and child from videos. Our multimodal CNN model uses a combination of 2D pose heatmaps, body part heatmaps, and cropped images. Two datasets (FlickrCI3D and YOUth PCI) are used to explore the generalization capabilities across different contact scenarios. Our experiments demonstrate that using 2D pose heatmaps and body part heatmaps yields the best performance in contact classification when trained from scratch on parent-infant interactions. We further investigate the influence of proximity on our classification performance. Our results indicate that there are unique challenges in parent-infant contact classification. Finally, we show that contact rates from aggregating frame-level predictions provide decent approximations of the true contact rates, suggesting that they can serve as an automated proxy for measuring the quality of parent-child interactions. By releasing the annotations for the YOUth PCI dataset and our code[1], we encourage further research to deepen our understanding of parent-infant interactions and their implications for attachment and development.[2]

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

Additional Key Words and Phrases: parent-child interaction; interaction analysis; contact detection; pose estimation; convolutional neural network

[1] Code and annotations: https://github.com/dmetehan/ContactClassification.git.
[2] This is the uncorrected author proof.

Manuscript submitted to ACM

ACM Reference Format:
Metehan Doyran, Ronald Poppe, and Albert Ali Salah. 2023. Embracing Contact: Detecting Parent-Infant Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3577190.3614147

1 INTRODUCTION

Contact is a prominent component of nonverbal communication, and plays a fundamental role in the communication of emotion [16]. This is especially true for infants, whose speech understanding and production skills are only in early development. In parent-infant interactions, physical contact has an important function in early child development, encouraging attachment and emotional regulation [1, 3]. The analysis of contact in parent-infant interactions is therefore valuable to assess the quality of the interaction. Yet, when considering contact between parent and infant, the predominant measurement method is self-report through questionnaires [4]. Such methods provide summary annotations at the level of the interaction and are inherently subjective. While several more objective observational measures have been developed and used, they lack consistency in their definition and operationalization [25]. Specifically, they either consider observable or functional touch and differ in the temporal granularity of the annotations.

We argue in this paper that even a binary assessment of contact (*i.e.*, there is contact or not) already provides a significant amount of insight into the interaction, especially when made continuously, *i.e.*, at each frame. As providing such annotations manually is very time-consuming and error-prone, we consider automated methods. However, compared to other modalities in face-to-face co-located interactions, the detection of inter-personal contact has received surprisingly little attention [2].

Two prior works on contact detection in parent-infant interaction videos are [7] and [8]. Both consider a fixed, seated interaction with limited freedom of movement for both interactants. In contrast, we focus on free play in a larger space with toys, see Fig. 1. Arguably, this constitutes a more complex scenario due to the more varied and dynamic positioning of both parent and infant. But especially in such dynamic interactions, subjective contact measurement methods are expected to be biased [4]. An automated, objective contact detection method is therefore a welcome alternative.

In this paper, we investigate the feasibility of a vision-based method to automatically provide such frame-level contact annotations. We use a multimodal convolutional neural network (CNN), adapted from [11], with three input modalities, namely 2D pose, image, and body parts. Our main contributions are:

(1) We demonstrate the feasibility of using automatic contact detection by providing both quantitative and qualitative results on a free play parent-infant interaction dataset.
(2) We perform systematic experiments with our multimodal CNN model to understand the contribution of the various components.
(3) We provide physical contact annotations for the YOUth PCI dataset [21] to encourage more research on automatic contact detection of parent-infant interaction during free play. We also release our code base.

The remainder of this paper is as follows. We first discuss related work on parent-child interaction analysis, with a focus on physical contact. In Section 3, we introduce the YOUth PCI dataset and detail our contact annotations. Our methodology, including the discussion of the architecture and training of our CNN model, appears in Section 4. We quantitatively and qualitatively evaluate our approach in Section 5 and discuss the results in Section 6. Section 7 concludes the paper.

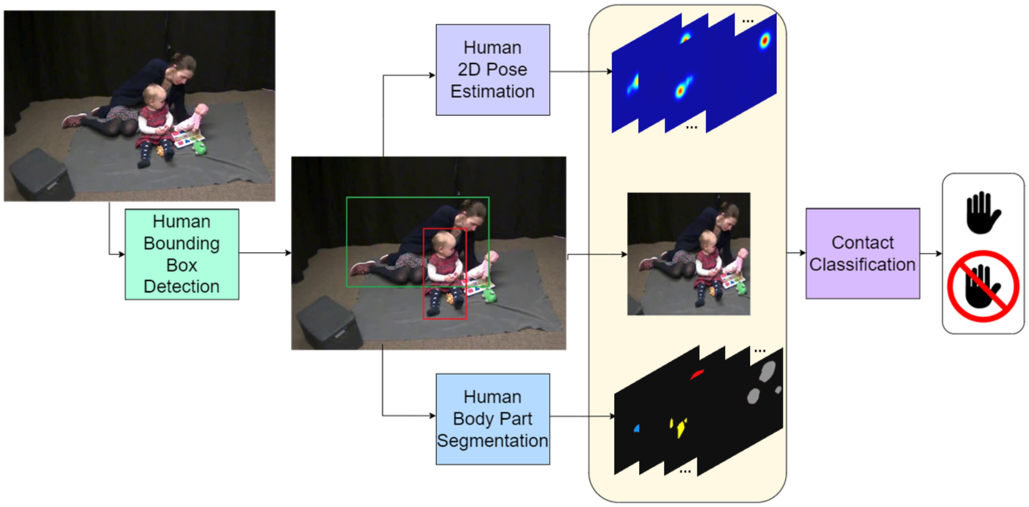

Fig. 2. Pipeline of our approach. YOLOx is used to detect the human bounding boxes in the video. This is followed by DARK Pose with HR-Net to generate heatmaps for body parts, and human body part segmentation to enable a tri-modal classification.

2 RELATED WORK

Our work builds upon several areas of research, including parent-child interaction studies, contact classification, pose detection, and body part segmentation. In this section, we discuss the relevant literature in these areas and highlight the contributions of our work in addressing the challenges in parent-child contact analysis.

Contact in parent-child interactions. Parent-infant interaction has been widely studied in the context of psychology, child development, and attachment theory [1, 3]. Inter-personal touch is also widely recognized to maintain a crucial role in a child's life during development [24], setting the groundwork for other communication forms that emerge later in life [16]. Touch influences physiological states, supports healthy biological growth, and serves a vital role in social development [20].

Contact classification. Despite the importance of touch in parent-child interactions, very little work has been done to analyze physical contact detection in parent-child interactions from a computer vision perspective, with the exception of [7, 8]. In both [8] and [7], a computer vision-based approach was used to detect touch events between a parent's hand and a child's body in a restricted interaction setting, where both infant and parent are seated facing each other. In [7], a precision of about 48% is obtained with a good recall (99%) for such a scenario, and an additional classification is made for the location of the touch on the infant's body. Our work is similar, in that a multi-step analysis is made for detecting touch events, but the scenarios we investigate are more challenging. We do not only consider hand-to-body touch events, therefore no specific attention is devoted to hand segmentation in this work.

Fieraru *et al.* [11] developed a novel approach for detecting and analyzing physical contact between people from visual data, addressing a critical gap in the literature. Although their main focus was 3D reconstruction, their methodology successfully enabled the classification of contact between adults in various settings, predominantly from the sports domain. Their dataset, FlickrCI3D, is used in our transfer learning experiments in Section 5.

2D pose estimation. Our proposed methodology in this paper leverages pose detection to facilitate contact classification. Several works have made significant advancements in this problem, such as OpenPose [5], AlphaPose [10], PoseNet [17], DarkPose [30], and HRNet [27]. These pose detectors have been widely used in various computer vision applications, including human activity recognition and human-object interaction analysis. In our work, we employ DARK Pose with HRNet to derive meaningful features for contact classification. Our early pilot experiments (not reported here) showed that these models work better than others for detecting the poses of infant bodies.

Body part segmentation. Body part segmentation has been a much-researched task in computer vision, with state-of-the-art models like DeepLab [6] and U-Net [23] reaching good performances. More recently, the Segment Anything Model (SAM, [18]) is developed with a massive training set of a billion segmentation masks for 11 million images, indicating that the improvements in visual segmentation will continue soon.

Lin *et al.* [19] proposed a cross-domain adaptation method for body part segmentation, highlighting its potential in improving the performance of related tasks. Segmentation allows more semantically structured modeling for interaction analysis, by distinguishing between different body parts which are used in different contexts.

Human-human interactions. Several studies have been conducted to analyze and predict human-human interactions [13, 14, 22, 28]. These works explore 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 use attention mechanisms, different initialization procedures, and network architectures to improve the performance and stability of interaction analysis (see [26] for a review). Research on contact detection might benefit from human-human interaction classification, and *vice versa*.

3 PARENT-INFANT INTERACTION DATASET

For our analyses, we use parent-child interaction (PCI) videos from the YOUth Cohort study [21]. In each video, a parent and a child freely play[3]. We focus on a set with 10-month-old children (average 11.4 months old with a standard deviation of 1.2 months) and use the terms *child* and *infant* interchangeably.

Recordings are made in a room with a play area and a box of toys. Parents and infants are free to play with toys on the ground (see Fig 1). Sometimes infants and parents play together and sometimes infants ignore the parents and play by themselves. Occasionally, infants crawl out of the play area and leave the view of the cameras, to be brought back by their parents. The interactions are unstructured as there are no requirements about which toy to play with and the variability of the toys available (car, doll, switch box, flower, book, baby bottle, shape box with four different colored shapes). This setup allows for a wide variety of interactions, ranging from infants playing with their parents, sitting on their parent's laps to look at a book, or playing individually with toys.

The YOUth Cohort study is ongoing, with interactions recorded each week. Interactions are recorded from four cameras close to ground level to capture the interaction more closely. We limit our analysis to a single view, to address a setting that is more common and does not require a synchronized multi-camera setup. A total of 94 interactions with unique parent-child pairs were randomly selected from the available pool. The videos were temporally segmented to begin when the experimenter exits the scene and conclude upon the experimenter's return. This resulted in an average video duration of 12:33 minutes (standard deviation of 31 seconds). Our data covers almost 20 hours of interaction.

3.1 Physical Contact Annotations

Our definition of contact includes both intentional and unintentional contact. Intentional contact mostly concerns touch with the hand (such as grabbing, caressing, hitting, or supporting), whereas unintentional contact can be between

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

any two body parts, such as a foot touching a leg. The annotation considers only the 2D view of the selected camera. Annotations are provided for individual frames so that the annotators are not informed of temporal information. In this way, our automated approach is not put at a disadvantage when being presented with a single frame, without temporal context.

For manual annotation of contact, a frame was selected every 5 seconds. The frames are annotated by a single coder. A second coder annotated a subset of 200 frames sampled randomly from the dataset to have an equal number of *contact* and *no contact* classes. The inter-annotator agreement between the two coders is calculated using Cohen's Kappa coefficient and results in an 85% agreement. The annotations included contact, no contact, and *ambiguous* classes following the convention from the FlickrCI3D dataset. Since annotations are made from a single view, frames of free play parent-infant interaction can include significant occlusion scenarios such as an infant sitting between the legs of the parent or a parent hugging the infant. Also, the scale difference between parents and infants causes infants to be partly occluded by their parents regularly. In cases where the annotator cannot judge whether there is contact or not between parent and infant, the frames are annotated as ambiguous. Consequently, after discarding 7, 185 ambiguous frames (similar to [11]), our experiments and analysis include 6, 915 frames, with 2, 983 (43.13%) of those frames containing instances of physical contact between the parent and the infant, and the rest of the frames (56.87%) being labeled as no contact. The data is divided into training, validation and test set based on the parent-child pairs considering the gender of the child. All the frames from each parent child pair are in either training (50 videos), validation (20 videos) or test set (20 videos).

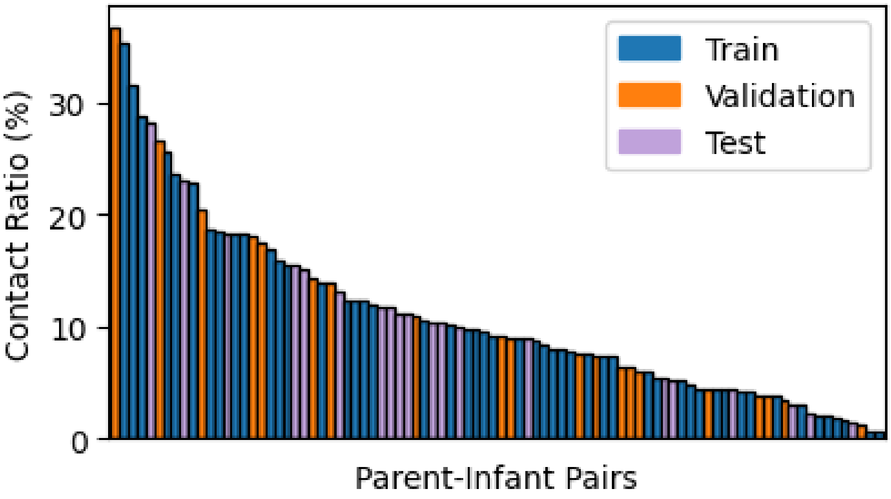

Fig. 3. Contact ratio per parent-infant pair, in decreasing order. Colors indicate to which set an interaction belongs.

Fig. 3 shows the ratio of physical contact per video, colored per set (training, validation, test). The contact ratio varies significantly with an average of 21.97% (standard deviation 15.96%). The maximum contact ratio in a video is 73.61% and the minimum contact ratio is 1.29%. These statistics also include the frames where the contact could be judged and was annotated as ambiguous.

To increase our understanding of the variation in the dataset, we investigate the proximity between the parent and the infant. From the output of our 2D pose detection, explained in detail in Section 4, we calculate the minimum distance

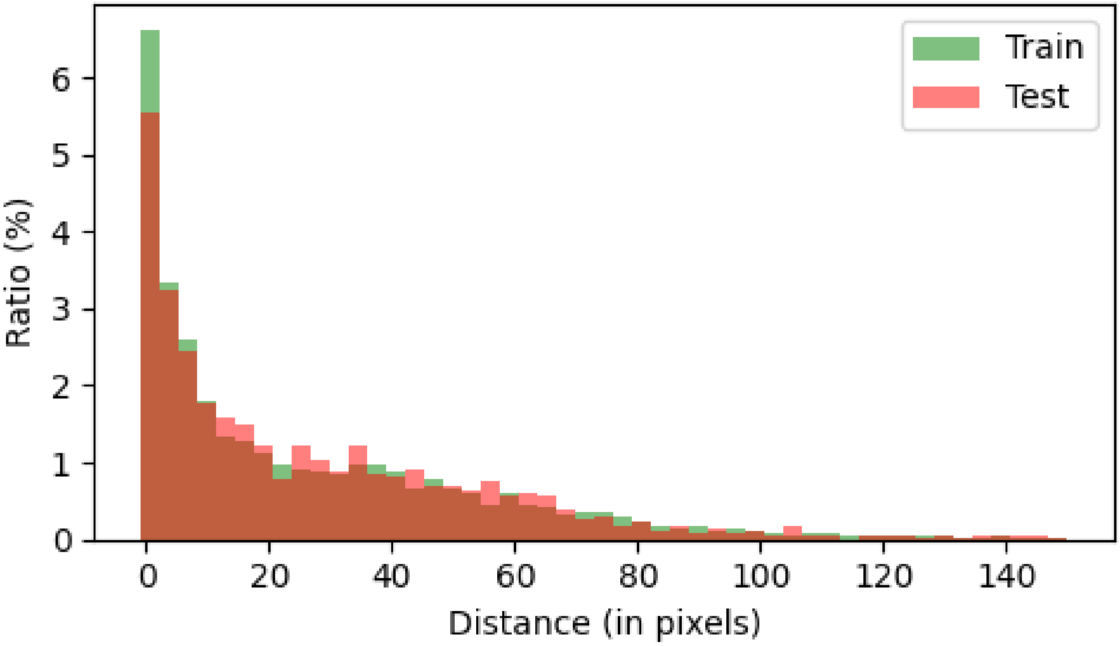

Fig. 4. 2D proximity distribution in pixels for training and test sets. The resolution of a frame (see Fig. 2) is **960** × **540**.

in pixels between all pairs of joints between the parent and the infant. Distances are measured in the original frame, with a resolution of 960 × 540. Fig. 4 shows the 2D proximity distribution across training and test sets, which follow a similar trend. Close interactions make up a significant part of the data, which provides evidence for the challenging nature of the dataset. Fig. 10 shows a random sample of poses for different proximity ranges. We discuss it in detail in Section 6.

4 CNN-BASED CONTACT CLASSIFICATION

For our automated contact classification, we develop and train a convolutional neural network (CNN). The architecture of our CNN largely follows [11], but since there is no public code for this approach, we re-implemented it and replaced some sub-modules with recent alternatives. We provide an open-source implementation of our code to allow other researchers to replicate our results, and to build on our work.

The pipeline of our method appears in Fig. 2. We consider three modalities extracted from a 960 × 540 frame: (1) RGB image cropped around the two interactants, (2) 2D pose heatmap, and (3) body part segmentation maps. We start by detecting the two interacting people. Unlike [11], we do not use a bottom-up body pose estimator. Our preliminary experiments showed that bottom-up pose detection methods struggle with close parent-infant interaction frames, whereas top-down pose detectors have fewer problems with the intertwined nature of individuals in the YOUth PCI dataset. In our pipeline, we replace the human detection with a YOLOx human bounding box detector [12].

In the pre-processing stage of our pipeline, we select the two bounding boxes with the highest confidence, as detected by the human detector. We then crop the tightest region that includes the bounding boxes of both people and add a margin of 11% in each direction (top, bottom, left, right) to alleviate issues with cropping. These cropped images are resized and padded to $N \times N$ (112×112) and used as our first input modality.

We process each of the two bounding boxes to estimate the 2D locations of the key joints of the persons. We use the DARK Pose model [29] with an HR-Net (W48) backbone [27] to obtain 17 body landmark heatmaps per person (*i.e.*, 34 maps). A heatmap reflects the probability that a specific body joint is found at each location in the image. These heatmaps are transformed and scaled to match the spatial dimensions of the crop. Unlike [11], we did not use landmark locations to generate new pose heatmaps with normal distributions around the landmark locations. Our preliminary experiments showed superior results when the 2D pose heatmaps are directly used as our second input modality.

To extract semantic human features, we replaced the 2D body part labeling of [11] with a recent, state-of-the-art model [19]. The output of this process is a labeled segmentation of each body part. A map contains the body parts of both interactants, which is the common output of body part segmentation algorithms if no masking step is applied. The segmentation maps provide information about the visibility and shape of the body parts. By considering all body part segmentations, information about proximity and overlap in the image is revealed. 2D body part labels are calculated on the crop, resulting in one background and 14 body part heatmaps. These heatmaps are binary encoded and used as our third input modality.

A modified ResNet-50 [15] is used as the backbone for our contact classifier. The first layer is adjusted to take not only the cropped image as the input but also the 34 body landmark heatmaps and the 15 body part heatmaps. The first layer convolutional weights corresponding to the image input are copied from a ResNet pre-trained on the ImageNet dataset [9]. The last layer of the ResNet is replaced by a fully connected layer with two outputs, for contact and no contact, respectively.

5 EXPERIMENTS AND RESULTS

To evaluate the performance of our model on parent-child interactions, we perform our experiments on the YOUth PCI data.

Metrics. To quantitatively assess our approach, we use three different metrics: accuracy, balanced accuracy, and F1 score. Accuracy can be misleading for datasets with imbalanced class distribution, whereas balanced accuracy is calculated by taking the average accuracy across each class without counting how often the classes occur in the dataset. We also use the F1 score for the contact class to identify if the method is successful in detecting contact cases.

Model verification. To verify that our implementation produces predictions in line with the models presented in [11], we first train and evaluate our model on the FlickrCI3D dataset. The dataset contains $55,095$ images, where each image contains two or more people and each pair of people are annotated for contact, no contact, and uncertain contact. The training set includes $49,372$ pairs annotated as contact, $14,733$ pairs annotated as no contact, and $17,197$ pairs discarded for ambiguity.

Our best model achieves a test accuracy of 82.79%. When we check the balanced accuracy (75.28%) and the F1 score (62.08%), however, we make two observations. First, the balanced accuracy is notably lower than the 84.6% that was reported in [11]. Second, this difference is caused by a lower accuracy for contact, with 84.4% and 61.4% for [11] and our model, respectively. Our model scores better on no-contact interactions, with 89.2% compared to 84.8% in [11]. The differences in performance between our model and that in [11] could be due to the different algorithms to produce the multimodal inputs, or due to different hyper-parameters such as batch size.

Baselines. We include three baselines in our main comparison. *All No Contact* is the result of always predicting the majority class: no contact. *Random Guess* predicts contact or no contact each with 50% probability. *Informed Guess* uses the prior training probabilities, 41.81% and 58.19% for contact and no contact, respectively.

Table 1. Performance comparison on the YOUth PCI dataset. The last three rows are the baselines. The standard deviation of 10 runs for each setting is given between parentheses.

Training	Acc.	Bal. Acc.	F1 Score
FlickrCI3D	57.68 (4.98)	58.31 (3.35)	54.71 (4.07)
FlickrCI3D+YOUth	79.95 (0.80)	79.77 (0.78)	76.63 (0.95)
YOUth	**80.41** (1.89)	**80.42** (2.07)	**77.39** (2.51)
All No Contact	58.19	50.00	0
Random Guess	50.00	50.00	50.00
Informed Guess	51.34	50.00	41.81

Main results. For our main results, we compare the aforementioned network that was only trained on FlickrCI3D to the same network that was subsequently fine-tuned on the YOUth PCI dataset, and to a network that was trained on YOUth PCI from scratch. A summary of the results, collected over 10 runs for each setting to minimize the effects of random initialization and batch order, appears in Table 1.

Table 2. Effect of different input modalities on the YOUth PCI dataset. The standard deviation for 10 runs is in parentheses.

2D Pose Heatmaps	Cropped Image	Body Part Maps	Accuracy	Balanced Accuracy	F1 Score
✓			80.39 (2.04)	80.47 (1.87)	77.54 (2.08)
	✓		76.74 (2.42)	76.27 (2.10)	72.44 (2.62)
		✓	80.98 (1.36)	80.09 (1.38)	76.60 (1.83)
✓	✓		77.96 (2.01)	78.34 (1.81)	75.35 (2.07)
✓		✓	**81.88** (1.20)	**81.72** (1.10)	**78.82** (1.34)
	✓	✓	79.83 (1.91)	79.74 (1.25)	76.65 (1.29)
✓	✓	✓	80.41 (1.89)	80.42 (2.07)	77.39 (2.51)

The cross-dataset test results on YOUth PCI with only FlickrCI3D as the training set demonstrate a performance that is above all three baselines, except the accuracy score that is comparable to the All No Contact baseline. Given that no YOUth PCI data was used in the training, it becomes clear that the model trained on FlickrCI3D generalizes to some extent, most specifically for the contact class. This also informs us that the two datasets have similarities.

However, these results are largely surpassed when domain-speci-fic data from YOUth PCI is involved. Comparing the second model (training on FlickrCI3D and finetuning on YOUth PCI) and the third model (training on YOUth PCI from scratch) shows that training from scratch on YOUth PCI yields slightly better results, whereas the fine-tuned model has more robust performance in terms of a lower standard deviation.

The predictions are statistically significantly above all baselines. As an indication, for the first run, paired t-tests between the model trained from scratch and the All No Contact ($t(1575)=29.50$, $p<0.001$), Random Guess ($t(1575)=-9.25$, $p<0.001$), and Informed Guess ($t(1575)=-6.79$, $p<0.001$) baselines demonstrate significantly higher performance for our model.

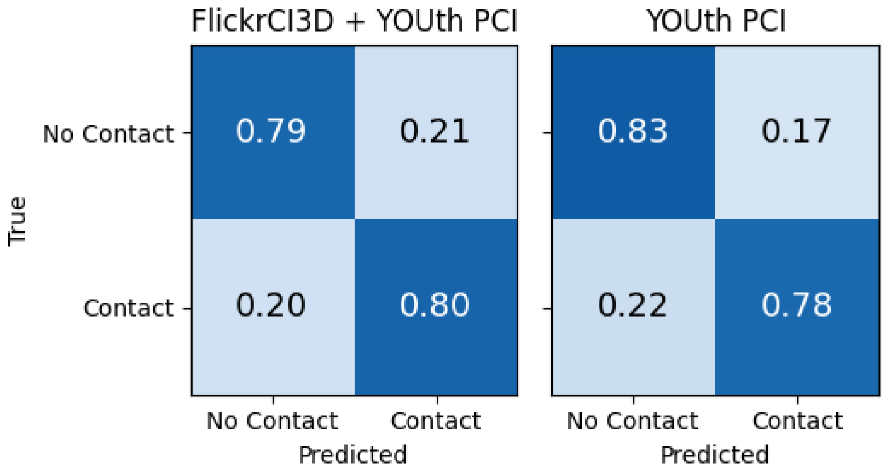

Fig. 5. Confusion matrices for the models; pre-trained on FlickrCI3D and fine-tuned on YOUth PCI (left), trained on YOUth PCI from scratch (right).

When comparing the two confusion matrices in Fig. 5 the model trained from scratch learned to identify the no contact class better, while maintaining a comparable accuracy for the contact class. However, the results are quite close to each other and no conclusive judgments can be made.

Effect of input modality. Our second quantitative analysis focuses on using different combinations of input modalities. Results are shown in Table 2. To minimize the effects of random initialization and batch order, we report the mean and standard deviations over 10 runs for each combination of the three input modalities. For each of the three evaluation metrics, the best performance on the YOUth PCI dataset is the model trained with a combination of 2D pose heatmaps and body part heatmaps. It's also the most robust configuration compared to the others based on the lower standard deviations per evaluation metric.

The additional availability of the image crops reduces the performance and makes the training less stable demonstrated by a higher standard deviation. Similar observations can be made when comparing the performance of 2D pose heatmaps and body part heatmaps individually, with and without the availability of image crops. In both cases, the performance is lower when image crops are available. This finding is valuable and shows the difficulty of training models on small datasets using raw inputs such as the cropped image in our case. Processed input modalities consistently outperform the results of the raw image modality. Overall, the differences between the models are modest, especially in terms of accuracy. Higher scores for the balanced accuracy and F1 require that the minority class (in our case, the contact class) is better predicted.

Influence of the amount of data. We perform a systematic evaluation of the effect of using less training data when training the model on the YOUth PCI dataset from scratch. Again, we aggregate results from 10 runs for each configuration. Results in Fig. 6 show the means and standard deviations. As expected, using less training data yields gradually worse performance. This is because more training data allows the model to learn more patterns and relationships within the dataset, resulting in better generalization to unseen data. However, it is surprising to see that the performance difference between models trained with 100% and 10% of the training data is only around 8% across all metrics. With at least 50% of the training data, the decrease in performance is limited to 2%. This suggests that the

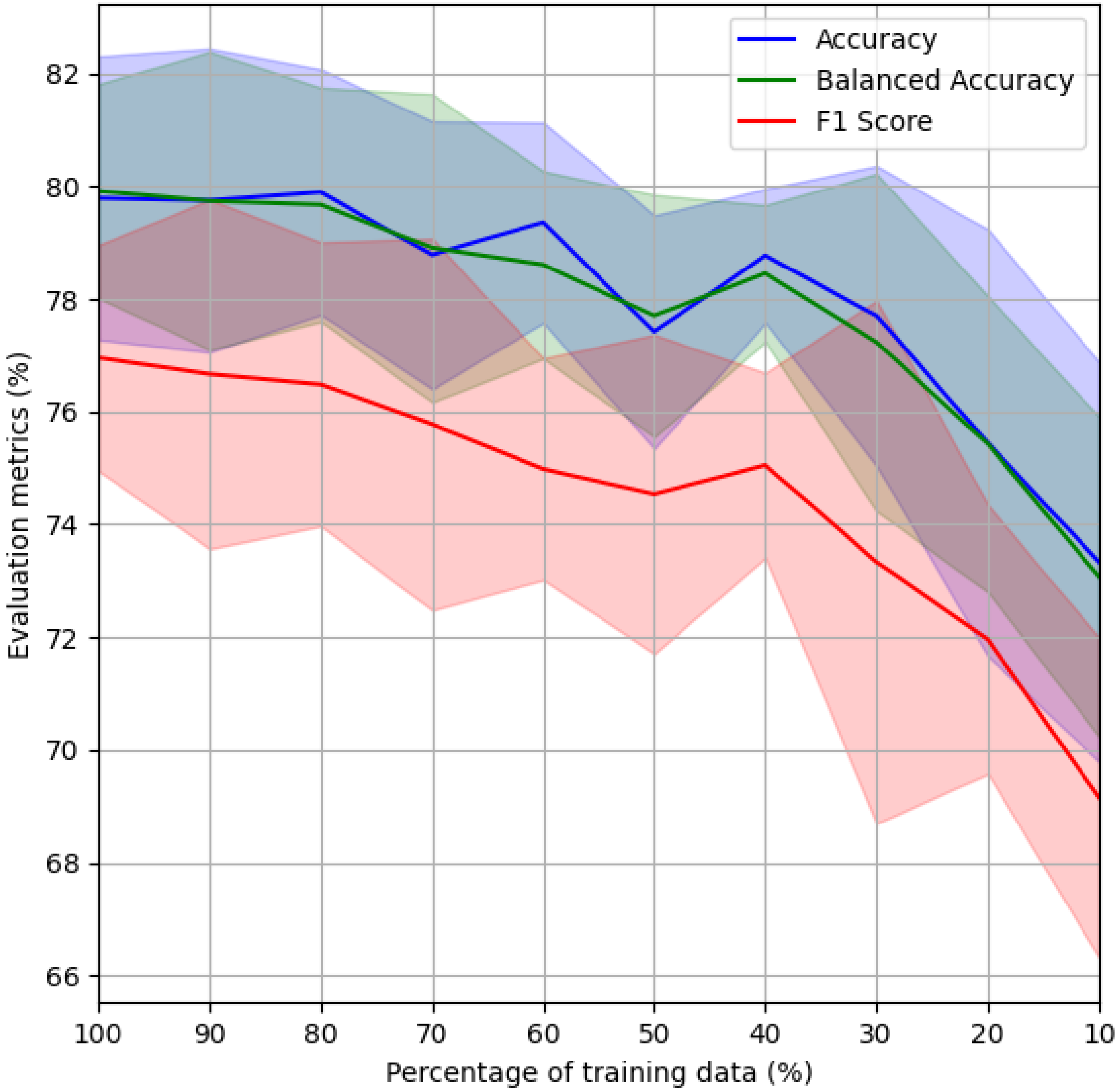

Fig. 6. Effects of using decreasing amounts of training data on the performance of the model. Graphs show means and standard deviations of 10 runs for each setting.

model may be robust to smaller training data sizes or that the data may contain redundant information. Sampling every 5 seconds might have resulted in similar frames.

Influence of proximity. One possible reason for the modest increase in performance when using more data is the existence of inputs that are particularly easy to classify. In this case, no substantial amount of data is required for training. We expect that proximity might be an influential variable in this respect. When the two interactants in our videos are far apart, there is likely no contact. However, when their joints are close together in the image, it becomes more difficult to predict whether they are in contact. In particular, the lost depth information might render distances between body parts small in 2D, whereas they are sizeable in 3D.

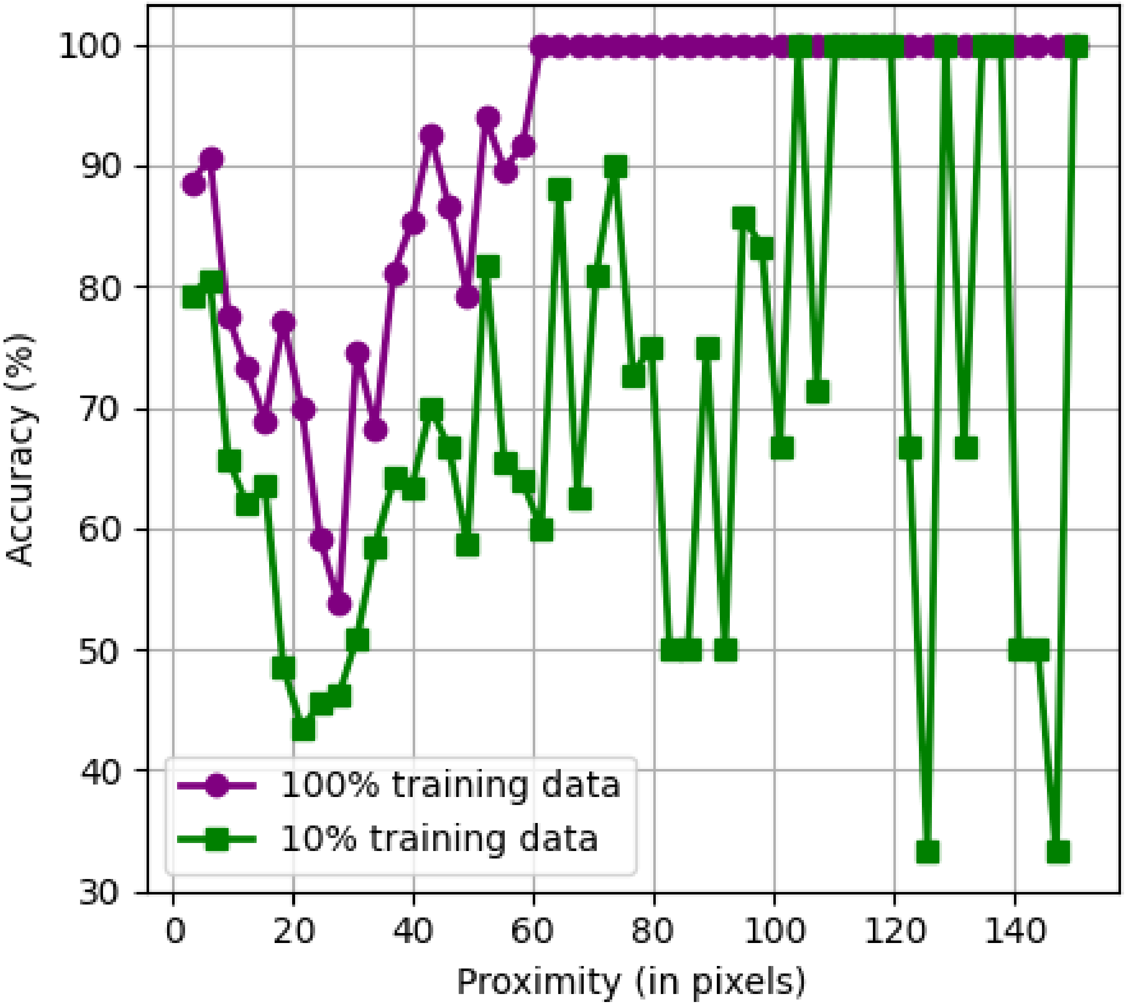

Fig. 7. Accuracy per proximity bin for two models trained with 100% and 10% of the training data, respectively.

From the proximity levels in Fig. 10, we can see different types of poses between parents and infants. The figure also shows depth ordering problems for some of the frames since the poses are detected in 2D. To investigate the role of proximity in our results, we have calculated the accuracy over each of a range of 2D proximity values, calculated as in Section 3 by dividing the range of distances into 50 bins and visualized in Fig. 4. A visualization of the accuracy scores per proximity bin appears in Fig. 7, for two models trained from scratch on 100% and 10% of the YOUth PCI data, respectively. The accuracy of the model trained with 10% training data is consistently lower than the model that uses all available data.

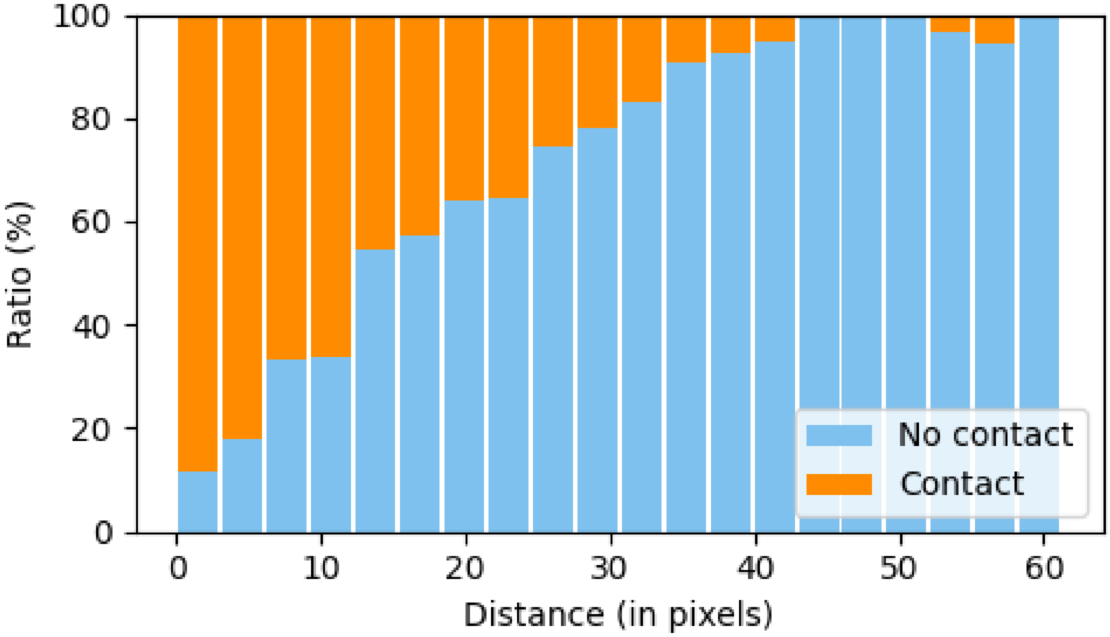

Fig. 8. Contact and no contact class counts per proximity bin in the test set. Higher than 60-pixel distance examples only include no contact class and are omitted from the plot.

We obtain perfect accuracy for the 100% data model when the distance between the two people is at least 60 pixels. These cases are less likely to contain contact. The model that used only 10% of the data occasionally has a low accuracy for these proximity values. We assume that this is caused by severe undersampling, as there are not too many distant cases (see Fig. 4).

We also observe high scores when the proximity is very low. These cases correspond to close proximity, in which physical contact is likely. Between these extremes, we observe lower accuracy scores for both the model trained on 100% of the data and the model that was trained with only 10% of this data. These cases, which may involve occlusions between the parent and the infant, can be considered more difficult to predict.

In Fig. 8, we inspect proximity by visualizing class distribution per proximity bin. Between 80-90% of the close proximity frames contain physical contact. In contrast, the samples of parent-infant pairs with 10 to 30-pixel distances show similar amounts of contact and no contact classes, which explains the dip in the performance of the 100% data model around those bins in Fig. 7.

Video-level classification. To characterize the interaction between a parent and a child, one might look at the percentage of contact over an entire video. In this case, since we aggregate frame-level classifications, confusion between the contact and no contact class might be partly mitigated if there is no systematic bias.

In Fig. 9, the predicted and true percentage of contact frames are shown for each test video. Most videos are relatively accurately predicted. The mean difference between the ratio of contact predictions and true labels across the test set is 10.74% with a standard deviation of 10.25%. There are only 3 videos out of 20, which have a difference of more than 20%. The correlation between predicted and true contact rates is 0.89 (Pearson's r: $r(18) = .89, p < 0.01$).

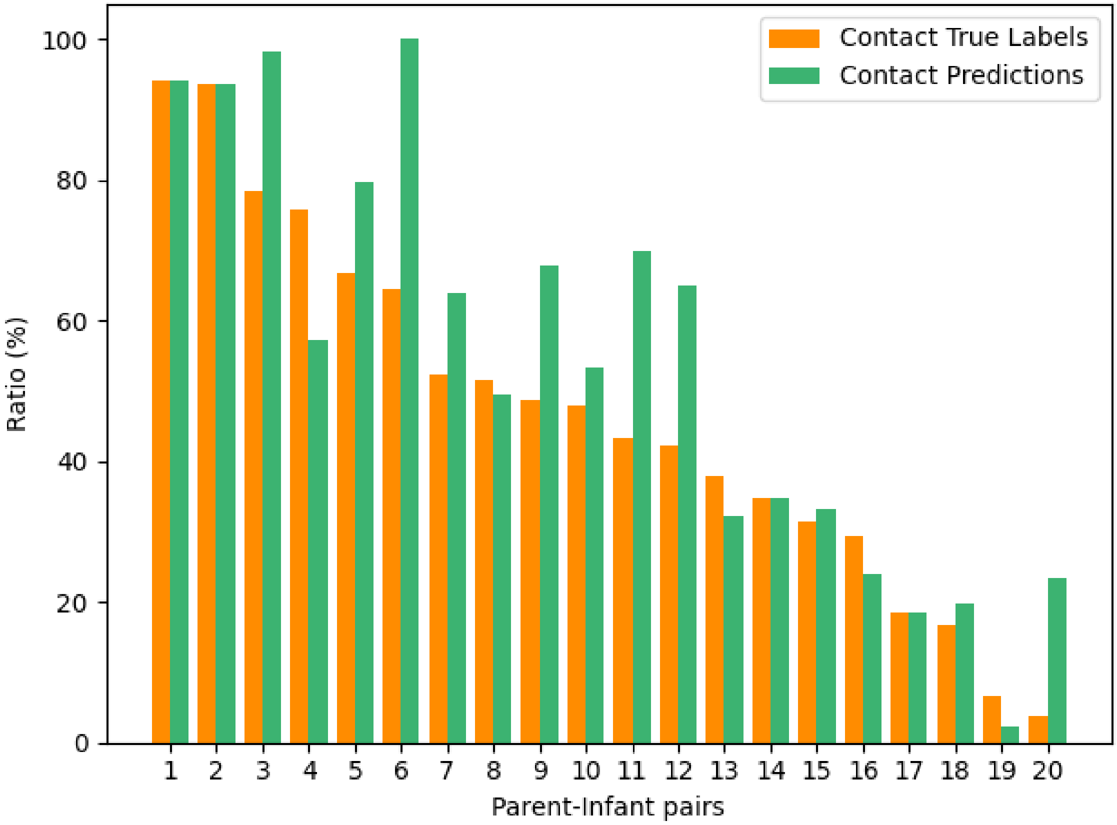

Fig. 9. Contact prediction and true counts per parent-infant pair. Test videos are numbered to allow referencing.

For several videos, *e.g.*, 3, 6, and 11, a much higher number of contact frames is predicted. These videos contain frames in which the interactants are usually occluding each other or in very close proximity without making contact. The first two rows of Fig. 11 show a variety of such false predictions. In the first row, second column, only a few landmarks of the infant are detected behind the parent, and they are further apart from each other in the depth ordering. Similarly, in the frame in the second row, the fourth column, the infant is occluding the parent. Fig. 11 shows a selection of frames where none of the 7 best models (selected from the input modality experiment, Table 2) were able to make a correct classification.

For videos 4 and 19, we see the opposite effect: a lower number of predicted frames compared to the actual number of contact frames. Visual inspection of the videos reveals that the frames include close proximity interaction but the body parts of the interactants are not significantly overlapping each other. The last two rows in Fig 11 show such examples where contact could not be predicted by the models. These frames look similar to one mode of error in the first two rows where there are false contact detections.

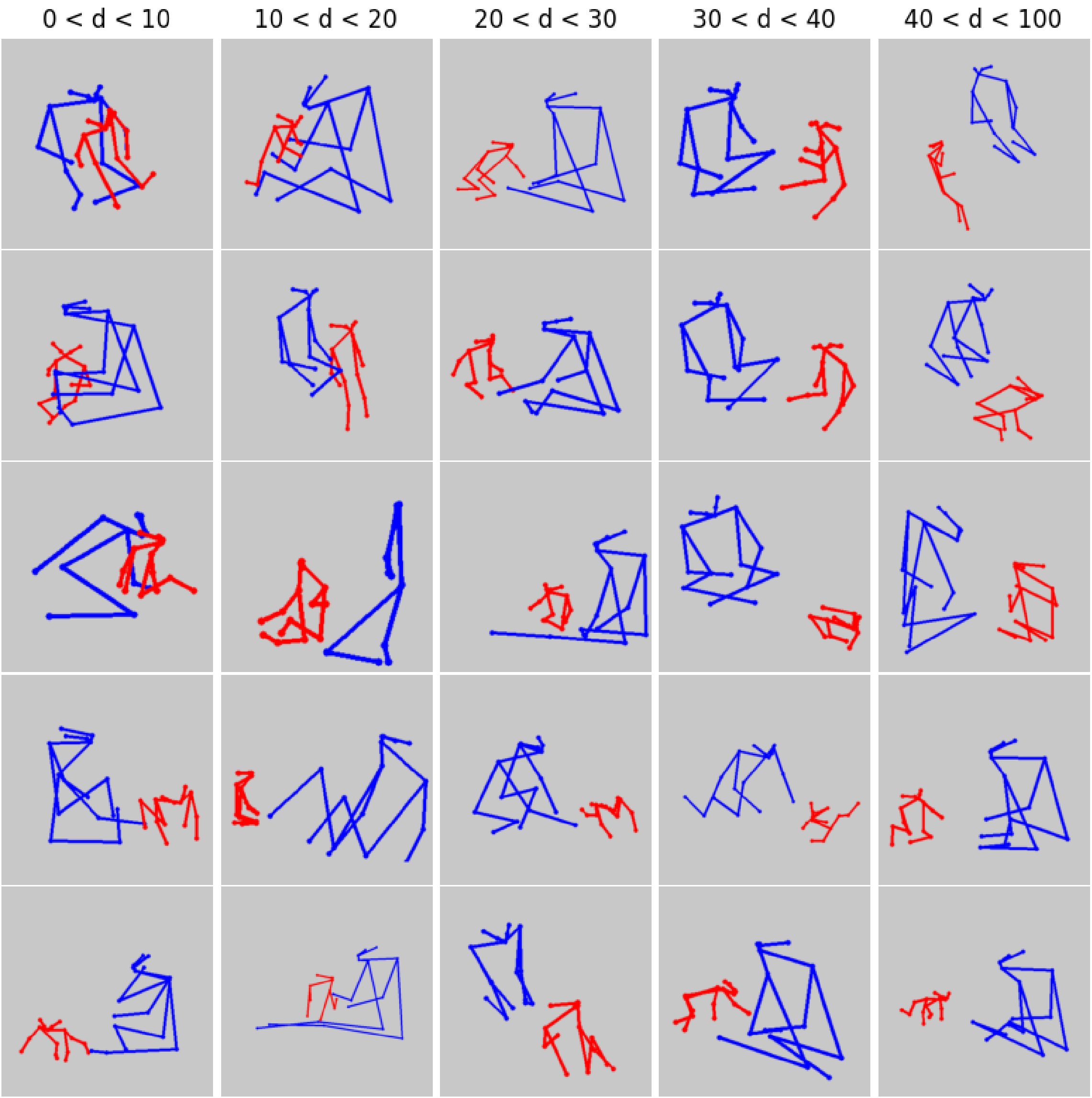

Fig. 10. Example poses for different proximity ranges. 'd' is the minimum pixel distance between 2D pose landmarks of the two interactants. Red: infant, blue: parent.

6 DISCUSSION

Compared to previous work on automated contact detection from images of interactions between adults [11], there is a fundamental difference in the nature of interaction between an infant and their parent compared to the interaction between two adults. In this paper, we have examined the feasibility of contact detection in free play interactions between an infant and a parent. We can detect their physical contact with reasonable accuracy using convolutional neural

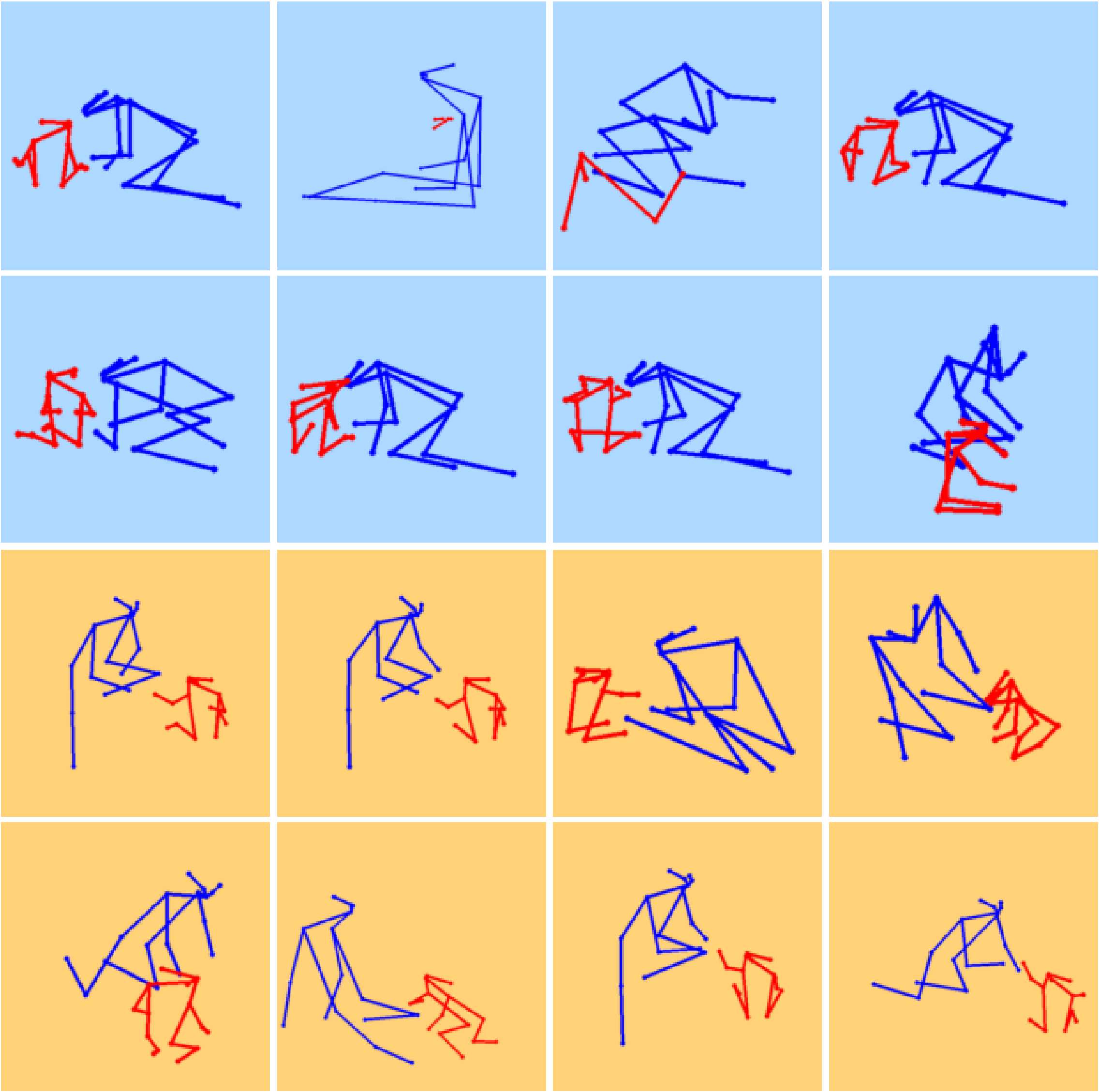

Fig. 11. Failure cases. First two rows show false detection of contact, last two rows show missing contact detections.

networks and processed input modalities such as 2D pose heatmaps and body part heatmaps. The performance achieved even with smaller amounts of training data (10%) demonstrates the potential for applying these models in real-world applications. Aggregation at the interaction-level showed good correlation between predicted and true contact rates.

Challenges. Several challenges remain in detecting contact in parent-child interactions. First, performance dips were observed when the parent and child were somewhat close. These situations include occlusions (Fig. 10, columns 2 and 3), which make it harder for the model to distinguish between contact and non-contact. One important cause is the use of a single view. During classification, no information about distances in depth is readily available. Similarly, by relying on a single view for the annotation, many challenging frames were labeled as ambiguous. Second, the differences between datasets, such as FlickrCI3D and YOUth PCI, highlight the need to consider domain-specific factors when training models for parent-infant contact detection. Models pre-trained on one dataset may not perform optimally when applied to another dataset.

Solutions. Several strategies can be employed to address the challenges in detecting physical contact in parent-child interactions. To handle difficult cases involving bodily occlusions, more advanced techniques, such as incorporating temporal information, using multiple different views, or reconstructing parent-infant interactions in 3D could be employed. Exploring additional input modalities such as eye gaze detection may also enhance the model's performance in detecting contact by informing models about the interaction dynamics. To account for the differences between datasets and domain-specific factors, more diverse and representative datasets can be developed. To tackle the time-consuming and costly annotation process, researchers could explore automated or semi-automated annotation methods, leveraging existing annotations or knowledge from related domains, ultimately resulting in larger and more comprehensive datasets for training. By publicly releasing our contact annotations and code, we hope to encourage more researchers to focus on this largely overlooked communication channel.

7 CONCLUSION

We have presented an approach to detect physical parent-infant contact in videos using CNNs. We examined the performance on the YOUth Parent-Child Interaction (PCI) dataset. As the first to target contact detection in parent-infant interaction in a free play setting, our findings provide valuable insights into the prospects and challenges and highlight avenues for improvement.

Our experiments highlight the importance of using processed input modalities such as 2D pose heatmaps and body part heatmaps. The results also emphasize the need to consider domain-specific factors when training models for parent-infant contact detection. An analysis of the proximity between the interactants revealed that medium-close interactions are the most challenging. Lower proximity or higher proximity levels are easier to predict since contact and no contact distributions are more distinct. Future work should be aimed at improving distinction for these arrangements. Also, we should address distinguishing between intentional and unintentional contact, potentially using temporal information or additional modalities such as eye gaze.

Our video-level aggregation experiment further demonstrates the feasibility of automatically classifying contact levels over an entire interaction. Such a measure can readily be used as a measure for the quality of parent-child interaction. Future research should investigate potential correlations and causality between physical contact and indicators of cognitive, and social development.

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] Cigdem Beyan, Alessandro Vinciarelli, and Alessio Del Bue. 2022. Face-to-Face Co-Located Human-Human Social Interaction Analysis using Nonverbal Cues: A Survey. arXiv:arXiv:2207.10574

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

[4] 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.

[5] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. 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.

[6] 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.

[7] 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.

[8] 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.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*. IEEE, Miami, FL, USA, 248–255.

[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.

[13] 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.

[14] Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2021. Multiperson extreme motion prediction with cross-interaction attention.

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

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

[23] 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.

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

[25] Juliana F. Serra, Isabel C. Lisboa, Adriana Sampaio, and Alfredo F. Pereira. 2023. Observational measures of caregiver's touch behavior in infancy: A systematic review. *Neuroscience & Biobehavioral Reviews* 150 (2023), 105160. https://doi.org/10.1016/j.neubiorev.2023.105160

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

[27] 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.

[28] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, and Otmar Hilliges. 2023. Hi4D: 4D Instance Segmentation of Close Human Interaction.

[29] 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.

[30] 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.

Received 8 May 2023