2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Survey of Automated Methods for Nonverbal Behavior Analysis in Parent-Child Interactions

Berfu Karaca[1,2], Albert Ali Salah[2], Jaap Denissen[1], Ronald Poppe[2], Sonja M.C. de Zwarte[1]
[1] Dept. Developmental Psychology, Utrecht University, Utrecht, The Netherlands
[2] Dept. Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

***Abstract*— Social interactions are fundamental for human beings, motivating the abundance of studies into the behavioral correlates of constructs such as personality and relationship. The primary drivers of this research are video-taped recordings of interactions. Recent advancements in automatic behavior analysis provide a cost-effective and more objective alternative to manual coding by trained experts. Still, the use of automated analysis is far from trivial. In this literature survey, we discuss the current state-of-the-art in automated parent-child interaction analysis, and critically assess opportunities and limitations. We focus on parent-child interactions as they reflect various aspects of a child's development, and provide distinct challenges for the automated measurement and interpretation of the interactive behavior. We briefly discuss single-person and dyadic nonverbal measurements, and identify measurement challenges. We then provide an overview of various developmental constructs that can be measured through the classification of extracted cues. Finally, we outline persistent limitations of the current state-of-the-art, and we highlight promising directions to bridge the gap between manual and automated measurements.**

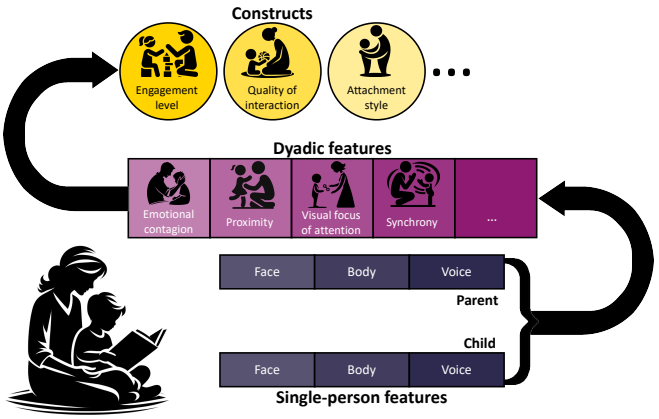

Fig. 1. Schematic overview of the process of measuring and interpreting interactive behavior in parent-child interactions. Recordings are analyzed for single-person behaviors, before combining these into dyadic features. Finally, interpretation takes places in terms of constructs for the child's development or the interaction quality.

I. INTRODUCTION

Parent-child interactions (PCIs) offer insights into many aspects of a child's development, including cognition, language acquisition, and socio-emotional growth, as well as the achievement of developmental milestones [1]–[3]. Observation of videotaped PCIs stands out as an integral assessment technique in tracking child development [4], [5].

Videos of PCIs contain numerous informative cues and the quantitative analysis of these cues provides insights into a wealth of constructs including the child's development. In conventional non-computational studies, videos are manually labeled by trained coders [6]–[8]. Manual labeling is typically laborious and requires extensive training [4]. Employing computer analysis techniques promises a paradigm shift, allowing for automated extraction of important features from PCIs, thus facilitating more objective analysis of larger amounts of data. Despite the prevalence of computer analysis in adult behavior analysis studies, research addressing PCIs remains limited. This review seeks to bridge the areas of developmental psychology and computer science, shedding light on opportunities to automate behavioral coding for the interpretation of developmental constructs and specific challenges in analyzing child behavior, and understanding the complexity of PCIs.

This work was supported by an AiNed Fellowship to SMCdZ (grant no NGF.1607.22.006).

We focus on nonverbal behavior because it is the primary mode of communication of young children, and remains important while speech is still under development [9]. Therefore, a substantial portion of PCI research with young children, in particular in infants, has concentrated on nonverbal behaviors [10]. An overview of the process of measuring and interpreting interactive behavior in parent-child interactions appears in Figure 1.

High-level constructs such as attachment style are too complex to measure directly and require expert interpretations. Therefore, researchers have focused on measurable features associated with developmental constructs. Analysis of the dyad requires the prior measurement of single-person behaviors of both parent and child, typically in terms of facial expressions, body movements, and vocalizations. The analysis of interaction dynamics must consider behaviors within the dyadic context [11], and involves techniques such as time-series analysis, or dyadic measures such as proximity and synchrony. Most of the surveyed studies aim at understanding the correlates of behavior cues and aspects of personality, the nature of the interaction, or developmental constructs such as attachment style. Others take a real-time approach, such as understanding the success of parent-child interaction therapy (PCIT) and improving the quality of interaction [12], [13]. Since these studies focus on speech modality, they are not included in this literature survey.

The remainder of this paper is structured as follows. Section II presents an overview of parent-child interaction paradigms. In Section III, we discuss automatic behavior measurement at the single-person and dyadic levels. We then turn to developmental constructs assessed through the interpretation of extracted behavioral cues in Section IV. In Section V, we identify persistent gaps in the field of automated PCI analysis and outline directions for future research. We conclude in Section VI.

II. PARENT-CHILD EXPERIMENTAL SETTINGS

Parent-child interaction videos have been conventionally explored in observational studies using behavioral tasks, including the Ainsworth Strange Situation, Manchester Child Attachment Story Task, and Still-Face paradigm. More recently, less controlled paradigms are explored, including structured play, free play, story reading, as well as psychodynamic therapy sessions. The physical and social setting in which the interaction is observed determines to a large extend the type of behaviors that we can expect to observe, and which type of measurement tools are appropriate. In this section, we discuss the most common experimental settings that have been addressed in automated PCI analysis research. Table I provides an overview of studies that have used a specific paradigm.

The Ainsworth Strange Situation Assessment (SSA) is the recognized measure of attachment quality via observing the infant's exploration behavior and responses to the separation and reunion with the parent, as well as the reaction to a stranger [14]. By coding attachment behaviors including proximity seeking, secure base behavior, stranger anxiety, separation anxiety, and reunion response, children's attachment security is classified into secure, insecure, or disorganized, depending on the reactions of the child in different phases of the task.

Manchester Child Attachment Story Task (MCAST) is another method to assess attachment style. While SSA is the standard method to measure attachment style in infancy, MCAST is designed for pre-school children [15]. During this assessment, children listen to five distress-related stories in different contexts: breakfast, nightmare, hopscotch, tummy-ache, and getting lost in a shopping mall. At the end of each story, children are asked to play with two dolls (baby doll and mommy doll) to complete the stories of the examiner. With this procedure, children are rated in terms of their attachment-related behaviors, narrative coherence, disorganized attahment behaviors, and the mentalizing ability.

Face-to-Face/Still Face (FFSF) is a validated procedure measuring socio-emotional regulation of infants facing a social stressor [16]. It consists of phases of face-to-face (FF) interaction such as a play episode, followed by a still-face (SF) period in which the caregiver stops communication, and ending with a reunion (RE) phase in which the infant and the caregiver resume normal face-to-face interaction. The SF phases typically induce stress in the infants. The FFSF paradigm is frequently employed to assess infants' emotional regulation development. The emotions, behaviors, and interaction dynamics observed during each episode offer valuable information on the PCIs, and contribute to the assessment of the infant's development of emotion regulation skills and resilience.

Structured and free play: While structured play includes inherent rules and goals, in free play, parents and the children engage together in their own preferred way, selecting toys or activities without strict guidelines [17]. For instance, cooperative and competitive games are being used in structured play settings to observe the behavioral patterns in different conditions of cooperation and competition [18]. Parents and children adjust their leading behaviors depending on the condition, in synchrony during cooperation and independently in competitive settings [19].

In semi-structured play settings, interactants are placed in a play situation to achieve a specific goal such as measuring joint attention. For example, the Three-Bag Procedure is a semi-structured play task in which parents are asked to play with their children using three bags of age-appropriate toys in a set order [20]. Joint storybook reading is another interaction setting used to observe the mutual engagement in PCIs [21]. Naturalistic observations might be in the context of structured or free play, as well as observation of routines between parents and children during various activities regardless of context and location [22], [23].

Parent-Child Interaction Therapy (PCIT) is a real-time application of PCI research. PCIT is an evidence-based treatment program designed to support caregivers and their children, aged between 2 and 7 years old, with behavioral and emotional difficulties. PCIT consists of two phases: a Child-Directed Interaction Phase (CDI) encouraging the child to lead the play activity, aiming at enhancing parent-child relationship during child-led play, and a Parent-Directed Interaction Phase (PDI) where parents enhance their behavior management strategies with the ultimate goal of creating a supportive home environment.

III. AUTOMATIC BEHAVIOR ANALYSIS

We discuss automatic analysis of nonverbal behaviors first at the single-person level, before discussing dyadic nonverbal measurements. Nonverbal dyadic features are those behavioral cues that relate to both parent and child, such as synchrony, proximity, visual focus of attention, and emotional contagion, that are derived from the single-person measurements. In our overview Table I, we summarize automated PCI analysis publications, and indicate the features, both single-person and dyadic, in the first column, whereas the methods themselves are in the "Tools" column.

A. Single Person Features

We identify facial expressions, body movements and non-linguistic vocalizations as the main sources of nonverbal behavior cues.

1) Facial Expressions: The predominant representations of facial expressions are rooted in the facial action coding system (FACS) [43] and its infant-specific adaptation, Baby-FACS [44]. FACS breaks down individual or groups of facial

Single-Person Features	Dyadic Features	Construct	Ref	Age	Tool(s)	Setting
Facial Expressions	Emotional Contagion	Engagement Level Attachment Style	[24]	6m	Face analysis	FFSF
		Attachment Style	[25]	4m	Face analysis	FFSF
Body Movements	Emotional Contagion	Attachment Style	[26]	4m	Face tracking	FFSF
	Synchrony	Attachment Style	[27]	4m	Head tracking	FFSF
			[28]	5.5m	Object tracking Gaze tracking	Free Play
		Engagement Level	[29]	41-100m	Pose estimation	Free Play
			[30]	2-7y	Movement tracking	Free Play
Vocal	Emotional Contagion		[31]	3-24m	Audio analysis	Free Play, FFSF
Multimodal: Body Movements Body part regions		Engagement Level	[32]	3-7y	Image analysis	Storybook Reading
Multimodal: Body Movements, Vocalizations	Synchrony	Engagement Level Interaction Quality	[33]	2-18m	Pose estimation Audio analysis	FFSF
		Attachment Style	[34]	5-8m	Image analysis Audio analysis	FFSF, SSA
Multimodal: Facial Expressions Vocalizations		Attachment Style	[35]	5-9y	Facial analysis Audio analysis	MCAST
Body Movements	**Synchrony**	Engagement Level	[10]	3-7y	Pose estimation	Storybook Reading
		Interaction Quality Attachment Style	[36]	12-24m	MEA	Free Play
		Attachment Style	[19]	8-12y	MEA	Structured Play
	Proximity	Interaction Quality Attachment Style	[37]		Image analysis	Seated FF
			[38]	10m	Pose estimation Image analysis	Free Play
Head Movements	**VFOA**	Engagement Level	[39]	14m	Image analysis	Semi-Structured Play
			[40]	16m-14y	Image analysis	Structured Play
		Interaction Quality	[41]	10m	Head tracking	Free Play
Multimodal: Body Movements, Vocalizations, Proximity		Attachment Style	[42]	6-18m	Pose estimation Audio analysis	Semi-Structured Play

TABLE I

OVERVIEW OF STUDIES FOCUSING ON AUTOMATED ANALYSIS OF PCIS. SEE TEXT FOR FEATURES, CONSTRUCTS AND EXPERIMENTAL SETTINGS. YELLOW CELLS REPRESENT STUDIES DIRECTLY FOCUSING ON CONSTRUCTS, WHILE GRAY CELLS REPRESENT STUDIES INDIRECTLY ADDRESSING CONSTRUCTS. M: MONTHS, Y= YEARS, MEA: MOTION-ENERGY ANALYSIS, FFSF: FACE-TO-FACE/STILL FACE, SSA: AINSWORTH STRANGE SITUATION ASSESSMENT, MCAST: MANCHESTER CHILD ATTACHMENT STORY TASK, FF: FACE-TO-FACE

muscles into Action Units (AUs), allowing the identification of specific facial expressions through the combination of these action units [45].

Facial expressions and their dynamics have been mainly associated to categories or dimensions of affect [46]–[49]. Due to importance of emotional development of the infant for positive parent-child relationships and high-level constructs, such as attachment style [50], studies exploring facial expressions frequently focus on emotional states or changes in these states.

Smile production in infants start at birth [51]. Social smile, emerging in the first three months, is one of the first observable behaviors to track the socio-emotional development of the infant, indicating positive emotion in early interactions [52], [53]. To differentiate different types of smiles, such as Duchenne vs non-Duchenne smiles, researchers explored components of smile intensity, mouth opening, and eye constriction [54], and explored different aspects of smile like reciprocating positive affect or exuberance [51].

The FFSF paradigm is commonly employed to measure the affect interplay throughout the episodes [25], [46]–[48]. Generally, facial expressions are naturally associated to positive emotions observed during the FF episode, while the SF episode leads to an overall increase in the occurrence of negative emotions [25], [55]. Exploring Duchenne cry-faces for the occurence of negative emotions during FFSF paradigm, researchers found increased Duchenne cry-faces during SF episode while there were no difference for non-Duchenne cry-faces [56]. These findings demonstrates the importance of differentiating intense Duchenne forms of expressions as a sensitive index of affective valence [25].

A substantial part of the studies classified facial expressions as positive, negative and/or neutral emotional states [46], [47], [49]. Shifting from categorical approaches, dimensional approaches are emerging [57], which typically consider valence and arousal as two orthogonal dimensions. However, the use of these methods in PCI is limited, potentially due to a scarcity of methods to analyze affect in children. Mang et al. adopted a dimensional approach, marking the first automated attempt to estimate the emotional valence of infant affect from facial expressions [48].

2) Body Movements: Body movements play a fundamental role in social interactions [58], including those between parents and their children. Isolated behaviors such as head and body gestures convey information about the dynamics between parent and the child related to emotional contagion, and joint engagement [26], [29], [30]. Moreover, quantiative representations of body movement can be used to analyze proximity and interaction synchrony [27], [30].

Head trackers [27] provide head orientation in three orthogonal directions. Body tracking systems [30] are used to record full-body poses. More recently, video-based methods

have been introduced as a cost-effective and less intrusive alternative to body-worn trackers [29]. Body tracking methods typically represent body poses as skeleton data, with the position of key joints in the body either in 2D image coordinates or in 3D world coordinates [59]. Popular video-based pose estimation algorithms include OpenPose [60], DensePose [61], and DarkPose [62]. Movement measures are obtained by considering pose measurements over time.

While analysis can focus on continuous variables such as the amount of movement or the orientation of the head, and body poses, movements can also be classified into specific actions or gestures [58]. Recent efforts are aiming to bridge the gap in infant action recognition. Yurtsever et al. conducted a pioneering study to classify the meanings of infants' activities [63]. They captured temporal information obtained from keypoints from different pose estimators and proposed a novel system for decoding infants' body language in videos. Dechemi et al. presented the BabyNet network, incorporating temporal inter-dependencies for video-based infant reaching action recognition [64].

3) Vocalizations: Because infants are not linguistically competent, parents and other caregivers modify their speech to them in a variety of ways to communicate [65]. Infant vocalizations, especially expressions of emotions, often serve communicative purposes in dyadic social processes. Basic emotional infant outbursts, such as cry, fuss, laugh, babble, and screech can convey meaningful information to parents. Additionally, parent vocalizations have been frequently explored as motherese/fatherese, adult-directed speech, infant-directed speech, playful noises, rhythmic sounds, laughter, and whispering [31].

Studies of infant vocalizations often utilize speech analysis tools such as openSMILE [66] to extract inter-vocalization intervals and acoustic features including pitch (level, range, and contour of the fundamental frequency F0), intensity as the energy of the voice, and duration [31], [67].

4) Multimodal: Human behavior is inherently multimodal, where dyadic interaction signals manifest through the interplay of body language, facial expressions, and vocalizations [68]. Fusion of features from various modalities enhances the efficacy of models in extracting meaningful information by resolving ambiguity and improving the overall quality of noisy data [33], [42], [69].

In an effort to enhance emotion recognition, Yang et al. investigated the fusion of gaze and head pose behaviors in interactions between infants and parents [69]. Through manual coding of the gaze behavior and emotions, they trained a deep neural network to capture the temporal information and image details, demonstrating improved recognition of gaze and affect with the incorporation of head pose and gaze features.

Furthermore, especially for young children whose communicative skills are developing, specific communicative functions could be expressed in different modalities as the child develops. Modalities are combined to explore developmental constructs including engagement level and attachment style [33], [34], [42].

Techniques to fuse multiple modalities can be broadly categorized into feature-level fusion and decision-level fusion. Feature-level fusion is performed by combining the extracted features from each modality and to integrate them into a large vector to reach a joint representation. Behavior classifications are subsequently made on these vectors. In contrast, decision-level fusion integrates the classification of each modality for a final output. In literature, decision-level fusion techniques are most common for various reasons such as to avoid overfitting as a result of the inbalance in the feature dimensionality and number of observations, and to reach better representations for temporarily correlated asynchronous modalities [70], [71].

B. Dyadic Features and Interaction Analysis

Single-person measurements are typically combined per modality, to reveal specific inter-personal measures. We discuss the most prominent ones.

1) Emotional Contagion: Affect plays a significant role in shaping and regulating our behaviors and interpersonal relationships with others in social interactions. The dynamic nature of affect, intimately tied to the ongoing dynamics of social interactions, exerts a continuous influence on behaviors throughout the interaction. Consequently, researchers have shown a keen interest in exploring emotional contagion in interactions, where an interactant's emotional state is a precedent of other interactant's emotional state in a responsive manner [72]–[74].

In the context of PCIs, facial expressions have been the predominant modality to explore emotional contagion during the interaction [24], [25]. Head movements also play a significant role in interpreting emotional states. Mother and infant angular displacement has been measured, shedding light on the perturbation and recovery of head movement coordination in emotion exchanges [26].

Vocal behaviors also convey information about the emotional contagion between the parent and the child. Studies exploring nonverbal vocalizations highlighted the importance of fundamental frequency compared to other acoustic features to recognize emotional state of the individuals [31], [67].

2) Proximity and Touch: The role of proximity and touch in parent-infant interaction is also critical, influencing infants' attention, arousal levels, behavioral and emotional states, as well as contributing to emotion regulation [37]. This tactile dimension has been also associated with the quality of interaction in previous studies [38].

Tracking systems are often employed to identify touch events within parent-infant interactions. Chen et al. introduced a touch event detection system integrating hand tracking and body analysis, as well as addressing a common concern of hand occlusions during tracking [75]. In this setting, they defined a touch event as the merging of contours of the adult's hand with the infant's contour. In a subsequent study, the authors shifted from motion trackers to a computer vision-based approach for the automated recognition of different types of caregiver touch classified based on the

infant's touched body parts [37]. Their model allowed the trained analysts to skip annotating a remarkable portion of frames while still capturing the vast majority of actual touch frames.

These studies have been conducted on seated interactions with limited freedom of movement for both participants, also only focusing on hand-to-body touch events. To address limitations of working within a narrow situational context, Doyran et al. provided an approach utilizing Convolutional Neural Networks (CNNs) to detect frame-level touch in free-play settings [38]. They also extended the focus to all physical touch events. Similar to [37], using body part segmentation allowed a more semantically structured modeling for interaction analysis, by distinguishing between different body parts that are used in different contexts [38].

3) Visual Focus of Attention: Gaze is another communication channel to convey emotional and mental states of individuals [76]. Investigating under the umbrella of Visual Focus of Attention (VFOA), gaze provides insights into the area an individual is looking at, providing a valuable aspect in understanding interactions. The dyadic exploration of VFOA extends to various contexts, including joint attention, mutual gaze, and gaze following [77]. While mutual gaze refers to the visual exchange between two individuals, joint attention refers to the shared focus of attention.

Mutual gaze, or eye contact, emerges as a significant factor in fostering effective communication and enlightening social interactions by signalling interest, attention, and active participation [40], [77]. Consequently, measurements of eye contact find application in diverse areas such as assessing the social communication skills of children at risk for developmental disorders like Autism Spectrum Disorder (ASD), as well as analyzing turn-taking and social roles [40].

Many studies opt for estimate VFOA through face and head pose detection. By identifying head pose angles, a vector representing the person's face direction can be calculated, allowing for the estimation of VFOA. Extracted cues of VFOA has been associated with developmental contents of emotional states [69] and quality of interaction [41] in PCIs.

4) Interactional Synchrony: Interactional synchrony is the temporal coordination of behavioral patterns between interactants [78]. It has been characterized with behaviors involving direct imitation or mirroring of others and congruency between interactants [79]. Mirroring occurs when one person consciously or unconsciously mimics the nonverbal communication of the other. Mirroring has been found to correlate with empathy between people and it is an early indicator of a positive outcome in an interaction [80], [81]. Infants copy caregiver's actions for social learning, to acquire social, communication, and emotion regulation skills [81].

Synchrony manifests in various forms, spanning behavioral, emotional, physiological, and neurological dimensions [82]. Its association with the development of social skills in early childhood has garnered significant attention. In the literature, synchrony has been defined in various ways and it is distinguished from mirroring by emphasizing the dynamic nature of timing over the nature of behaviors [83]. Behaviors in an interaction setting are synchronized in both timing and form in a patterned way which is providing insight into interpersonal coordination [79]. As such, interpersonal coordination has been associated with the engagement level and quality of interaction [83].

While timing in synchrony quantification is often ignored, other methods explicitly take into account the relative timing of the behaviors of both child and parent. Temporal methods such as time-lagged cross-correlation and recurrence analysis, and spectral methods such as cross-spectral coherence and power spectrum overlap, are commonly employed for assessing interaction synchrony [83].

Researchers have also explored synchrony in terms of leader-follower dynamics. PCIs can be globally classified into leader-follower dynamics "parent-led" and "infant-led". The latter is associated with higher levels of parent-child synchrony while parent-led interactions are typically stronger associated with the fulfillment of needs throughout the course of the interaction [84]. Yarmolovsky et al. distinguished between in-phase and anti-phase synchrony and identifying leaders in synchronous interactions [19]. While in-phase synchrony refers to movements that are in perfect harmony being associated with cooperative actions, anti-phase synchrony is characterized by alternating movements [85].

Synchrony has also been explored in turn-taking behavior within the context of affective synchrony of facial expressions. Understanding the preconditions of infant intentionality, studies have modelled dyadic state transitions and turn-taking behavior, unveiling early patterns of simultaneous responsivity [86]. In the context of turn-taking behavior, vocal features have also been frequently explored [87].

Behavioral synchrony is commonly analyzed using trackers during face-to-face interactions. While head-trackers are often utilized in seated settings [27], movement trackers have been used for less constrained settings such as free play, unveiling patterns of synchrony crucial for understanding interaction dynamics [28]. Increasingly, free play settings have been investigated, such as by Hammack et al. who measured dyadic movement synchrony using motion energy analysis (MEA) in at-home free play sessions [36].

Recent studies also explored inter-modal synchrony in PCIs. Klein et al. integrated features such as head position, arm position, and vocal fundamental frequency, achieving more robust models with expanded feature space against missing data [33].

C. Measurement Challenges

Despite significant advances in automated behavior analysis, several measurement challenges persist in the unobtrusive measurement and interpretation of behavior.

Analyzing infants: Computer vision and audio processing algorithms predominantly concern adults. Extending their use to children, especially young ones, is non-trivial. For example, infants have significantly different body dimensions, which complicates the estimation of their poses when adult-trained algorithms are used [89]. Also, infants exhibit more sudden and rapid movements, posing a challenge for

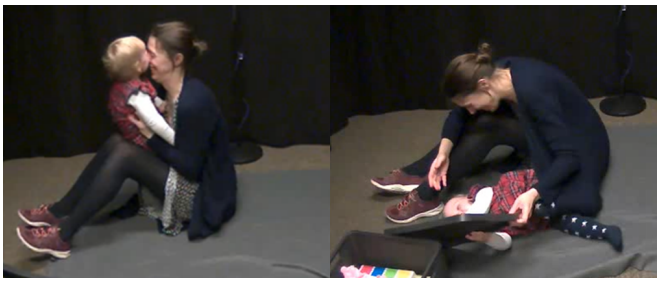

Fig. 2. Examples of challenging measurement situations regarding partial visibility and occlusions (see Section III-C) in free play PCIs. Images taken from the YOUth PCI dataset [88], with permission.

traditional computer vision models [90]. Inherent differences in the dynamics of facial expressions between infants and adults further present challenges [48], [91].

Partial visibility: Video cameras are unobtrusive but they might not be able to record all essential information in the scene. Especially in setups with a single or limited number of cameras, faces or parts of the body might be turned away from the cameras. Another issue is the (partial) occlusion of faces and bodies by either the same person or the interaction partner. Examples appear in Figure 2. For facial expression recognition, researchers have proposed methods to deal with occlusions like employing generative models for learning [92], weight adjustment methods [93] or region-based attention mechanism to deal with facial occlusions [47]. Occlusions are particularly evident in dyadic interactions with close proximity [94]. Issues related to proximity can cause early commitments due to the failure of the person detector in top-down pose detection [60], effectively complicating the robust assessment of physical contact in these setting.

In studies involving toys, existing algorithms designed to detect faces and body parts may struggle to discern between toys and body parts. This limitation highlights the need for specialized algorithms capable of accurately identifying and tracking toys and objects during interactions.

While the use of multiple cameras can alleviate some of these issues, combining partial information from multiple views requires a robust setup and more extensive processing.

Data scarcity: With the increasing sophistication of algorithms, in particular deep learning models, there is an increased need for data to train these algorithms sufficiently. Limited labeled data reduce the effectiveness of algorithms. One prominent risk is overfitting, where the performance of trained models does not extend to novel situations. Currently, many algorithms in automated behavior analysis are typically trained on large amounts of relatively general data, such as images of faces and bodies in a wide variety of situations. More specific algorithms are further adapted to perform well in more specific settings such as a seated parent-child interaction, e.g., [95]. For example, Huang et al. have explored invariant representation learning settings for pose estimation, and have fine-tuned adult models for infant pose estimation [89].

Adaptation is sometimes challenging because domain-specific data are required. Especially for infants, such data might not always be available due to the lack of consent to share video material of minors. The relative scarcity of infant data complicates the development of robust algorithms, but also prevent proper benchmarking. In turn, the rate of advances in algorithm development specifically targeting infants and children is lower compared to work focusing on adults.

One solution to deal with the limited availability of infant data is to resort to synthetic data. For example, Huang et al. created a hybrid dataset that consists of synthetic and real infant pose data [89]. The synthetic data is based on a 3D human model that is animated based on pose data. By varying the viewpoint, the body appearance and poses, more data can be obtained that are to some extend representative for the target application domain.

Another solution is to make the most out of the limited data that is available. Researchers have explored data augmentation techniques to create variation in the time dimension such as randomly cropping a segment or frame sampling in the dimension of time [48].

IV. DEVELOPMENTAL AND INTERACTION CONSTRUCTS

Using video-based nonverbal features, researchers have focused on various indicators for the interaction state or the development of, mainly, the child. In the social sciences, many instruments including surveys have been developed and validated to measure these various constructs. Methods from computer science attempt to provide quantitative measurements for these constructs [96]. Achieving this can enhance the robustness of behavior coding, with the ultimate goal of advancing our understanding of PCIs. The "construct" column in Table I summarizes the various constructs that have been the focus on research on automated PCI analysis.

A. Engagement Level

Social engagement stands as a critical indicator of an individual's socio-emotional and cognitive states, and helps to understand interpersonal dynamics and communication [97]. Within the domain of PCI, considerable attention has been directed towards studying child engagement and joint engagement. Studies leverage body pose features as insightful nonverbal cues to predict engagement. While parent nonverbal cues, such as pose features, are suggested as predictive of child's engagement, studies also revealed that child's engagement and disengagement ratings were better predicted with dyadic features such as proximity [10].

Given the neccessity of affective states for social interactions and their profound links to learning and socio-emotional development, researchers have explored the interplay of affective states to measure joint engagement [32]. In a recent study, researchers classified engagement into positive affect, neutral affect, negative affect, and object engagement. Leveraging the high classification accuracy of deep neural

networks, promising results were obtained for coding infant engagement in FFSF recordings [98].

To enhance the recognition of affect interplay, action recognition techniques have been seamlessly integrated with various video augmentation techniques, introducing a hybrid model for improved joint engagement recognition. The incorporation of video augmentation techniques demonstrated increased performance, showcasing sensitivity to subtle social cues indicative of interaction dynamics [32].

Researchers also used movement and pose features to obtain information about parental responsiveness and sensitivity, which are instrumental to assess engagement levels. In a study using features from movement tracking systems, the event of reaching to a toy was used as a proxy for initiating or responding to toy play, as well as for caregivers' responsiveness [29]. By automating the measurement of caregiver responsiveness, researchers provided a promising monitoring technique for the engagement level in dyadic interactions.

B. Quality of Interaction

Understanding the quality of parent-child interactions can facilitate the investigation of healthy parent-child relationships [82]. In early studies, researchers used statistical analyses to characterize interaction quality of different groups of children by age and diagnosis, e.g., ASD, mental retardation, versus typically developed children [99]. Also, focusing on age groups, Egmose et al. investigated quality of mother-infant interactions at 4 and 13 months using motion features [84]. They found stronger correlations between the interaction quality and motion features at 4 months compared to 13 months, which emphasizes the importance of motion features to predict quality of interaction.

In later studies, rather than direct measurements, researchers focused on dyadic measurements such as proximity and joint attention, which are suggested to be informative to detect quality of interactions [37], [38], [41]. Even though there are studies paving the way by automatically extracting important behavioral cues such as body pose, joint attention, and proximity, there are no fully automated detection of interaction quality in PCI settings.

C. Attachment Style

Attachment is an enduring emotional bond that connects one person to another across time [42]. Early attachment styles are established in childhood through the interaction between infants and caregivers, in particular parents.

Research focusing on feature extraction within PCI offers insights into attachment styles. Responsivity and parental sensitivity are associated to attachment security. For example, researchers measured smile parameters, including smile strength and eye constriction of the mother to investigate the influence of interactants on each other's positivity as a signal of responsivity [24]. In another study using human posture analysis and voice activity, researchers focused on three components of maternal sensitivity as positive regards, intrusiveness, and sensitivity which are critical for the development of infant attachment security [42].

Body movements also provide information for the attachment style. Chen et al. used a vector autoregressive model with time-varying parameters, capturing temporal dependencies [26]. In this study they found variations in infant-mother head dynamics based on infants' attachment security, which demonstrates the role of coordination for secure attachment, and emphasises the importance of taking into account the dynamics of the behaviors.

Researchers also integrated different modalities with body movements, such as vocalizations. Employing a multimodal approach, Li et al. fused classifications of motion and acoustic features during FFSF paradigm and achieved to distinguish between different attachment styles of secure and insecure attachment [34]. Alsofyani et al. combined facial expressions and vocalizations for attachment prediction in school age children [35]. Overall, their results demonstrated better performance when using multimodal approach.

V. DISCUSSION AND FUTURE STEPS

The exploration of interaction dynamics through nonverbal behavior analysis has become a thriving research area, particularly in interdisciplinary studies examining developmental indicators in PCI using state-of-the-art computer vision techniques. While promising progress has been made, several limitations and gap still exist.

A. Standardized behavior measurement

As seen from Table I, there is little overlap in the types of tools and methods that have been used to analyze PCIs. Part of the pragmatic introduction and deployment of tools is the variety of experimental settings that have been addressed. But with the increasing sophistication of these algorithms, and consequently a more flexible and robust way of using them, also provides the opportunity for the development of a standardized set of behavior measures. We argue that common tools would ensure good uptake of advances in automated analysis, and would consequently improve the potential to directly compare the outcomes of studies.

In particular, the availability of common tools would allow the automated re-analysis of previously manually coded recordings. By examining differences in manual and automated coding, especially between studies, the validation and development of theory could be improved, in line with [100].

As a prerequisite for standardized measurement, there is a need for better tools to measure infant facial expressions and body pose. The distinct differences of children compared to adults in terms of the physical appearance and dynamics complicate the use of algorithms that have been developed predominantly with adult data.

B. Interaction dynamics

The focus on limited dyadic features such as proximity and synchrony, or behavior frequency distracts from examining the dynamic patterns inherent in interactions. The interactional contingency between interactants cannot be solely

represented by counting interactants' behaviors [101]. Yet, the majority of current studies are still focused on single-person rather than dyadic features. There is a need for temporal analyses that take into account causal relations between behaviors, such as a parent's smile in response to a child's movement, or the pick-up of a toy in response to a child's gaze.

The inclusion of the function of behaviors, potentially independent of the modality in which they are expressed, might be a way forward. This avenue also opens up possibilities for a more symbolic analysis, in contrast to the purely data-driven methods that are currently popular. Examples are graph-based methods or attention maps to provide local explanations [39].

In addition to only considering the two interactants in the PCI, there should be more focus on interactions with objects such as toys, or the environment. How children interact with the objects, and how the individuals manipulate the objects could also be explored.

C. Measurable constructs

Future studies should strive for standardized definitions of developmental constructs to enhance the understanding of PCIs. Bridging gaps in defined developmental constructs can facilitate studies exploring other complex constructs, such as empathy and Theory of Mind, and ensure better understanding of complex dynamics of PCIs [102], [103].

While a significant portion of the studies focus on sub-components of higher-level developmental constructs, there are only a few studies directly focusing on developmental constructs. In the future, these higher-level constructs should be targetted more by employing state-of-the-art computer analysis techniques. We argue that this requires a critical re-analysis of the more subjective and interpretation-focused aspects of currently popular assessment instruments. A good understanding of the relation between objective dyadic behaviors and common constructs would also support the interpretability of automated methods, and provide explanations to predicted outcomes.

D. Benchmark datasets

A significant necessity is a large multimodal dataset that researchers can use to develop and benchmark their algorithms. Methods cannot be directly compared because of a lack of standard, publicly available data. This prevents a good assessment of the relative strengths and weaknesses of approaches.

There are some public datasets provided to explore social interactions of children such as Multimodal Dyadic Behavior (MMDB) [104], Play Therapy 13 (PT13) [105], and Dyadic Affect in Multimodal Interaction - Parent to Child (DAMI-P2C) [106]. While these public datasets have valuable contribution to the field, the need for a large multimodal dataset generalizable to unstructured experimental settings, such as free-play PCIs, persists. Consequently, the iterative improvement that is common in algorithm development is significantly hindered.

While a domain-specific multimodal dataset might initially bias the research towards a specific physical or task-related setting, a common focus would aid in the consolidation of tools and standardized measures, as discussed before. When common, robust tools are available and there is more confidence in the potential and limitations of applying these tools, application in broader contexts is more straightforward.

E. Application outside the lab

In the long term, automated analysis of PCIs has many applications outside the confined experimental settings that are currently common. The majority of studies rely on fixed settings, neglecting naturalistic environments and tasks such as home settings or unstructured free-play scenarios. However, there is a trade-off between the real-world applicability of the interaction settings and the quality of the data. As mentioned in Section III-C, low-quality recording settings and the variations in the environment, such as lighting differences, cause challenges for efficient feature extraction from videos [39].

Naturalistic observation of interactions requires naturalistic experimental settings, such as free play. While the less confined setting poses challenges in analyzing more complex behavior patterns that involve potential others and the environment, there are still stable interactive behaviors that can be focused on. For example, Jayaraman et al. [107] analyze gaze at the parent, including mutual gaze, by analyzing head-worn cameras. Such measurements, albeit somewhat obtrusive, can be made over extended periods of time, thereby providing a more complete picture of a child's development.

VI. CONCLUSION

We have discussed the state-of-the-art in automated analysis of nonverbal behavior in parent-child interactions (PCIs). Increasingly, automated methods provide a low-cost, objective alternative to manual coding. At the same time, we observe a trend in full automatically providing a qualitative or quantitative assessment of higher-order constructs regarding the interaction quality or development of the child.

Despite significant progress in the robustness of measurement tools, there is comparatively little focus on the analysis of children. We have identified the lack of public datasets that can generalize to unstructured PCI settings as a main obstacle. There is an urgent need for robust, broadly applicable tools that can aid in the standardization of the measurement across studies and settings.

Furthermore, there is room for improvement in terms of the assessment of temporal dyadic behaviors such as leader-follower dynamics and the investigation of behavior patterns across modalities. By pursuing a more multimodal approach, we can increasingly shift from form to function. This shift will aid in the interpretability of the measurements, and will bridge the gap between objective coding by algorithms, and the more subjective assessment of higher-level constructs.

REFERENCES

[1] J. P. Shonkoff and P. Hauser-Cram, "Early intervention for disabled infants and their families: A quantitative analysis," *Pediatrics*, vol. 80, no. 5, pp. 650–658, 1987.

[2] N. Minick, "Mind and activity in Vygotsky's work: An expanded frame of reference," *Cultural dynamics*, vol. 2, no. 2, pp. 162–187, 1989.

[3] S. J. Meisels and J. P. Shonkoff, *Handbook of early childhood intervention*. Cambridge University Press, 1990.

[4] A. Sommer, C. Hachul, and H.-G. Roßbach, "Video-based assessment and rating of parent-child interaction within the national educational panel study," in *Methodological issues of longitudinal surveys: The example of the National Educational Panel Study*, pp. 151–167, Springer, 2016.

[5] E. Ceulemans, N. Bodner, S. Vandesande, K. Van Leeuwen, B. Maes, *et al.*, "Parent-child interaction: A micro-level sequential approach in children with a significant cognitive and motor developmental delay," *Research in Developmental Disabilities*, vol. 85, pp. 172–186, 2019.

[6] T. Grebelsky-Lichtman, "Children's verbal and nonverbal congruent and incongruent communication during parent–child interactions," *Human Communication Research*, vol. 40, no. 4, pp. 415–441, 2014.

[7] S. Schroer, L. Smith, and C. Yu, "Examining the multimodal effects of parent speech in parent-infant interactions.," in *CogSci*, pp. 1015–1021, 2019.

[8] S. E. Schroer and C. Yu, "The real-time effects of parent speech on infants' multimodal attention and dyadic coordination," *Infancy*, vol. 27, no. 6, pp. 1154–1178, 2022.

[9] U. Liszkowski, "Two sources of meaning in infant communication: preceding action contexts and act-accompanying characteristics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1651, p. 20130294, 2014.

[10] J. Shen, Y. Li, J. Hassan, S. Alghowinem, H. W. Park, C. Breazeal, and R. Picard, "Fostering parent-child interactions through behavioral understanding of synchrony," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 1–4, IEEE, 2023.

[11] R. Trabelsi, J. Varadarajan, L. Zhang, I. Jabri, Y. Pei, F. Smach, A. Bouallegue, and P. Moulin, "Understanding the dynamics of social interactions: A multi-modal multi-view approach," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1s, pp. 1–16, 2019.

[12] B. W. Funderburk and S. Eyberg, "Parent–child interaction therapy.," in *History of psychotherapy: Continuity and change, 2nd ed.*, pp. 415–420, American Psychological Association, 2011.

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

[14] M. Ainsworth and S. B. abd D.J. Stayton, "Individual differences in strange-situational behaviour of one-year-olds," in *The origins of human social relations* (H. Schaffer, ed.), Academic Press, 1971.

[15] J. Green, C. Stanley, V. Smith, and R. Goldwyn, "A new method of evaluating attachment representations in young school-age children: The manchester child attachment story task," *Attachment & Human Development*, vol. 2, no. 1, pp. 48–70, 2000.

[16] E. Tronick, H. Als, L. Adamson, S. Wise, and T. B. Brazelton, "The infant's response to entrapment between contradictory messages in face-to-face interaction," *Journal of the American Academy of Child Psychiatry*, vol. 17, no. 1, pp. 1–13, 1978.

[17] L. J. Krijnen, M. Verhoeven, and A. L. van Baar, "Observing mother-child interaction in a free-play versus a structured task context and its relationship with preterm and term born toddlers' psychosocial outcomes," *Frontiers in Child and Adolescent Psychiatry*, vol. 2, p. 1176560, 2023.

[18] M. Eriksson, B. Kenward, L. Poom, and G. Stenberg, "The behavioral effects of cooperative and competitive board games in preschoolers," *Scandinavian Journal of Psychology*, vol. 62, no. 3, pp. 355–364, 2021.

[19] J. Yarmolovsky and R. Geva, "Follow the leader: Parent-and child-led synchrony in competitive and cooperative play," *Journal of Nonverbal Behavior*, pp. 1–17, 2023.

[20] NICHD Early Child Care Research Network and others, "Early child care and mother–child interaction from 36 months through first grade," *Infant Behavior and Development*, vol. 26, no. 3, pp. 345–370, 2003.

[21] J. N. Kaderavek and E. Sulzby, "Parent-child joint book reading: An observational protocol for young children," *American Journal of Speech-Language Pathology*, vol. 7, no. 1, pp. 33–47, 1998.

[22] S. Bai, R. L. Repetti, and J. B. Sperling, "Children's expressions of positive emotion are sustained by smiling, touching, and playing with parents and siblings: A naturalistic observational study of family life," *Developmental Psychology*, vol. 52, no. 1, pp. 88–101, 2016.

[23] C. A. Ewin, A. Reupert, and L. A. McLean, "Naturalistic observations of caregiver–child dyad mobile device use," *Journal of Child and Family Studies*, vol. 30, pp. 2042–2054, 2021.

[24] S.-M. Chow, L. Ou, J. F. Cohn, and D. S. Messinger, "Representing self-organization and nonstationarities in dyadic interaction processes using dynamic systems modeling techniques," *Innovative Assessment of Collaboration*, pp. 269–286, 2017.

[25] Y. A. Ahn, I. Önal Ertuğrul, S.-M. Chow, J. F. Cohn, and D. S. Messinger, "Automated measurement of infant and mother duchenne facial expressions in the face-to-face/still-face," *Infancy*, vol. 28, no. 5, pp. 910–929, 2023.

[26] M. Chen, S.-M. Chow, Z. Hammal, D. S. Messinger, and J. F. Cohn, "A person-and time-varying vector autoregressive model to capture interactive infant-mother head movement dynamics," *Multivariate Behavioral Research*, vol. 56, no. 5, pp. 739–767, 2021.

[27] Z. Hammal, J. F. Cohn, and D. S. Messinger, "Head movement dynamics during play and perturbed mother-infant interaction," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 361–370, 2015.

[28] D. López Pérez, G. Leonardi, A. Niedźwiecka, A. Radkowska, J. Raczaszek-Leonardi, and P. Tomalski, "Combining recurrence analysis and automatic movement extraction from video recordings to study behavioral coupling in face-to-face parent-child interactions," *Frontiers in Psychology*, vol. 8, p. 2228, 2017.

[29] D. Y. Isaev, M. Sabatos-DeVito, J. M. Di Martino, K. Carpenter, R. Aiello, S. Compton, N. Davis, L. Franz, C. Sullivan, G. Dawson, *et al.*, "Computer vision analysis of caregiver–child interactions in children with neurodevelopmental disorders: A preliminary report," *Journal of Autism and Developmental Disorders*, pp. 1–12, 2023.

[30] A. L. Bey, M. Sabatos-DeVito, K. L. Carpenter, L. Franz, J. Howard, S. Vermeer, R. Simmons, J. D. Troy, and G. Dawson, "Automated videottracking of autistic children's movement during caregiver-child interaction: an exploratory study," *Journal of Autism and Developmental Disorders*, pp. 1–13, 2023.

[31] J. Li, M. Hasegawa-Johnson, and N. L. McElwain, "Analysis of acoustic and voice quality features for the classification of infant and mother vocalizations," *Speech Communication*, vol. 133, pp. 41–61, 2021.

[32] Y. Kim, H. Chen, S. Algohwinem, C. Breazeal, and H. W. Park, "Joint engagement classification using video augmentation techniques for multi-person HRI in the wild," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 698–707, 2023.

[33] L. Klein, V. Ardulov, Y. Hu, M. Soleymani, A. Gharib, B. Thompson, P. Levitt, and M. J. Matarić, "Incorporating measures of intermodal coordination in automated analysis of infant-mother interaction," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 287–295, 2020.

[34] H. Li, J. Cui, L. Wang, and H. Zha, "Infant attachment prediction using vision and audio features in mother-infant interaction," in *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*, pp. 489–502, Springer, 2020.

[35] H. Alsofyani and A. Vinciarelli, "Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 221–228, 2021.

[36] J. Hammack, M. Sharma, L. Riera-Gomez, H. Z. Gvirts, and T. Wilcox, "When i move, you move: Associations between automatic and person-coded measures of infant-mother synchrony during free-play using virtual in-home data collection," *Infant Behavior and Development*, vol. 72, p. 101869, 2023.

[37] Q. Chen, R. Abu-Zhaya, A. Seidl, and F. Zhu, "CNN based touch interaction detection for infant speech development," in *2019 IEEE*

Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 20–25, IEEE, 2019.

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

[39] M. Fraile, C. Fawcett, J. Lindblad, N. Sladoje, and G. Castellano, "End-to-end learning and analysis of infant engagement during guided play: Prediction and explainability," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 444–454, 2022.

[40] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg, "Detecting gaze towards eyes in natural social interactions and its use in child assessment," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 3, pp. 1–20, 2017.

[41] P. Li, H. Lu, R. W. Poppe, and A. A. Salah, "Automated detection of joint attention and mutual gaze in free play parent-child interactions," in *Companion Publication of the 25th International Conference on Multimodal Interaction*, pp. 374–382, 2023.

[42] A. Jebeli, L. K. Chen, K. Guerrerio, S. Papparotto, L. Berlin, and B. J. Harden, "Quantifying the quality of parent-child interaction through machine-learning based audio and video analysis: Towards a vision of ai-assisted coaching support for social workers," *ACM Journal on Computing and Sustainable Societies*, 2023.

[43] P. Ekman, "Facial expressions of emotion: New findings, new questions," *Psychological Science*, vol. 3, no. 1, pp. 34–38, 1992.

[44] H. Oster, "Baby FACS: Facial action coding system for infants and young children," *Unpublished monograph and coding manual. New York University*, 2006.

[45] I. Onal Ertugrul, Y. A. Ahn, M. Bilalpur, D. S. Messinger, M. L. Speltz, and J. F. Cohn, "Infant AFAR: Automated facial action recognition in infants," *Behavior research methods*, vol. 55, no. 3, pp. 1024–1035, 2023.

[46] N. Zaker, M. H. Mahoor, D. S. Messinger, and J. F. Cohn, "Jointly detecting infants' multiple facial action units expressed during spontaneous face-to-face communication," in *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 1357–1360, IEEE, 2014.

[47] S. Yan, W. Zheng, C. Tang, Y. Zong, N. Qiu, and X. Ke, "ARL-IL for automatic facial expression recognition of infants under 24 months of age," *Journal of Physics: Conference Series*, vol. 1518, no. 1, p. 012027, 2020.

[48] M. Ning, I. O. Ertugrul, D. S. Messinger, J. F. Cohn, and A. A. Salah, "Automated emotional valence estimation in infants with stochastic and strided temporal sampling," in *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8, IEEE, 2023.

[49] R. Burgess, I. Culpin, I. Costantini, H. Bould, I. Nabney, and R. M. Pearson, "Quantifying the efficacy of an automated facial coding software using videos of parents," *Frontiers in Psychology*, vol. 14, p. 1223806, 2023.

[50] U. Pauli-Pott and B. Mertesacker, "Affect expression in mother–infant interaction and subsequent attachment development," *Infant Behavior and Development*, vol. 32, no. 2, pp. 208–215, 2009.

[51] D. A. Sauter, N. M. McDonald, D. N. Gangi, D. S. Messinger, *et al.*, "Nonverbal expressions of positive emotions," *Handbook of Positive Emotions*, pp. 179–198, 2014.

[52] D. Messinger and A. Fogel, "The interactive development of social smiling," *Advances in Child Development and Behavior*, vol. 35, pp. 327–366, 2007.

[53] V. Wörmann, M. Holodynski, J. Kärtner, and H. Keller, "The emergence of social smiling: The interplay of maternal and infant imitation during the first three months in cross-cultural comparison," *Journal of Cross-Cultural Psychology*, vol. 45, no. 3, pp. 339–361, 2014.

[54] D. S. Messinger, M. H. Mahoor, S.-M. Chow, and J. F. Cohn, "Automated measurement of facial expression in infant–mother interaction: A pilot study," *Infancy*, vol. 14, no. 3, pp. 285–305, 2009.

[55] J. Mesman, M. H. van IJzendoorn, and M. J. Bakermans-Kranenburg, "The many faces of the still-face paradigm: A review and meta-analysis," *Developmental Review*, vol. 29, no. 2, pp. 120–162, 2009.

[56] S. J. Ahn, J. Bailenson, J. Fox, and M. Jabon, "Using automated facial expression analysis for emotion and behavior prediction," *The Routledge Handbook of Emotions and Mass Media*, pp. 349–367, 2010.

[57] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[58] R. Poppe, "Automatic analysis of bodily social signals," in *Social Signal Processing* (J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, eds.), ch. 12, p. 155–167, Cambridge: Cambridge University Press, 2017.

[59] R. Poppe, S. Van der Zee, D. Heylen, and P. Taylor, "AMAB: Automated measurement and analysis of body motion," *Behavior Research Methods*, vol. 46, pp. 625–633, 2014.

[60] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[61] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306, 2018.

[62] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7093–7102, 2020.

[63] M. M. E. Yurtsever and S. Eken, "BabyPose: real-time decoding of baby's non-verbal communication using 2D video-based pose estimation," *IEEE Sensors Journal*, vol. 22, no. 14, pp. 13776–13784, 2022.

[64] A. Dechemi, V. Bhakri, I. Sahin, A. Modi, J. Mestas, P. Peiris, D. E. Barrundia, E. Kokkoni, and K. Karydis, "Babynet: A lightweight network for infant reaching action recognition in unconstrained environments to support future pediatric rehabilitation applications," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, pp. 461–467, IEEE, 2021.

[65] C. Cox, C. Bergmann, E. Fowler, T. Keren-Portnoy, A. Roepstorff, G. Bryant, and R. Fusaroli, "A systematic review and bayesian meta-analysis of the acoustic features of infant-directed speech," *Nature Human Behaviour*, vol. 7, no. 1, pp. 114–133, 2023.

[66] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

[67] X. Yao, D. He, T. Jing, and K. de Barbaro, "Measuring mother-infant emotions by audio sensing," *arXiv preprint arXiv:1912.05920*, 2019.

[68] M. Jover and M. Gratier, "Toward a multimodal and continuous approach of infant-adult interactions," *Interaction Studies*, vol. 24, no. 1, pp. 5–47, 2023.

[69] B. Yang, J. Cui, Y. Tong, L. Wang, and H. Zha, "Recognition of infants' gaze behaviors and emotions," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3204–3209, IEEE, 2018.

[70] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3437–3443, IEEE, 2005.

[71] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pp. 1–6, IEEE, 2020.

[72] N. H. Frijda and D. Moffat, "A model of emotions and emotion communication," in *Proceedings of 1993 2nd IEEE International Workshop on Robot and Human Communication*, pp. 29–34, IEEE, 1993.

[73] G. Gergely, "The role of contingency detection in early affect–regulative interactions and in the development of different types of infant attachment," *Social Development*, vol. 13, no. 3, pp. 468–478, 2004.

[74] L. Crucianelli, L. Wheatley, M. L. Filippetti, P. M. Jenkinson, E. Kirk, and A. K. Fotopoulou, "The mindedness of maternal touch: An investigation of maternal mind-mindedness and mother-infant touch interactions," *Developmental Cognitive Neuroscience*, vol. 35, pp. 47–56, 2019.

[75] Q. Chen, H. Li, R. Abu-Zhaya, A. Seidl, F. Zhu, and E. J. Delp, "Touch event recognition for human interaction," *Electronic Imaging*, vol. 28, no. 11, pp. 1–6, 2016.

[76] N. J. Emery, "The eyes have it: the neuroethology, function and

evolution of social gaze," *Neuroscience & Biobehavioral Reviews*, vol. 24, no. 6, pp. 581–604, 2000.

[77] L. Schilbach, "Eye to eye, face to face and brain to brain: novel approaches to study the behavioral dynamics and neural mechanisms of social interactions," *Current Opinion in Behavioral Sciences*, vol. 3, pp. 130–135, 2015.

[78] N. E. Dunbar, J. K. Burgoon, and K. Fujiwara, "Automated methods to examine nonverbal synchrony in dyads," *Understanding Social Behavior in Dyadic and Small Group Interactions*, pp. 204–217, 2022.

[79] F. J. Bernieri, J. S. Reznick, and R. Rosenthal, "Synchrony, pseudosynchrony, and dissynchrony: measuring the entrainment process in mother-infant interactions," *Journal of personality and social psychology*, vol. 54, no. 2, pp. 243–253, 1988.

[80] J. R. Terven, B. Raducanu, M.-E. Meza, and J. Salas, "Evaluating real-time mirroring of head gestures using smart glasses," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 60–68, 2015.

[81] M. Verde-Cagiao, C. Nieto, and R. Campos, "Mother-infant coregulation from 0 to 2 years: The role of copy behaviors. a systematic review," *Infant Behavior and Development*, vol. 68, p. 101749, 2022.

[82] T. Horowitz-Kraus and C. Gashri, "Multimodal approach for characterizing the quality of parent–child interaction: A single synchronization source may not tell the whole story," *Biology*, vol. 12, no. 2, p. 241, 2023.

[83] M. Chetouani, E. Delaherche, G. Dumas, and D. Cohen, "Interpersonal synchrony: From social perception to social interaction," in *Social Signal Processing* (J. K. Burgoon, N. Magnenat-Thalmann, M. Pantic, and A. Vinciarelli, eds.), p. 202–212, Cambridge University Press, 2017.

[84] I. Egmose, G. Varni, K. Cordes, J. Smith-Nielsen, M. S. Væver, S. Køppe, D. Cohen, and M. Chetouani, "Relations between automatically extracted motion features and the quality of mother-infant interactions at 4 and 13 months," *Frontiers in Psychology*, vol. 8, p. 2178, 2017.

[85] P. J. Sullivan, K. Rickers, and K. L. Gammage, "The effect of different phases of synchrony on pain threshold," *Group Dynamics: Theory, Research, and Practice*, vol. 18, no. 2, pp. 122–128, 2014.

[86] D. M. Messinger, P. Ruvolo, N. V. Ekas, and A. Fogel, "Applying machine learning to infant interaction: The development is in the details," *Neural Networks*, vol. 23, no. 8-9, pp. 1004–1016, 2010.

[87] H. Yoo, D. A. Bowman, and D. K. Oller, "The origin of protoconversation: An examination of caregiver responses to cry and speech-like vocalizations," *Frontiers in Psychology*, vol. 9, p. 1510, 2018.

[88] N. C. Onland-Moret, J. E. Buizer-Voskamp, M. E. Albers, R. M. Brouwer, E. E. Buimer, R. S. Hessels, R. de Heus, J. Huijding, C. M. Junge, R. C. Mandl, *et al.*, "The youth study: Rationale, design, and study procedures," *Developmental Cognitive Neuroscience*, vol. 46, p. 100868, 2020.

[89] X. Huang, N. Fu, S. Liu, and S. Ostadabbas, "Invariant representation learning for infant pose estimation with small data," in *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, 2021.

[90] N. Zaker, M. H. Mahoor, W. I. Mattson, D. S. Messinger, and J. F. Cohn, "A comparison of alternative classifiers for detecting occurrence and intensity in spontaneous facial expression of infants with their mothers," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, IEEE, 2013.

[91] C. Tang, W. Zheng, Y. Zong, Z. Cui, N. Qiu, S. Yan, and X. Ke, "Automatic smile detection of infants in mother-infant interaction via CNN-based feature learning," in *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, pp. 35–40, 2018.

[92] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3359–3368, 2018.

[93] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pp. 211–214, IEEE, 2014.

[94] A. Stergiou and R. Poppe, "Analyzing human–human interactions: A survey," *Computer Vision and Image Understanding*, vol. 188, p. 102799, 2019.

[95] C. Beyan, A. Vinciarelli, and A. Del Bue, "Face-to-face co-located human-human social interaction analysis using nonverbal cues: A survey," *ACM Computing Surveys*, pp. 1–41, 2023.

[96] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[97] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder—a personalized computational modeling approach," *Frontiers in Robotics and AI*, p. 43, 2020.

[98] M. Faltyn, J. E. Krzeczkowski, M. Cummings, S. Anwar, T. Zeng, I. Zahid, K. O.-B. Ntow, and R. J. Van Lieshout, "Coding infant engagement in the face-to-face still-face paradigm using deep neural networks," *Infant Behavior and Development*, vol. 71, p. 101827, 2023.

[99] A. Mahdhaoui and M. Chetouani, "Understanding parent-infant behaviors using non-negative matrix factorization," *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues: Third COST 2102 International Training School, Caserta, Italy, March 15-19, 2010, Revised Selected Papers*, pp. 436–447, 2011.

[100] T. Yarkoni and J. Westfall, "Choosing prediction over explanation in psychology: Lessons from machine learning," *Perspectives on Psychological Science*, vol. 12, no. 6, p. 1100–1122, 2017.

[101] J. T. Grace, *The assessment of the mother-newborn interaction*. University of Rochester, 1989.

[102] M. Licata, M. Paulus, C. Thoermer, S. Kristen, A. L. Woodward, and B. Sodian, "Mother–infant interaction quality and infants' ability to encode actions as goal-directed," *Social Development*, vol. 23, no. 2, pp. 340–356, 2014.

[103] M. Hoogenhout and S. Malcolm-Smith, "Theory of mind predicts severity level in autism," *Autism*, vol. 21, no. 2, pp. 242–252, 2017.

[104] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, *et al.*, "Decoding children's social behavior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3414–3421, 2013.

[105] J. Li, A. Bhat, and R. Barmaki, "Improving the movement synchrony estimation with action quality assessment in children play therapy," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 397–406, 2021.

[106] H. Chen, S. M. Alghowinem, S. J. Jang, C. Breazeal, and H. W. Park, "Dyadic affect in parent-child multi-modal interaction: Introducing the DAMI-P2C dataset and its preliminary analysis," *IEEE Transactions on Affective Computing*, pp. 3345–3361, 2022.

[107] S. Jayaraman, C. M. Fausey, and L. B. Smith, "Why are faces denser in the visual experiences of younger than older infants?," *Developmental Psychology*, vol. 53, no. 1, pp. 38–49, 2017.