Automated Detection of Joint Attention and Mutual Gaze in Free Play Parent-Child Interactions

Peitong Li
p.li1@uu.nl
Utrecht University
Utrecht, NL

Hui Lu
h.lu1@uu.nl
Utrecht University
Utrecht, NL

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

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

ABSTRACT

Observing a child's interaction with their parents can provide us with important information about the child's cognitive development. Nonverbal cues such as joint attention and mutual gaze can indicate a child's engagement, and have diagnostic value. Since manual coding of gaze events during child-parent interactions is time-consuming and error-prone, there is a need for automatic assessment tools, capable of working with camera recordings without specialized eye-tracking equipment. There are few studies in this setting, and accessing naturalistic parent-child videos is difficult. In this paper, we investigate the feasibility of detecting joint attention and mutual gaze in videos. We test approach on challenging data of a child and a parent engaged in free play. By combining multiple off-the-shelf approaches, we manage to create a system that does not require much labeling and is flexible to use for view-independent interaction analysis.[1]

CCS CONCEPTS

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

KEYWORDS

joint attention, mutual gaze, parent-child interaction, cognitive development

ACM Reference Format:
Peitong Li, Hui Lu, Ronald Poppe, and Albert Ali Salah. 2023. Automated Detection of Joint Attention and Mutual Gaze in Free Play Parent-Child Interactions. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23 Companion), October 9–13, 2023, Paris, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3610661.3616234

[1]This is the uncorrected author proof, see ACM Copyright notice for full copyright information of the paper and the DOI of the published version.

ICMI '23 Companion, October 9–13, 2023, Paris, France

ACM ISBN 979-8-4007-0321-8/23/10…$15.00
https://doi.org/10.1145/3610661.3616234

1 INTRODUCTION

The physical interaction between children and their parents has a significant impact on a child's development and well-being [9]. Positive interactions between parents and their children can improve children's self-esteem and social skills [11, 27], as well as children's academic achievement and cognitive development [22]. Manual assessment of such interactions by experts can provide in-depth insights and indicators about the child's development. However, it requires significant time and human resources, and subjective biases of researchers may affect the results. There is a need for accurate and scalable approaches to support the experts.

There are a number of cues that are relevant for such analysis, including mutual gaze, joint attention, touching and body positioning, affective cues such as vocalizations and facial expressions. In this paper, we propose an automatic approach to determine mutual gaze and joint attention during child-parent interactions (see Figure 1). Joint attention refers to the shared focus of attention between two individuals on an object or event, and has been shown to play a crucial role in children's language development and social skills [35]. Mutual gaze refers to the visual exchange between two individuals, where they both look at each other's faces. It has been linked to a range of positive outcomes in children such as increased social competence and empathy [15]. Dynamics of mutual gaze and joint attention can reflect the quality of the relationship between parents and children, and provide diagnostics on developmental problems, such as Autism Spectrum Disorders [17].

Figure 1: Joint attention (left) and mutual gaze (right)

Dedicated eye tracking systems are typically employed for attention studies, but such equipment may not be available in all settings or in legacy datasets. Furthermore, wearable eye trackers reduce the ecological validity of the interactions. In this paper, we therefore focus on freely recorded interaction videos, as this is an unobtrusive way of observation. We propose a novel approach that works in free parent-child play settings to estimate mutual gaze and joint attention. By leveraging off-the-shelf machine learning algorithms, our approach enables large-scale studies that are more accessible and cost-effective than traditional eye-tracking methods[2].

The remainder of the paper is organized as follows. We discuss related work on detection of nonverbal cues in parent-child interactions, as well as on gaze attention estimation in Section 2. Details of our method are introduced in Section 3. In Section 4, the video data and annotations are described. Section 5 contains our experimental results and discussion. We conclude in Section 6.

2 RELATED WORK

Researchers have used both manual and automated methods to analyze parent-child interactions. Manual methods include coding schemes, observational techniques, and self-report measures. Coding schemes are based on various behavioral categories, which can include warmth, control, negativity, positive affect, negative affect, and communication patterns [1, 7, 18]. Observational techniques systematically observe parent-child interactions and document different aspects of the interaction, such as gaze direction, facial expressions, body language, and tone of voice. They are often used to capture the emotional tone of parent-child interactions and how parents and children interact with each other under various circumstances [10, 21]. Self-report measures involve collecting data directly from parents and children about their perceptions of their relationships and interactions. For example, questionnaires have been used to assess parental warmth or control [33].

2.1 Nonverbal cues for parent-child interaction

Automatic coding methods are used to analyze both verbal and nonverbal components of human interactions, with the latter being especially important during infancy and toddlerhood, as they play a critical role in facilitating parent-child interactions during early stages of development [8, 13, 26]. Vocal behavior, face expression, body activity, proxemics, physical appearance, eye gaze, and visual focus of attention have been widely investigated.

Facial expressions have been analyzed using computer vision, but infant faces are different compared to adults, and methods trained with adults perform poorly on infant facial expression recognition tasks [29]. Similarly, pose detection and body activity analysis needs to be adopted for infants. Body activity refers to physical movements and gestures. Body gestures and head movements can help understand the interaction styles between parents and children, and thus provide insights into their relationship dynamics [2]. Proxemics refers to the physical distance between individuals during the interaction. Avril et al. [3] used skeletal tracking to monitor the proximity between an interacting parent and child seated at a table. Physical appearance refers to the observable visual traits of an individual that are present during social interactions, such as height, weight, body shape, skin/hair color, clothing style, and the use of makeup or accessories. The appearance of individuals and contextual objects can be used to distinguish between various types of social relationships Liu et al. [20]. Vocal behavior refers to all aspects of speech, such as the use of vocalizations like fillers, laughter, and sobbing, as well as pauses and turn-taking in conversation. Nguyen et al. [28] used Bayesian meta-analysis method to analyze the development of turn-taking in adult-child vocal interactions. As it can be seen from this brief overview, there are plenty of non-verbal cues for interaction analysis. In this paper, we focus on two gaze-related indicators.

[2]Our code is open source: https://github.com/Chelseapt/Joint-Attention-and-Mutual-Gaze-in-Free-Play.

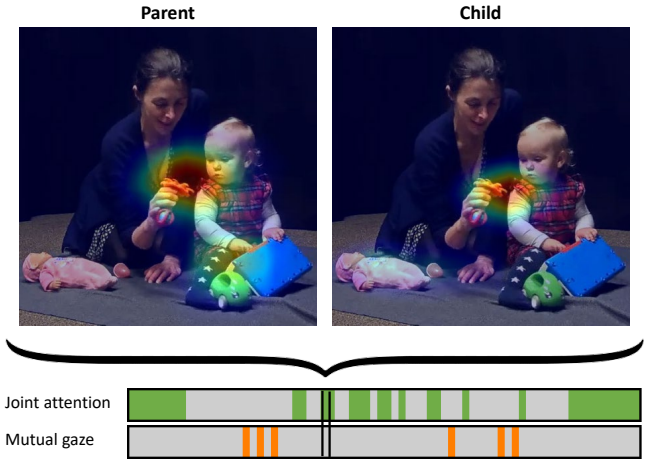

Figure 2: Conceptual overview of our approach. Based on the estimated visual focus of attention of parent and child, we estimate their mutual gaze and joint attention over the entire interaction. Stills are from an interaction not used in this study. Permission for reproduction granted.

2.2 Gaze attention estimation

Eye gaze and visual focus of attention (VFOA) refer to the direction of individuals' gaze during the interaction. Analyzing gaze behavior can help understand social dynamics in parent-child interactions and reveal individual differences [12].

Head pose estimation is important for gaze estimation in free settings, but we will not focus on this task here. Liu et al. [20] provides a comprehensive overview of different head pose estimation techniques, including their advantages and limitations. Piccardi et al. [32] previously analysed an infant's gaze patterns for the focus of attention, using a wearable camera. This decouples the gaze estimation from head pose estimation, but wearable cameras are difficult to use, and can cause ecological validity issues.

Zhang et al. [39] provides a list of recent databases for gaze estimation, but these are all focused on frontal face-based estimation, which is a common scenario in human-computer interaction, where a person is facing a screen. An example system that works in such a setting is EyeShopper [5], which is designed to track shoppers' gaze in surveillance systems. Kodama et al. [14] employed two

non-overlapping cameras fixed on opposite sides of an audience to identify the target of attention for multiple individuals.

A technique for estimating and tracking visual focus of attention during multi-party social interactions is proposed in Massé et al. [25]. The method uses head movements and Bayesian state-space modeling to infer VFOA and gaze, suggesting potential applications in situations where conventional eye-tracking techniques may not be practical. However, our experimental setting presents a more challenging task in which individuals can move around freely. Their eyes are also not always visible, so we cannot rely on precise estimation of the gaze target. As a potential solution, Chong et al. [6] recently presented an approach for tracking individuals' VFOA (in images) without explicitly relying on eye gaze, which is adopted here for estimating attention distribution. The method works even when the target is beyond the frame boundaries, and produces a heat map of VFOA estimations.

For detecting and localizing joint attention, Sumer et al. [34] proposed a novel method called attention flow, which does not require the use of face detectors or head pose/gaze estimators and is based solely on the raw input image. Their data comes from TV programs, which may not represent real-world social scenes.

Yücel et al. [37] analyzed the focus of attention of adults from frontal videos in a limited interaction setting, where joint attention is established with a robot on an object of interest. Because the video resolution in their setting was too low to track the eyes accurately -which is a common problem- they have used head pose estimation to interpolate the gaze, using a bottom-up computational model to find salient objects in the estimated gaze cone. If the object of focus can be determined for both participants of a dyadic interaction, joint attention can be established. Kwon et al. [16] proposed a method for inferring a common focus of attention based on visual (object and face detection) and linguistic clues. While our work shares similarities with this approach, our application domain is more challenging as people could move around freely. Moreover, we limit the analysis to only visual cues, as speech capibilities of the infants in our study are just developing.

Mutual gaze is of particular interest when studying interactions. In an early work on social interactions, Ba and Odobez [4] detected mutual gaze based on head pose in meeting scenarios. More recent approaches with similar goals use deep neural networks that take detected and cropped head images as input [23, 24]. Zhang et al. [38] also considered a meeting scenario and used OpenFace and gaze inputs to detect whom of the other participants was looked at. When working with video data, the temporal duration of the interaction and spatial localization of the relevant individuals is a factor that can be helpful. Palmero et al. [31] focused on identifying mutual gaze occurrences in face-to-face dyadic interactions using two calibrated monocular RGB cameras, but, similarly to [4, 38], cameras are placed in front of each participant, which is a more restrictive setting than our free play application scenario.

3 METHODOLOGY

3.1 Overview of our approach

Our approach is schematically visualized in Figure 3. Based on a video of a scene in which the interaction takes place, we detect the heads of the parent and child, and the objects in the scene. Heat maps of likely 2D gaze locations in the image are obtained from a video frame and corresponding head detection of both the parent and the child. Based on the heat maps, we calculate likelihood scores for the detected head and object regions. By combining these scores for the parent and child, we produce a final binary classification for both joint attention and mutual gaze. The only step that requires training in our current setup is object detection; off-the-shelf models are used for other modules.

3.2 Head tracking

We perform head detection on both parents and children, followed by head tracking. We employ LAEO-Net [23], which uses a head detector trained as Single Shot Multi-box Detector (SSD, [19]). The head detector was designed to detect the entirety of the head, not just the face, which is more robust with cases where the face is turned away from the camera, and has fewer missed detections compared to face detection.

We use DeepSort [36] to track and consistently link the head detections to the parent or the child. DeepSort associates detections with existing tracks through a combination of position and appearance information. We use the trained model that was provided by the authors for a pedestrian tracking scenario, without making adjustments to the appearance term. As a result, DeepSort occasionally adds new tracks instead of prolonging existing ones. This typically happens when changes in the appearance of the head region are significant during quick changes in the head orientation, or when head occlusions occur. To perform detailed evaluations, we provide manual annotation for the tracks as coming from the parent or the child. This tracking can be automated relatively easily, for example by relying on the substantial significant age difference between our subjects or differences in facial identity.

Finally, we interpolate the missing head detections. Since we operate within an free-play setting, occlusions and turned heads will lead to missed head detections. Especially for children, more erratic head movements occasionally lead to detection or tracking failures. The vast majority of such gaps are short in duration, and the location differences before and after the gap are modest. We linearly interpolate gaps less than two seconds (corresponding to 50 or 60 frames in different videos) and with a Euclidean distance between the head detection centers of less than 45 pixels (with frames of 960×540 pixels).

3.3 Object detection

In this study, we are interested in joint attention to any of the toys that are present in the scene. To this end, we employ object detection to locate these toys. We consider 12 object categories, summarized in Table 2. The top part of the table contains independent toys; the objects in the lower part of the table are parts of a shape box that can be independently manipulated. Typically, objects in the latter category are smaller and are less often visible as they are frequently inside the shape box. Since the toys are specific for our scene, we train object detectors for each class. We use YoloV5[3] as our convolutional neural network (CNN) object detection model. YoloV5 is widely used and it shows competitive performance without complex parameter tuning during the training process. Specifically, we

[3]Online available at: https://github.com/ultralytics/yolov5

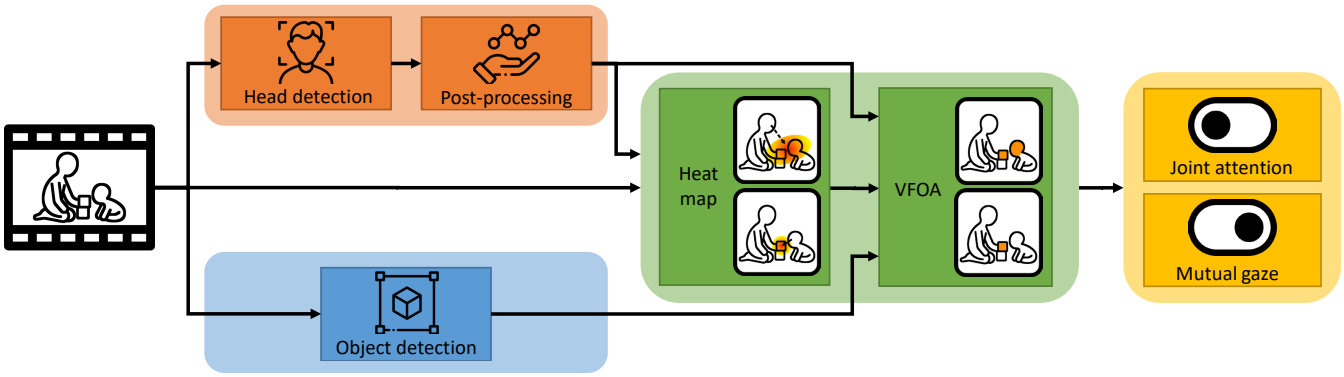

Figure 3: Schematic overview of our approach: heads and objects are detected first. Heat maps for 2D gaze attention are estimated from a frame and the head regions using [6]. We then calculate the visual focus of attention by considering the heat map values for each region of interest. Finally, we combine the outcomes of parent and child to classify whether there is mutual gaze or joint attention to an object.

use the default training parameters with SGD optimization with a weight decay of 0.01 and a momentum of 0.937.

3.4 Joint attention and mutual gaze estimation

As a first step to assess joint attention or mutual gaze between parent and child, we examine where either is looking at. We obtain heat maps of the gaze attention for the parent and the child independently, using the method presented in [6]. When provided with an image frame and the region of the head, the algorithm determines the likelihood that a 2D location in the frame is attended to. Since depth information is lost, particular attention is paid to salient regions, building on the assumption that we tend to attend to objects. At the core of this method is a spatio-temporal model based on a CNN, consisting of two main components: a head feature extractor and a scene feature extractor, respectively.

The head feature extractor extracts facial regions from video frames and generates feature vectors related to the head. To take advantage of the temporal continuity in head movement when analyzing videos, we use our interpolated head regions instead of the head detection in [6]. The scene feature extractor extracts scene regions from video frames and generates feature vectors related to the gaze. The outputs of these two components are fed into a multi-layer perceptron (MLP) to predict where a person is looking in each frame. The output of this process is a gaze likelihood heat map.

Since we are interested in the visual focus of attention, we combine the heat map with the regions of the objects and the other's head. To this end, we combine the heat map values within each region. We average the heat map values within the region to calculate the VFOA value: VFOA_{avg}. To determine whether a region containing a head or toy is attended to, we apply a threshold on the VFOA values. For VFOA_{avg}, we empirically determined a threshold of 80 on the normalized heat maps produced from [6].

By finally combining the VFOA of both parent and child, we obtain a binary indication of mutual gaze and joint attention. For mutual gaze, the parent should look at the child, and *vice versa*. Joint attention at a toy requires that both parent and child look at it, determined by the VFOA classification. For each object that we consider, we calculate the joint attention. We adopt a multiple-attention strategy. During manual annotation and automated detection, we allow parents or children to attend to several targets simultaneously. We choose this strategy since it proved to be difficult to confidently identify one target when other toys were close. For joint attention, our output is therefore a vector of binary indicators for each object.

Note that the processing of the parent's and child's VFOA to classify joint attention and mutual gaze is identical. Strictly speaking, we therefore do not need to know which head belongs to the parent and which to the child. It's only for reporting the performance separately here that we distinguish between the two.

4 DATA COLLECTION AND ANNOTATION

4.1 Video data description

For analyses, we use parent-child interaction videos from the YOUth Cohort study [30], which are freely available to researchers after an ethical approval process[4]. In each video, a parent and a child of approximately 10 months old play together with toys of different sizes in the playground. Parent occasionally introduce a new toy, but the interactions are relatively unstructured, as the use of specific toys is not required, and show show significant variation in play.

Recordings in the YOUth Cohort are currently still ongoing. Each interaction is recorded from four cameras, but we restrict our analyses to a single overhead view, see Figure 2. We selected 20 interactions from a pool of videos that have been recorded with the highest spatial resolution (i.e., 960×540 pixels). The total length of 20 videos is around 250 minutes. In 18 videos, the parent is the mother. In the remaining 2 videos, the father plays with the child.

We have temporally segmented the videos to start when the experimenter left the scene, and stopped our analyses when the experimenter returned. The part of the interaction that we analyze is approximately 12–13 minutes per video. For the manual annotations, we select a frame every 10 seconds. In total, our analyses cover 1522 frames of which 1486 contain two heads.

[4]More information at: https://www.uu.nl/en/research/youth-cohort-study

Table 1: Number of head annotations and detections with either child or parent, or both. Agreement is the number of frames in which manual annotation and interpolated detection overlap.

Type	Manual	Detected	Agreement
Both	1486	1026	990
Child only	2	31	2
Parent only	20	413	3

4.2 Annotation and detection of heads

We provide detailed manual annotations to compare automatic and manual estimation approaches, as well as to assess accuracy. For the manual annotation of the heads of parent and child, we used the DarkLabel 2.4 annotation tool[5], which allows for drawing bounding boxes for a number of pre-specified classes. We distinguished between the head of the parent and the head of the child, to facilitate subsequent analyses for each. A single coder annotated all heads, and a second coder verified the annotations.

For the automated detection, we used [23] and interpolated missing frames. Interpolation increased parents' head frames by 2.57% and children's head frames by 15.14%. The total number of manually annotated and automatically detected frames are given in Table 1. The automated detection misses a number of heads and occasionally generates false positives. From the table, it can be observed that the majority of the missed head detections are from the child.

Table 2: Number of object annotations and detections. Agreement is the number of frames in which manual annotation and interpolated detection overlap. Percentage is relative to the manual annotations.

Type	Manual	Detected	Agreement
Car	1182	1154	1124 (95.09%)
Doll	1173	1034	996 (84.91%)
Switch box	1126	1080	1053 (93.52%)
Flower	988	841	770 (77.94%)
Book	507	438	363 (71.60%)
Baby bottle	203	59	37 (18.23%)
Shape box	700	646	618 (88.29%)
Green star	234	167	143 (61.11%)
Yellow cylinder	246	139	101 (41.06%)
Blue cube	188	110	78 (41.49%)
Red triangle	562	228	188 (33.45%)
Shape box lid	583	248	126 (21.61%)

4.3 Annotation and detection of objects

We annotated all the toys in the 1522 frames used in this study. A single coder made the annotations, which were checked by a second coder. To train the YoloV5 model for toy detection, we annotated a non-overlapping set of 10 videos in the same interaction setting and with the same resolution. Our training set consisted of 3K training and 1.5K test images.

A summary of the manually labeled and automatically detected toys appears in Table 2. We consider a manual and detected region to be in agreement if they have the same label and their Intersection over Union (IoU) overlap is at least 0.5.

The agreement is generally high for the larger objects but is significantly lower for smaller shapes in the shape box and for the baby bottle. This effect is partly due to the difficulty in detecting partly occluded small objects, and partly because inaccurate localization has a larger effect on the IoU for smaller objects.

4.4 Annotation of the visual focus of attention

Two coders independently annotated mutual gaze and joint attention to specific toys. In each frame, multiple VFOA annotations per person could be made, for example when toys were in close proximity or when a person attended to a larger area. Because the VFOA annotations are based on the head and object detections, the number of targets is the same for both coders. Moreover, a VFOA annotation per target is binary. We can therefore calculate the agreement between the two coders as the percentage of matching VFOA labels per frame, averaged over all frames. For joint attention and mutual gaze, the inter-annotator agreement (Cohen's kappa) are 82.93% and 73.51%, respectively.

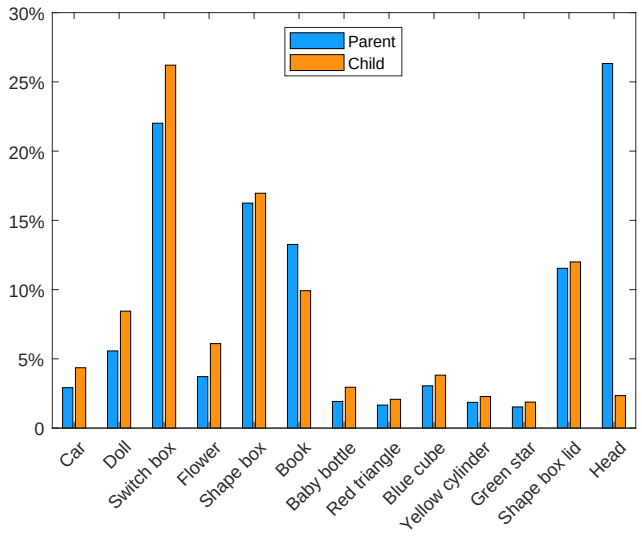

Figure 4: Percentage of time attending to each object and head. Percentages from Coder 1 for parent (blue) and child (orange).

Since the annotations of the two coders are largely similar, we focus on Coder 1 in this section. The percentage of frames with mutual gaze is 1.46%. This is a relatively low number but it can be understood by observing the percentage of time that parent and child spend looking at each other. In Figure 4, it becomes clear that parents look at their child in 26.33% of the time, whereas children only look at their parents 2.35% of the time. From Figure 4, no large differences in the VFOA for different toys are observed.

[5]Online available at: https://github.com/darkpgmr/DarkLabel

Table 3: Joint attention and mutual gaze evaluation results (%) for combinations of labeled/detected heads/objects and baselines. C1/C2: Coders 1 and 2. GT: ground truth, manually labeled. D: automatically detected.

		Joint attention						
		C1			C2			Average
Heads	**Objects**	Recall	Precision	F1 score	Recall	Precision	F1 score	F1 score
GT	GT	63.52	38.92	46.57	63.65	35.84	44.13	45.35
Baseline_GT	50/50	26.27	14.52	16.77	26.86	13.44	15.92	16.35
Baseline_GT	Prior	0.72	5.95	1.24	0.79	5.91	1.36	1.30
GT	D	45.11	44.15	41.99	44.17	40.79	39.74	40.87
D	GT	46.18	38.41	40.27	47.46	36.24	39.27	39.77
D	D	30.42	40.31	33.25	31.19	39.41	33.24	33.25
Baseline_D	50/50	10.32	13.06	10.20	10.52	12.32	10.11	10.16
Baseline_D	Prior	0.29	4.80	0.74	0.33	4.73	0.76	0.75
		Mutual gaze						
		C1			C2			Average
Heads		Recall	Precision	F1 score	Recall	Precision	F1 score	F1 score
GT		25.00	4.00	6.90	11.54	2.40	3.97	5.44
Baseline_GT	50/50	30.00	1.68	3.18	34.62	2.52	4.70	3.94
D		20.00	4.00	6.67	11.54	3.00	4.76	5.72
Baseline_D	50/50	20.00	1.56	2.89	15.38	1.56	2.83	2.86

5 EXPERIMENT AND RESULTS

In this section, we discuss our baselines and metrics, followed by the results for joint attention (Section 5.1) and mutual gaze (Section 5.2). A discussion appears in Section 5.3 and we present qualitative results (Section 5.4) and an analysis at the level of the entire interaction (Section 5.5), to further understand the potential of our approach.

Baselines. We are predicting the visual focus of attention for each object and head, from both the child and the parent perspective. This corresponds to a series of binary decisions. We conducted two types of baseline experiments for comparison. The first was based on a 50% probability of visual focus of attention to a given target (*50/50*). Since there are many potential targets, this naive baseline will be a significant over-representation of the actual amount of VFOA. To this end, we use a second baseline using a prior (*Prior*), incorporating prior knowledge about the proportion of different objects and heads viewed from the parent's and the child's perspectives, respectively.

We divided each baseline experiment into two groups based on the different input data: manually provided ground truth annotations (GT) and automated detections (D). For the manual annotations (GT), more heads and objects are available. Consequently, we expect higher recall rates compared to the baseline based on automated detections (D).

It's important to note that the results from the prior baseline experiments were quite poor due to low probabilities of actial VFOA. For mutual gaze, the results of the prior baseline becomes zero so we don't report it. Therefore, in subsequent comparisons, we only compare our results with the random baseline. For completeness, we report the evaluation metrics to both coders individually, as well as the average over both.

Metrics. We utilized recall, precision, and F1 scores as our evaluation metrics. For joint attention, we calculated these metrics individually for each of the 12 objects, obtained from a coder and from automatically processing with our approach. Then, we averaged the results across all 12 objects to obtain an overall measure of joint attention performance For each frame, we only consider the objects that were actually annotated, since the set of visible objects is possibly different. Moreover, different numbers of objects can be attended to by each person. F1 scores proved to be effective in handling the issue of sample imbalance, providing a comprehensive objective measure of performance. Regarding mutual gaze, we had a single binary output. Therefore, by iterating through all frames, we directly obtained the results for recall, precision, and F1 scores.

5.1 Joint attention

Results for joint attention are given in Table 3 (top part). Our approach was evaluated using manually annotated (GT) heads and objects, yielding an average F1 score for joint attention classification of 45.35%. The difference between the VFOA annotations of Coder 1 and Coder 2 is also minimal. Although the results are not as high as we might have wished, we outperform the baseline (Baseline_GT 50/50) at 16.35%. Compared to the baseline, we have achieved an improvement of 30% in F1 score, indicating that the adopted algorithm has a fairly effective performance in joint attention detection.

When changing the manual labels to automated detections, we observe that the results are slightly more affected by the detection of heads than objects. When head detections are used together with manually annotated objects, the score decreases to 39.77%. This decrease is predominantly caused by the lack of two detected heads. In Table 1, we already observed that many heads, especially of

the child, are not detected automatically. Consequently, we cannot make joint attention classifications. This is reflected is the much lower recall.

When we use manually labeled heads but automatically detected objects, the score compared to ground truth input decreases to 40.87%. This decrease is caused by the missing detections. Still, the decrease is not dramatic, and mainly because both parent and child are predominantly looking at the larger objects, see Figure 4. These objects are also better detected (see Table 2).

When applying our VFOA method using only automatically detected heads and objects, so using only automated detections, the score is lowest, at 33.15%. The missing heads and missing object detections both contribute to the missing VFOA detections, which consequently lower the detection of mutual attention.

5.2 Mutual gaze

We now proceed to the detection of mutual gaze, which is a binary classification task. Since our analysis is solely focused on heads, the presence of objects is not relevant. The F1 score when using manually annotated heads is 5.44%. This result might seem poor, which is largely due to the physical setting. When parent and child look at each other, typically they are facing each other. In this setting, at least one of the heads cannot be well observed. This has consequences for the estimated VFOA, as well as for the annotations. Upon inspection of the annotation, we found that both coders were conservative indicating VFOA when the face of the parent or child was not visible. Since this was true for both coders, the inter-annotator agreement for mutual gaze was good at 73.51%. At the same time, this consequently lowers the precision for our approach because VFOA at the other's face might have been classified for heads that are significantly turned away from the camera view. Still, when compared to the baseline, which has an F1 score of 3.94%, our experimental results show an improvement in accuracy. The small percentage of actual mutual gaze complicated drawing stronger conclusions, as is also witnessed by the significantly different recall values between the two coders.

Interestingly, when using detected heads as input, the score for mutual gaze actually increases. This is because in the case of using GT (ground truth) heads, there are more false positives compared to using D (detected) heads (heads not detected are recorded as no gaze attention), resulting in a slight overall accuracy increase of 0.28%, bringing it to 5.72%.

5.3 Discussion

Overall, we see that a deterioration in the quantity (more than quality) of the detections causes the largest drop in VFOA detection, for both joint attention and mutual gaze. If one of the heads is not detected, it is not possible to obtain VFOA estimations. Missing object detections have a smaller effect, especially since the detection performance for the most common objects is relatively good.

Several factors caused differences between the results of our automated approach and the manually coded joint attention and mutual gaze. First, heads were sometimes turned away significantly, complicating the estimation of the gaze heat map. Second, some objects were barely visible, for example, due to occlusions by either a person or other toys. From the manual annotation, we could make out the presence of the object. But, in certain cases, the object appeared too small to be reliably detected. This often happened for the shapes in the shape box. If these are in a person's hand, it's very difficult to detect them automatically. Third, we used only a single view of a cluttered play area. From the perspective of the camera, toys would typically be overlapping, thus complicating the distinction between them. Finally, while the agreement between the two coders was high, it was not perfect.

The results of joint attention and mutual gaze scores show a high variation between different videos, for which there are various reasons. First, each interaction has a different distribution of joint attention and mutual gaze, causing different baselines for both measures. Second, there is a significant difference in the seating arrangement. Both parent and child could be standing, crawling, or sitting. Not every situation allows for a good assessment of the visual focus of attention. Third, some videos showed more dynamic interactions, with more frequent switching of attention for different toys, the parent, and other targets in the environment. Especially when parent and child would observe an area with toys, rather than focusing on a specific toy, we observed lower inter-annotator agreement and lower agreement with the automatically estimated VFOA.

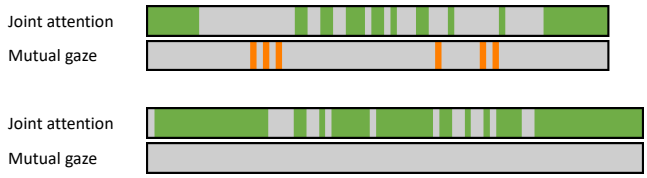

Figure 5: Joint attention (green) and mutual gaze (orange) distribution in two videos show the differences across sessions.

5.4 Qualitative analysis

We have analyzed our VFOA detection approach on a collection of 1522 frames from 20 videos. Our analyses have provided insights into the performance of different components. Here, we additionally demonstrate how our approach can be used to understand the nature of parent-child interactions. To this end, we have selected two out of the 20 videos. We summarize the joint attention and mutual gaze annotated by Coder 1 for these videos over time in Figure 5. Similar visualizations could be produced based on automatic detections of heads and objects. These visualizations would have a higher temporal resolution but would be less accurate due to the missing detections.

For the first video in Figure 5 (top), we notice multiple periods of joint attention. In the first period, the child is exploring the toys. In the second period, roughly halfway into the interaction, the parent and the child play with the shape box. In the final minutes of the interaction, the parent reads a soft book to the child. The joint attention is mainly on this book. There are also two periods in which there is mutual gaze. In the first period, the parent holds up the toy and proposes how they could play with it. The second period is marked by the parent explaining that she will read a book.

In the second video in Figure 5 (bottom), we didn't observe any mutual gaze. The child is predominantly focused on the toys.

Similarly, the parent follows the actions of the child. We see roughly the same periods of increased joint attention as in the first video. These correspond to the phases of initial exploration, playing with the shape box, and reading a book.

Here, we didn't distinguish between the joint attention for the various objects. Moreover, we didn't look at the timing of the individual visual focus of attention. We expect that a more fine-grained temporal analysis could reveal patterns of exploration. Moreover, we expect such analyses will allow for the investigation of leading and following, for example when a parent shows how a toy can be used.

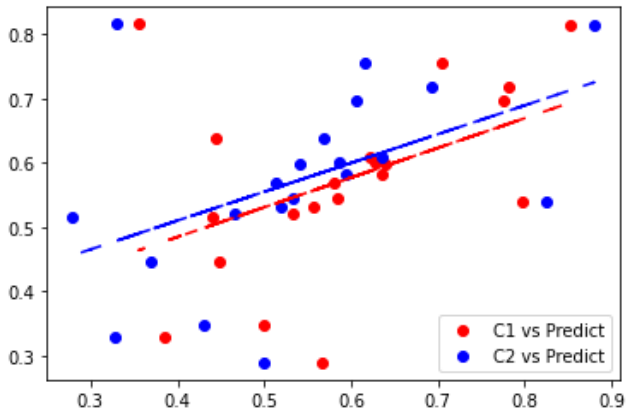

Figure 6: Correlation between the joint attention predictions based on ground truth head and object annotations, and the visual focus of attention annotations of Coder 1 (C1, in red) and Coder 2 (C2, in blue), respectively. Trendlines are superimposed.

5.5 Whole-interaction analysis

Finally, we explore the potential of our automated joint attention estimation at the level of an entire interaction. While predictions at the frame level can be inaccurate, when there is no systematic bias, these could be averaged out over an entire interaction and still yield accurate estimates of the overall amount of joint attention.

To investigate whether our method can be used to predict joint attention over an entire interaction, we have aggregated the joint attention values over all objects and frames as the average number of objects that is attended to by both parent and child per frame. In Figure 6, we visualize these joint attention predictions in relation to the VFOA annotations of Coder 1 and Coder 2, respectively.

We observe that the predictions follow the annotations. The Pearson correlation between the predictions and the annotations of both coders is (marginally) significant, respectively $r(19) = 0.410$, $p = 0.051$ and $r(19) = 0.447$, $p = 0.037$ for C1 and C2.

The trendlines of the two coders are similar, with slightly higher VFOA scores for C1. The trendlines for C1 and C2, respectively, are $0.300 + 0.461x$ and $0.331 + 0.447x$, with the x the annotated average number of objects per frame that receive joint attention. We observe that the correlation is mainly skewed by one interaction with a relatively high predicted score (81.58%), whereas the averaged annotated VFOA of all objects is markedly lower with 35.53% and 32.89% for C1 and C2, respectively. In this interaction, the father looks down in a significant part of the interaction. The attention heat map is more diffuse due to the limited visibility of the face. Therefore, the object that the child interacts with typically is predicted as being looked at by the father. Instead, both coders have predominantly annotated gaze at the child's face.

The correlation between predictions and ground truth shows that our method might be suitable for screening of interactions that contain a lot or, conversely, little joint attention. As such, it can be a proxy to understand the quality or type of interaction.

6 CONCLUSIONS

In this paper, we propose an automatic method to detect joint attention and mutual gaze during free play parent-child interactions. Our approach combines head detection, object detection, and visual focus of attention classification. Our experiments are conducted on naturalistic parent-child videos, which do not require specialized eye-tracking equipment. We manually annotate and analyze 250 minutes of interaction videos to evaluate our approach and compare our results with those obtained from automated face and object detection methods. Finally, we offer qualitative insights into how continuously measured visual focus of attention can improve our understanding of parent-child interactions. Our approach requires minimal training and overcomes some of the challenges of finding and annotating large amounts of interaction data.

Our work also has some limitations. In terms of the setting, significant head movements make it difficult to precisely estimate the gaze heat map. Additionally, detecting small objects (like shapes within the shape box) is seen to be unreliable using automated methods. Regarding our approach, using a single-view camera introduces difficulties due to toys overlapping with each other. In future work, we plan to technically improve the VFOA prediction by including multiple viewpoints and actively take into account the confidence of the head detection in promoting the best view. In addition, we expect that leveraging temporal continuity will also aid in improving the predictions. Finally, we plan to combine the VFOA estimations with a notion of human action, in particular regarding object use. We expect that these advances help us to more thoroughly measure the nature of parent-child interactions.

ACKNOWLEDGEMENTS

The work is partly funded by the China Scholarship Council (CSC).

REFERENCES

[1] Nazan Aksan, Grazyna Kochanska, and Margaret R Ortmann. 2006. Mutually responsive orientation between parents and their young children: toward methodological advances in the science of relationships. *Developmental psychology* 42, 5 (2006), 833.

[2] Sharifa Alghowinem, Huili Chen, Cynthia Breazeal, and Hae Won Park. 2021. Body Gesture and Head Movement Analyses in Dyadic Parent-Child Interaction as Indicators of Relationship. In *16th IEEE Int. Conf. on Automatic Face and Gesture Recognition*. 1–5.

[3] Marie Avril, Chloë Leclère, Sylvie Viaux, Stéphane Michelet, Catherine Achard, Sylvain Missonnier, Miri Keren, David Cohen, and Mohamed Chetouani. 2014. Social signal processing for studying parent–infant interaction. *Frontiers in psychology* 5 (2014), 1437.

[4] Sileye O Ba and Jean-Marc Odobez. 2008. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 1 (2008), 16–33.

[5] Carlos Bermejo, Dimitris Chatzopoulos, and Pan Hui. 2020. Eyeshopper: Estimating shoppers' gaze using CCTV cameras. In *Proc. ACM Multimedia*. 2765–2774.

[6] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. 2020. Detecting attended visual targets in video. In *Proc. CVPR*. 5396–5406.

[7] Kirby Deater-Deckard, Maria V Pylas, and Stephen A Petrill. 1997. *Parent-Child Interaction System (PARCHISY)*. Institute of Psychiatry, London, UK.

[8] Soumia Dermouche and Catherine Pelachaud. 2019. Engagement modeling in dyadic interaction. In *Proc. ICMI*. 440–445.

[9] Allyson Funamoto and Christina M Rinaldi. 2015. Measuring parent–child mutuality: A review of current observational coding systems. *Infant Mental Health Journal* 36, 1 (2015), 3–11.

[10] Frances Gardner. 1997. Observational Methods for Recording Parent-Child Interaction: How Generalisable Are the Findings? *Child Psychology and Psychiatry Review* 2, 2 (1997), 70–74.

[11] Ronald W Henderson. 2013. *Parent-Child interaction: Theory, research, and prospects*. Academic Press.

[12] Gijs A Holleman, Ignace TC Hooge, Jorg Huijding, Maja Deković, Chantal Kemner, and Roy S Hessels. 2021. Gaze and speech behavior in parent–child interactions: The role of conflict and cooperation. *Current Psychology* 42 (2021), 1–22.

[13] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.

[14] Yuki Kodama, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, Ichiro Ide, Hiroshi Murase, Hidehisa Nagano, and Kunio Kashino. 2019. Localizing the gaze target of a crowd of people. In *Proc. ACCV*. Springer, 15–30.

[15] Murray Krantz, Susan Wanska George, and Kathleen Hursh. 1983. Gaze and mutual gaze of preschool children in conversation. *The Journal of Psychology* 113, 1 (1983), 9–15.

[16] Taeahn Kwon, Minkyung Jeong, Eon-Suk Ko, and Youngki Lee. 2022. Captivate! contextual language guidance for parent–child interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[17] Susan R Leekam and Christopher AH Ramsden. 2006. Dyadic orienting and joint attention in preschool children with autism. *Journal of autism and developmental disorders* 36 (2006), 185–197.

[18] Eric W Lindsey, Jacquelyn Mize, and Gregory S Pettit. 1997. Mutuality in parent-child play: Consequences for children's peer competence. *Journal of Social and Personal Relationships* 14, 4 (1997), 523–538.

[19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single shot multibox detector. In *Proc. ECCV*. Springer, 21–37.

[20] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proc. CVPR*. 3566–3574.

[21] Annett Lotzin, Xiaoxing Lu, Levente Kriston, Julia Schiborr, Teresa Musal, Georg Romer, and Brigitte Ramsauer. 2015. Observational tools for measuring parent–infant interaction: A systematic review. *Clinical child and family psychology review* 18 (2015), 99–132.

[22] Rebecca A Marcon. 1999. Positive relationships between parent school involvement and public school inner-city preschoolers' development and academic performance. *School psychology review* 28, 3 (1999), 395–412.

[23] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. 2019. Laeo-net: revisiting people looking at each other in videos. In *Proc. CVPR*. 3477–3485.

[24] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. 2014. Detecting people looking at each other in videos. *International Journal of Computer Vision* 106 (2014), 282–296.

[25] Benoît Massé, Silèye Ba, and Radu Horaud. 2017. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Trans. PAMI* 40, 11 (2017), 2711–2724.

[26] James C McCroskey, Virginia P Richmond, Aino Sallinen, Joan M Fayer, and Robert A Barraclough. 1995. A cross-cultural and multi-behavioral analysis of the relationship between nonverbal immediacy and teacher evaluation. *Communication Education* 44, 4 (1995), 281–291.

[27] David J McDowell and Ross D Parke. 2009. Parental correlates of children's peer relations: An empirical test of a tripartite model. *Developmental psychology* 45, 1 (2009), 224–235.

[28] Vivian Nguyen, Otto Versyp, Christopher Cox, and Riccardo Fusaroli. 2022. A systematic review and Bayesian meta-analysis of the development of turn taking in adult–child vocal interactions. *Child Development* 93, 4 (2022), 1181–1200.

[29] Itir Onal Ertugrul, Yeojin Amy Ahn, Maneesh Bilalpur, Daniel S Messinger, Matthew L Speltz, and Jeffrey F Cohn. 2022. Infant AFAR: Automated facial action recognition in infants. *Behavior research methods* 55 (2022), 1–12.

[30] 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, Pascal Pas, Matthijs Vink, Juliëette JM van der Wal, Hilleke E Hulshoff Pol, and Chantal Kemner. 2020. The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience* 46 (2020), A100868.

[31] Cristina Palmero, Elsbeth A van Dam, Sergio Escalera, Mike Kelia, Guido F Lichtert, Lucas PJJ Noldus, Andrew J Spink, and Astrid van Wieringen. 2018. Automatic mutual gaze detection in face-to-face dyadic interaction videos. In *Proceedings of Measuring Behavior*, Vol. 1. 2.

[32] Lorenzo Piccardi, Basilio Noris, Olivier Barbey, Aude Billard, Giuseppina Schiavone, Flavio Keller, and Claes von Hofsten. 2007. Wearcam: A head mounted wireless camera for monitoring gaze attention and for the diagnosis of developmental disorders in young children. In *Proc. ROMAN*. IEEE, 594–598.

[33] Carly AY Reid, Lynne D Roberts, Clare M Roberts, and Jan P Piek. 2015. Towards a model of contemporary parenting: The parenting behaviours and dimensions questionnaire. *PloS one* 10, 6 (2015), e0114179.

[34] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2020. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 3327–3336.

[35] Michael Tomasello and Michael Jeffrey Farrar. 1986. Joint attention and early language. *Child development* 57, 6 (1986), 1454–1463.

[36] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *Proc. ICIP*. IEEE, 3645–3649.

[37] Zeynep Yücel, Albert Ali Salah, Çetin Meriçli, Tekin Meriçli, Roberto Valenti, and Theo Gevers. 2013. Joint attention by gaze interpolation and saliency. *IEEE Trans. on cybernetics* 43, 3 (2013), 829–842.

[38] Lingyu Zhang, Mallory Morgan, Indrani Bhattacharya, Michael Foley, Jonas Braasch, Christoph Riedl, Brooke Foucault Welles, and Richard J Radke. 2019. Improved visual focus of attention estimation and prosodic features for analyzing group interactions. In *2019 International Conference on Multimodal Interaction*. 385–394.

[39] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2017. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. PAMI* 41, 1 (2017), 162–175.