The AI4Autism Project: A Multimodal and Interdisciplinary Approach to Autism Diagnosis and Stratification

Samy Tafasca
stafasca@idiap.ch
Idiap Research Institute, Martigny

Anshul Gupta
agupta@idiap.ch
Idiap Research Institute, Martigny

Nada Kojovic
Nada.Kojovic@unige.ch
University of Geneva, Geneva

Mirko Gelsomini
mirko.gelsomini@supsi.ch
SUPSI, Manno

Thomas Maillart
thomas.maillart@unige.ch
University of Geneva, Geneva

Michela Papandrea
michela.papandrea@supsi.ch
SUPSI, Manno

Jean-marc Odobez
odobez@idiap.ch
Idiap Research Institute, Martigny

Marie Schaer
Marie.Schaer@unige.ch
University of Geneva, Geneva

ABSTRACT

Nowadays, 1 in 36 children is diagnosed with autism spectrum disorder (ASD) according to the Centers for Disease Control and Prevention (CDC) [52], which makes this condition one of the most prevalent neurodevelopmental disorders. For children on the autism spectrum who face substantial developmental delays, the trajectory of their cognitive growth can be markedly improved by interventions if the condition is identified early. Therefore, there is a critical need for more scalable screening and diagnostic tools, as well as the need to improve phenotyping to refine estimates of ASD symptoms in children. Here, we introduce AI4Autism: a 4-year project funded by the Swiss National Science Foundation, which aims to address the needs outlined above. In this project, we examine the potential of digital sensing to provide automated measures of the extended autism phenotype. This is accomplished using multimodal techniques based on computer vision and Internet of Things sensing, for the purpose of stratifying autism subtypes in ways that would allow for precision medicine. We present an overview of our main results so far, introducing datasets and annotations that we intend to make publicly available, as well as methods and algorithms for analyzing children’s behaviors and producing an ASD diagnosis.

KEYWORDS

Autism, computer vision, internet of things, datasets, neural networks, gaze detection, behavior analysis.

ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ICMI ’23 Companion, October 9–13, 2023, Paris, France
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8321-8/23/10...
https://doi.org/10.1145/3610661.3616239

1 INTRODUCTION

Autism is a prevalent lifelong neurodevelopmental disorder characterized by deficits in communication (e.g. language delay) and reciprocal social interactions (e.g. turn-taking in conversations), and by the presence of restricted and repetitive behaviors and interests (e.g. hand-flapping) [1]. The term spectrum in ASD is used to refer to the wide range of possible symptoms and their perceived severity. The level of intellectual and adaptive functioning is highly variable, as well as the level of functioning in the different areas of life in adulthood (e.g. relationships, education, employment).

Autism often becomes evident in the first 3 years of childhood, although it can sometimes remain undiagnosed until adulthood when the symptoms are more subtle. For children with a larger developmental delay, getting a proper and early diagnosis is of critical importance for access to early intervention programs that have a tremendous impact on the child’s long-term outcome [27]. It is indeed now widely recognized that early intervention has the potential to substantially transform the future of affected children [27, 59, 77, 93], and that there is a direct relationship between the age at the start of the intervention and the outcome [93]. Even in the absence of a specialized intervention program, an earlier diagnosis is associated with a significantly better outcome, as parents can learn how to optimally support the development of their child.

In addition to screening, there is also an important need to better understand and delineate the whole spectrum of ASD. Clinicians currently lack insights and methods to best stratify between different autism sub-types, which could then be associated with a different prognosis or sensitivity to treatment [6]. This motivates the need for having a more granular strategy for autism characterization, taking into account the intricacies of symptoms, hence moving towards precision medicine, i.e. tailoring intervention strategies to the specific characteristics of each individual [20].

Recent developments in digital sensing and machine learning have offered opportunities for seamless sensing of body movement, social scene capture, and measure of object manipulation. Despite considerable efforts, most studies in computational behavior coding and digital phenotyping for autism have suffered from the following...
Aim 1: better stratification of ASD subtypes. From a clinical perspective, one goal is to study whether the fine-grained quantification of the child’s behavior in a controlled setting can actually be exploited for better stratifying different autism sub-types, which is not possible with the current gold standard [35]. We expect that such stratification that goes beyond current clinical approaches can trigger the discovery of previously undetected autism sub-types, associated with a different prognosis, sensitivity to treatment, or neuro-biological mechanisms.

Aim 2: build a large-scale curated database of manual annotations of behaviors in preschoolers with autism. Intensive manual annotations still represent the only way to obtain a fine-grained quantification of specific autism symptoms. Drawing inspiration from the ADOS coding protocol, we set out to perform intensive fine-grained annotations of non-verbal behaviors, like joint attention, gestures, coordination of visual contact with other nonverbal behaviors, or play behaviors. Such annotations will be used for the study in aim 1, and serve as a basis for training computer-vision and IoT ASD profiling machine learning approaches (aim 3).

Aim 3: designing tools for automated ASD identification and digital phenotyping in young children under controlled conditions. Considering the need for robust automated and unbiased screening approaches, we investigate interpretable multimodal deep learning techniques (computer vision and IoT) for the recognition of ASD behavioral patterns related to free play activities and the coordination between gaze and posture/gestures, and obtain an automated classification between ASD and non-ASD, as well as a digital ASD profile. To build these tools, we will primarily work on data collected to analyze specific behaviors in clinical conditions, following specific assessment structures (ADOS), based on study samples that are clearly either ASD or typically developing (TD). Different methodological frameworks will be investigated here, as described in the next subsections.

Aim 4: validating and generalizing models and tools for ASD screening at large. While clinical data represents an important scientific step to obtain relevant datasets with annotations for both ASD profile and ground-truth behaviors, our aim is to design models working in more challenging conditions, from several perspectives: sensing, with more versatile recording structures such as mobile cameras or IoT set-ups and variable viewpoints or lighting conditions; different assessment protocols (less structured, shorter, etc); or less clearly defined clinical populations, e.g. children suspected of having autism seeking diagnosis in an outpatient clinic. These individuals might have autistic traits, or other comorbid conditions, departing from clear-cut research samples composed of well-separated positives and negatives. To demonstrate the scalability of our approach, we aim to validate models in the “real life” conditions of an outpatient clinic or directly in daycare centers.

2 PROJECT OBJECTIVES

The AI4Autism project has been designed to develop tools for automated behavioral coding and digital phenotyping of autism. The objective is to extract and comprehend the most relevant features for detecting non-verbal social and motor behaviors, which are key for diagnosing autism [2, 15, 30, 72, 91]. This approach is particularly suited for screening very young children (age 1 to 4), as the features that distinguish children with ASD from their typically developing peers are largely non-verbal at this age [26, 50, 51]. Our main aims are the following:

**Aim 1: better stratification of ASD subtypes.** From a clinical perspective, one goal is to study whether the fine-grained quantification of the child’s behavior in a controlled setting can actually be exploited for better stratifying different autism sub-types, which is not possible with the current gold standard [35]. We expect that such stratification that goes beyond current clinical approaches can trigger the discovery of previously undetected autism sub-types, associated with a different prognosis, sensitivity to treatment, or neuro-biological mechanisms.

**Aim 2: build a large-scale curated database of manual annotations of behaviors in preschoolers with autism.** Intensive manual annotations still represent the only way to obtain a fine-grained quantification of specific autism symptoms. Drawing inspiration from the ADOS coding protocol, we set out to perform intensive fine-grained annotations of non-verbal behaviors, like joint attention, gestures, coordination of visual contact with other nonverbal behaviors, or play behaviors. Such annotations will be used for the study in aim 1, and serve as a basis for training computer-vision and IoT ASD profiling machine learning approaches (aim 3).

**Aim 3: designing tools for automated ASD identification and digital phenotyping in young children under controlled conditions.** Considering the need for robust automated and unbiased screening approaches, we investigate interpretable multimodal deep learning techniques (computer vision and IoT) for the recognition of ASD behavioral patterns related to free play activities and the coordination between gaze and posture/gestures, and obtain an automated classification between ASD and non-ASD, as well as a digital ASD profile. To build these tools, we will primarily work on data collected to analyze specific behaviors in clinical conditions, following specific assessment structures (ADOS), based on study samples that are clearly either ASD or typically developing (TD). Different methodological frameworks will be investigated here, as described in the next subsections.

**Aim 4: validating and generalizing models and tools for ASD screening at large.** While clinical data represents an important scientific step to obtain relevant datasets with annotations for both ASD profile and ground-truth behaviors, our aim is to design models working in more challenging conditions, from several perspectives: sensing, with more versatile recording structures such as mobile cameras or IoT set-ups and variable viewpoints or lighting conditions; different assessment protocols (less structured, shorter, etc); or less clearly defined clinical populations, e.g. children suspected of having autism seeking diagnosis in an outpatient clinic. These individuals might have autistic traits, or other comorbid conditions, departing from clear-cut research samples composed of well-separated positives and negatives. To demonstrate the scalability of our approach, we aim to validate models in the “real life” conditions of an outpatient clinic or directly in daycare centers.

3 RESEARCH DIRECTIONS

Several key elements need to be investigated to address the above aims. The first one, which we discuss in Section 4, is related to the creation of useful annotated datasets to enable clinical research, developing machine learning methods, and open science. The second one is the overall strategies for designing clinical and computational methods. In the following, we introduce the main limitations of the state-of-the-art from different perspectives and present the general approach we take to address them.
3.1 Clinical Perspective

Practices and limitations in autism screening and profiling. Autism spectrum disorder is a clinical diagnosis that relies on criteria described in the Diagnostic and Statistical Manual of Mental Disorders [1]. Early symptoms usually include altered modulation of eye contact and diminished engagement in social interactions, whereas non-social symptoms include altered quality of play, repetitive behaviors, mannerisms, and sensory issues. As the child develops language, he/she can often express himself with stereotyped language, echolalia, altered prosody, and show difficulties in turn-taking conversations. In short, ASD symptoms can present in a large variety of ways, depending on age and cognitive abilities.

Regarding screening, approaches largely rely on parent questionnaires given by primary healthcare practitioners, which are thus influenced by the subjective assessments in parents’ reports [82], yielding suboptimal screening sensitivity [10]. Furthermore, their use depends on the level of training of these practitioners [25, 40]. To ensure unbiased access for each child, several authors advocate that screening should happen directly in childcare centers rather than in healthcare practices [23, 41, 46, 89]. This would, however, imply large-scale training of daycare workers. In this context, behavioral sensing tools clearly hold a promise to support autism screening [21, 75], in an automated, unbiased, and scalable manner.

Another critical challenge is to provide clinicians with relevant autism sub-types that could be associated with different prognoses [6]. Currently, the best clinical and research practice for autism phenotyping relies on an observational scale, the Autism Diagnostic Observation Schedule (the ADOS-2, [50]). It consists of a 45-60 min observational assessment completed by a highly-trained clinician that aims at creating a playful environment to elicit social behaviors, and rating behaviors on a scale of 0 to 3 following very strict and standardized procedures. While extremely useful for clinical practice standardization, the relatively coarse symptom granularity measure prevents and restricts the use of these tools to monitor the change over time. Therefore, digital tools for extended phenotyping carry the promise to gather more comprehensive sets of measures for autism symptoms on a continuous scale.

ASD digital stratification. As stated above, the current gold-standard diagnostic assessment ADOS relies on manual coding by expert clinicians, and only provides a semi-quantitative assessment of autism symptoms. Following the deep phenotyping principle, which aims to gather information about disease manifestations in a more individual and granular way [22, 87], we can achieve a more fine-grained quantification of these symptoms. This approach holds the potential for delineating autism sub-types and developing personalized medicine for autistic individuals [7]. Currently, such fine-grained descriptions of autism symptoms however largely rely on labor-intensive manual annotations by experts. A better delineation could help to provide more individualized intervention strategies for autistic children. Indeed, current intervention strategies in younger autistic children rely on play-based behavioral intervention. Despite the general positive gain associated with such methods [27, 78], their benefits remain variable from one child to another [64, 85]. There is currently little knowledge on which type of intervention better works for each child [84]. We thus critically need to identify relevant autism subgroups that better respond to each intervention type. In this endeavor, we advocate deep phenotyping, in particular the one provided by digital phenotyping, which has the potential to provide us with meaningful autism subgroups. We thus aim to develop a comprehensive stratification strategy for automated quantification of a large variety of autistic symptoms, with a particular focus on preschool age.

3.2 Computational Perspective

Computer vision and behavior recognition methodology. The use of computer vision to obtain objective and quantitative standardized observations supporting the diagnosis of ASD children is relatively young. Earlier work used to rely on high-end screen-based eye trackers [43, 44, 67, 73]. Recently, research works have emerged to analyze children’s behavior in more ecological settings, including real interactions [72]. Due to the prevalence of attention and emotional deficit in ASD children, a vast majority of research has been dedicated to the measurement of these two cues [2, 16, 65, 72]. Most earlier works used head pose as a proxy for gaze [18] and were facing the sensor placement issue with moving children. To alleviate this problem, wearable sensors have been proposed either by instrumenting the child with an eye-tracking sensor [58], facing issues with the compliance of the kids, or using a sensor worn by the experimenter/clinician [16], allowing for the extraction of more accurate attentional cues (e.g. eye contact). Regarding motor gesture analysis for ASD, [92] is the only work we are aware of, but it relies on a very contrived diagnosis pipeline (i.e. voice prompt, imposed starting hand pose and position), and with school-aged children (i.e. 10 years old).

In terms of settings, most research targeted standard face-to-face clinician-child or parent-child interaction sessions. Other diagnosis settings include social robots, seen as simpler interaction partners that can elicit desired behaviors to train or test children’s responses [2]; Kinect sensors to measure social deficits in children with ASD [11]; or as an attempt to dramatically scale the screening process of ASD for children, mobile healthcare sensors combining affective, attention, and basic name-call stimuli head rotation reactions [39]. None of the above has considered natural interactions in open settings like ADOS, as we do.

Investigated strategies. In this project, we investigate two different approaches for autism analysis, as shown in Fig 2:

1. (1) one-stage detection from raw inputs;
2. (2) a compositional approach which first identifies (or exploits as a supervision signal) low-level cues (e.g. gestures, gaze), which can then be used to predict ADOS-2 score items and the ASD diagnosis.

The idea behind the first method is to operate directly on image sequences from ADOS-2 sessions or on standard modalities extracted from them (e.g. pose, depth) using off-the-shelf pre-trained models, and to classify whether the patient is ASD or TD (see Section 5.1 and [45]). While it offers the benefit of avoiding the need for behavioral annotations and discovering specific patterns, this approach sacrifices interpretability. There’s also a risk that models trained this way might rely on contextual elements from the session environment (e.g. the clinician’s attitude toward the child) that could indicate a diagnosis, rather than solely on the child’s observable behaviors. The second approach adopts a two-step process where we first model relevant motor coordination behavioral patterns related to posture, gestures, gaze, attention and
communication, interactions or play activities, which are used in a second stage to infer the ASD binary label or build a relevant ASD profile from the ADOS-2 items. A major benefit is to enable the quantification of these behaviors, which in turn can help uncover semantically meaningful ASD profiles.

**IoT sensing.** One of the methodologies we investigated involves the analysis of play behaviors. Play, in fact, is an instinctive need for humans. It correlates with a healthy development process. Playing development and types of play – functional and symbolic – are strictly related to children’s abilities and development. Many models exist to describe the relationship between children’s play and their development [56, 61]. However, the literature focuses mostly on populations older than two years old [5] with a strong focus on the child’s social, emotional, and cognitive development [14, 86], with little attention devoted to the sensory-motor aspect of play [80].

In this project, we focus on studying children’s play behaviors from the perspective of sensory-motor developments: we target very small children, exploit a non-invasive innovative approach based on toys’ inertial measurements, UWB-based localization, and video analysis, and collect statistically significant data to provide meaningful results directly applicable in clinical practice. This approach allows for a more fine-grained, objective, and automated analysis of play behavior, converging in the recognition of play patterns relevant to the analysis of children’s neurodevelopmental disorders. Our goal is to exploit this IoT-based play analysis methodology as a support for traditional social and health research. It provides the technological tools allowing quantitative answers to questions like: “How and for how long does a child interact with a toy?”, and “How do the child-toy interactions evolve with time?”, hence deriving precise children play behavioral models and measures that can be related to clinical data.

4 DATASETS

Datasets are the cornerstone of our research. In this section, we introduce the different datasets that we have collected, along with their purposes and annotation protocols. Given our aims, on the one hand, we need clinical data in sufficient amounts to assess methods as well as research novel ASD stratifications. To this end, we mainly rely on the Geneva Autism cohort (see description in the introduction and Section 4.1). On the other hand, we need data with behavioral annotations in order to train machine learning models. To address the latter, we rely on both publicly available benchmarks (ChildPlay dataset, Section 4.3) which can be disseminated for research, as well as autism-relevant data, as behaviors exhibited in such cases can be pathological (see sections 4.1).

4.1 UniGe ADOS Dataset

**Motivation.** ADOS, the current gold-standard diagnostic assessment for autism, relies on manual expert coding and provides a semi-quantitative assessment of autism symptoms. As our goal is to investigate the development of a fully automated methodology to better quantify autism symptoms following the principles of deep phenotyping, we aim to more densely annotate such ADOS sessions to investigate our research questions, from both the clinical

Figure 2: Two different approaches for ASD analysis, which can both be trained end-to-end. In green (top), directly inferring the ASD diagnosis and scores from multiple features extracted from the raw video sequences. In blue (bottom), the network architecture includes intermediate modeling of several behavioral cues, which can lead to more interpretable diagnoses.

One Stage

Image(s)

Feature Extraction

Depth Pose Semantic

One Stage

1st Stage

Image(s)

Modeling Behavioral Cues

Gesture Labels Gaze Labels Interaction Labels

1st Stage

2nd Stage

Modeling ASD

ADOS Diagnosis

ADOS Scores

2nd Stage

Modeling ASD

ADOS Diagnosis

ADOS Scores

Figure 2: Two different approaches for ASD analysis, which can both be trained end-to-end. In green (top), directly inferring the ASD diagnosis and scores from multiple features extracted from the raw video sequences. In blue (bottom), the network architecture includes intermediate modeling of several behavioral cues, which can lead to more interpretable diagnoses.
perspective (see ASD stratification in aim 1 and Section 3.1) and behavior computational perspective (see aim 3 and Section 3.2).

**Data Collection.** Children included in this project are assessed within the context of the Geneva Autism Cohort, [31, 47, 68], which employs a comprehensive phenotyping strategy combining standardized clinical tools with neuroimaging and digital phenotyping tools. Within this framework, autistic symptoms are evaluated using the ADOS protocol [26, 50, 51]. The ADOS comprises different modules specifically designed to elicit and observe autism symptoms at various developmental and language levels of individuals. These modules assess approximately 30 specific behaviors and assign scores on a scale ranging from 0 (no evident symptom) to 3 (very strong evidence of symptomatology). The coding process is closely intertwined with the assessment, as clinicians take notes during the evaluation, and the coding progression actively shapes the session flow. The ultimate objective is to obtain a fully scored session based solely on live interaction.

**Annotation Protocol.** Taking inspiration from the ADOS coding protocol, we designed an annotation protocol focusing on behaviors primarily characterized by nonverbal components, which is well suited to be used with computer vision tools. In the ADOS, coding thresholds for each behavior are determined based on their frequency, with more stringent thresholds applied to behaviors that are less prevalent in the general population and vice versa. For instance, for high-frequency behaviors like the initiation of joint attention, the absence of at least one clear example of a three-step joint attention behavior (involving visual contact and/or vocalization or gesture) within a single ADOS session is considered indicative of a symptom in this area. Once a child demonstrates such behavior or multiple repeated instances, the ADOS does not differentiate further in its scoring. With our adapted annotation protocol, our aim is to enhance the measurement granularity by documenting all instances of a given behavior, along with their duration. A video annotation sheet sample is provided in the appendix (see Fig. 9). The annotation is carried out using the tool BORIS [32].

### 4.2 The Geneva Pose for Autism dataset

**Motivation.** To demonstrate the capacity of computer vision tools to support automated identification, we built a dataset using solely the information contained in the body pose of the participants in the social interaction. Note that in this way, the data are anonymized, thus overcoming the sensitive question of personal data protection.

**Data.** For this purpose, we used available ADOS videos collected between 2013 and 2020 in the Geneva Autism Cohort. Sixty-eight children with autism (2.8 ± 0.93 years) and 68 typically developing children (2.55 ± 0.97 years) were included, divided into two equal and balanced training and test sets. We used the multi-person 2D pose estimation OpenPose technology [13] to extract the pose of all people present in the room (child, parent, clinician) and generate videos of only the skeletal keypoints, which formed our dataset (see Fig. 4A and B), which constitutes the data. We refer the interested reader to our previously published work [45] for more details.

**Annotations.** As these videos are coming from the Geneva Autism Cohort, following an ADOS protocol, each video comes from the corresponding behavioral annotations in a 0-3 range, as well as a global ASD vs TD score.
gaze points is highly challenging, if not impossible. To avoid this, our gaze label was defined to include 7 non-overlapping classes to account for special scenarios: inside-frame, outside-frame, gaze-shift, occluded, eyes-closed, uncertain, not-annotated.

We are also exploring extra possible layers of annotation related to gestures and interactions with objects such as holding, operating, and pointing (see right column of Fig. 3). This will allow the joint modeling of gaze and interactions, which is instrumental in analyzing motor-gaze coordination patterns.

**Comparison with the literature.** The Multimodal Dyadic Behavior (MMDB) dataset [72], the Self-Stimulatory Behaviors dataset (SSBD) [69], DREAM [8] and 3D-AD [74] are all datasets meant to tackle different aspects of autism, be it stimulating behaviors (arm flapping, head banging), speech and vocalizations, communicative gestures (e.g. pointing, reaching, etc.), or gaze patterns (e.g. shared attention). However, they are either anonymized, limited in terms of behaviors, or restricted to lab environments (e.g. screening or therapy sessions). In contrast, ChildPlay boasts a higher diversity of scenes, people, gestures, viewing angles, and lighting conditions.

5 METHODS AND ANALYSIS

5.1 End-to-end ASD Detection from Raw Inputs

This approach relates to the first category described in Fig. 2. That is, given only videos of the pose of people, can we directly identify ASD vs TD children without identifying atomic behavioral cues? The method is described in more detail in [45].

**Model architecture.** To this end, we relied on the Geneva pose dataset (see Sec. 4.2). We split the videos of this dataset into 5-second segments to train a neural network. This network relied on a VGG16 convolutional neural network pre-trained on Imagenet to extract high-dimensional semantic features from individual frames, and use them as input to an LSTM temporal model. The trained model was applied to make predictions on 5s video segments of the test set, which were further aggregated over the entire duration of the video for each subject to make a global prediction.

**Results.** Using the above network, we reached a prediction accuracy of 80.9% and a F1 score of 0.82 on the unseen test set (sensitivity: 85.3%, specificity: 76.5%). Further tests on an independent and unbalanced set of 106 preschoolers (105 with ASD) led to a similar level of accuracy (0.81) demonstrating the robustness of our approach. We further observed that the confidence probability for ASD obtained was highly correlated with many clinical parameters (severity of autism symptoms measured with the ADOS, p<0.001; measures of level of daily functioning, p<0.001; cognitive functioning as measured with best estimate IQ, p=0.012). Strikingly, the ASD probability was also highly correlated with many individual items from the ADOS, particularly those related to the coordination between gaze and gestures (see Fig.4C). We further tested how much the prediction accuracy was influenced by the video length, and the results demonstrate that an average 0.7 accuracy is already obtained with 10 min video segments (see Fig.4D). The prediction consistency is also very high across the video of a single individual even with relatively short video segments where for instance, for half of the ASD samples, our method achieves a 100% consistency in prediction based on randomly selected segments summing up to only 10 minutes. These important results demonstrate the feasibility of video-based automated identification of autism symptoms.

5.2 Attention Modeling

**Motivation.** Understanding attention behaviors is a key component for autism diagnosis, and we have investigated this topic. Prior works have focused on estimating gaze directions or proxies for it [34, 49], and from there, people’s Visual Focus of Attention (VFOA), defined as looking at a specific person or object [3, 33, 60, 76], but such methods required access to frontal views of people and/or knowledge of the 3D scene structure and were contrived to specific face-to-face interaction settings. To analyze gaze in more general scenes, we have followed [71], which introduced the Gaze Following task, defined as predicting the gaze target of a person (i.e. defined as 2D coordinates in the image). Predictions of all people can then be further processed to infer their social gaze patterns. Below we detail prior works in this direction, as well as ours.

**Gaze Following from Images.** Typical gaze following methods have a two-branch architecture (such as our approach, see Fig. 5): the first one processes the gaze information of a person (typically, a head crop and the person’s image location), and its output is processed by the second branch along with the full image to identify salient items related to the person’s gaze and generate a heatmap highlighting the candidate visual attention target of the person [17, 36, 42, 48, 57, 90]. More recent works have incorporated additional modalities like temporal information [19], depth [4, 29], or a combination of pose and depth as in our work, which had set the state-of-the-art in gaze following benchmarks [38].
Depth, in particular, is important for ruling out salient items that lie along a person’s 2D line of sight but are not visible to the person in 3D space. As standard gaze following datasets do not come with depth information, some works [4, 29] relied on depth (more specifically, disparity) estimated from pre-trained models [70, 88]. However, due to unknown camera parameters, the latter suffer from stretched and distorted reconstructed 3D scenes (point clouds) which are not so suitable for 3D geometric scene analysis.

To address this issue, we recently proposed [81] a more geometrically consistent approach, leveraging a depth estimation algorithm [63] yielding geometry-preserving point clouds, and learning to predict a 3D gaze vector congruent with the estimated point cloud, and allowing to predict the 3D Field of View (3DFOV) of a person. Fig. 5 presents our approach, while Fig. 6 shows qualitative results of our method on ChildPlay samples.

**Social Gaze Prediction.** Predicting gaze 2D locations is not very informative for autism analysis (and other social tasks), where obtaining attention labels is more relevant. In particular, three social gaze tasks are of general interest:

- **LAEO.** Looking at Each Other (or eye contact): binary label indicating whether a pair of people look at each other;
- **SA.** Shared Attention (or joint attention): binary label indicating whether a pair of people look at the same item;
- **LAH.** Looking at Head: binary label indicating whether a person looks at another person’s face/head.

LAEO was first studied in computer vision by [55]. Their model used geometric information to predict LAEO. Since then, several deep learning-based works have been proposed to solve this task [12, 24, 53, 54], including the use of a gaze following approach [37]. Regarding the SA task, authors in [28] were the first to study it. Their method combined the predicted 2D gaze cones of people in the scene with a heatmap of object region proposals. A more recent work [79] improved over this by directly inferring shared attention from the raw image. Several gaze following methods [19, 83] also evaluated their method on the SA task and seemed to provide better performance than methods meant specifically for shared attention detection (protocols were not exactly the same).

Despite its importance, the LAH task had never been proposed in the context of analyzing general scenes. Indeed, LAH can be viewed as a generalization of LAEO, indicating one-way eye contact. This is relevant for autism diagnosis as children may look at the clinician and not vice-versa, and detecting such cases could help provide an extended characterization of children’s gaze behavior. We thus introduced it in [81], and processed gaze following benchmark datasets to obtain LAH annotations. Sample performance (accuracy of looking or not at a person) is shown in Table 1. We see that models trained on standard gaze following datasets, which consist mostly of adults, provide poor LAH performance for children, and improve significantly after fine-tuning on ChildPlay. This highlights (i) the difficulty of the task, with performance reaching only up to 0.7, and (ii) the need for children-specific datasets like ChildPlay to obtain gaze models more applicable to children.

Finally, note that we recently developed a social gaze prediction methodology, summarized in Fig. 7 for the 3 tasks, which addressed previous limitations by processing all people jointly, leading to state-of-the-art performance.
whose level of granularity is in the order of the milli-second. To
understand of children’s neurodevelopment, only a few
works have investigated the sensory-motor aspect of play in the
IoT literature. Developing an ASD digital phenotyping method calls
for defining manipulation measurement methodologies enabling
an autonomous and unbiased play observation, which is privacy-
preserving and non-intrusive for the children. Additionally, chil-
dren’s gestures and play activities are intrinsically different from
human activities well-studied in literature: the difference between
fine-grained and coarse-grained activity drives the need for new
methodologies for human activity recognition. Below we describe
how we addressed these challenges: measurement of autonomous
play activity and recognition of play-related fine-grained activities.

**Autonomous Play.** Our methodology exploits inertial and indoor-
localization sensors to measure the movements of toys manipulated
by children, allowing the inference of the driving manipulation
activities during play in a privacy-preserving and non-intrusive
manner. The measurements rely on the AutoPlay toys-kit that we
had developed, which includes commonly used toys (e.g. ball, car,
cube, doll, spoon) [9, 30] and extended to incorporate the ADOS toys
(see Fig. 8): in both cases, sensors are embedded within the toys in a
child-safe way. Each toy embeds one sensor node, comprising both
an Inertial Measurement Unit (IMU) and an Ultra-Wide Band sensor
(UWB) recording the 3D position and orientation of the embedding
object in a room. The toys allow for the analysis of the main sensory-
motor classes of play: mouthing, simple manipulation, functional,
relational, and functional-relational [66]. The measurement of play
behaviors with different settings (ADOS augmented toys and ADOS
protocol, AutoPlay toys-kit for autonomous play) allows for the
comparison of play behaviors analysis in autonomous situations
versus in a semi-standardized protocol, and can provide clinicians
with indicators related to children’s motor behavior (i.e. giving
gesture) assessment, and validates the potential scaling of the IoT
based methodology in a non-controlled scenario. The augmented
ADOS setting has been chosen because ADOS is the gold standard
for autism diagnosis. AutoPlay and ADOS data allow us to model
autonomous play behavior by extracting behavioral quantitative
features from the observations and investigate which of them and
of play behavior patterns are specifically associated with autism,
in a way that could help remote screening.

**Fine-Grained Human Activity Recognition.** Human manipulation-
related activities are often composed of sequences of micro-movements
whose level of granularity is in the order of the milli-second. To
recognize those micro-movements and activities, we use supervised
learning methodologies applied to the inertial and 3D indoor lo-
calization (UWB) data collected from the toy-embedded sensors,
using as labels video-based annotation of play activities related
to the children-toys interactions. An initial list of micro-activities
for analyzing play behaviors has been identified [30] through an
initial study involving typically developing children in the age
of 9 to 18 months: functional activities such as drag, stack, and
push; exploratory activities like bite and knock; rotational activities
like overturn and rollover. In [62], we demonstrated the intrinsic
difference between coarse-grained (widely investigated in litera-
ture) and fine-grained (characterizing children’s motor behavior)
human activities, showing the performance drop (around 10% for
f1-score) of the well-established methodologies from the Human
Activity Recognition literature when applied to children play activ-
ities, demonstrating the need for dedicated approaches which we
will further investigate during this project.

**5.3 Toy-Embedded Sensors**

**Motivation.** As mentioned earlier, although play behavior is a
natural demonstration of children’s neurodevelopment, only a few
works have investigated the sensory-motor aspect of play in the
IoT literature. Developing an ASD digital phenotyping method calls
for defining manipulation measurement methodologies enabling
an autonomous and unbiased play observation, which is privacy-
preserving and non-intrusive for the children. Additionally, chil-
dren’s gestures and play activities are intrinsically different from
human activities well-studied in literature: the difference between
fine-grained and coarse-grained activity drives the need for new
methodologies for human activity recognition. Below we describe
how we addressed these challenges: measurement of autonomous
play activity and recognition of play-related fine-grained activities.

**Autonomous Play.** Our methodology exploits inertial and indoor-
localization sensors to measure the movements of toys manipulated
by children, allowing the inference of the driving manipulation
activities during play in a privacy-preserving and non-intrusive
manner. The measurements rely on the AutoPlay toys-kit that we
had developed, which includes commonly used toys (e.g. ball, car,
cube, doll, spoon) [9, 30] and extended to incorporate the ADOS toys
(see Fig. 8): in both cases, sensors are embedded within the toys in a
child-safe way. Each toy embeds one sensor node, comprising both
an Inertial Measurement Unit (IMU) and an Ultra-Wide Band sensor
(UWB) recording the 3D position and orientation of the embedding
object in a room. The toys allow for the analysis of the main sensory-
motor classes of play: mouthing, simple manipulation, functional,
relational, and functional-relational [66]. The measurement of play
behaviors with different settings (ADOS augmented toys and ADOS
protocol, AutoPlay toys-kit for autonomous play) allows for the
comparison of play behaviors analysis in autonomous situations
versus in a semi-standardized protocol, and can provide clinicians
with indicators related to children’s motor behavior (i.e. giving
gesture) assessment, and validates the potential scaling of the IoT
based methodology in a non-controlled scenario. The augmented
ADOS setting has been chosen because ADOS is the gold standard
for autism diagnosis. AutoPlay and ADOS data allow us to model
autonomous play behavior by extracting behavioral quantitative
features from the observations and investigate which of them and
of play behavior patterns are specifically associated with autism,
in a way that could help remote screening.

**Fine-Grained Human Activity Recognition.** Human manipulation-
related activities are often composed of sequences of micro-movements
whose level of granularity is in the order of the milli-second. To
recognize those micro-movements and activities, we use supervised
learning methodologies applied to the inertial and 3D indoor lo-
calization (UWB) data collected from the toy-embedded sensors,
using as labels video-based annotation of play activities related
to the children-toys interactions. An initial list of micro-activities
for analyzing play behaviors has been identified [30] through an
initial study involving typically developing children in the age
of 9 to 18 months: functional activities such as drag, stack, and
push; exploratory activities like bite and knock; rotational activities
like overturn and rollover. In [62], we demonstrated the intrinsic
difference between coarse-grained (widely investigated in litera-
ture) and fine-grained (characterizing children’s motor behavior)
human activities, showing the performance drop (around 10% for
f1-score) of the well-established methodologies from the Human
Activity Recognition literature when applied to children play activ-
ities, demonstrating the need for dedicated approaches which we
will further investigate during this project.

**6 CONCLUSION AND FUTURE WORK**

This paper is an effort to review the AI4Autism project and con-
trast our work with the available literature. In essence, it is an
interdisciplinary research collaboration aiming to develop tools
and strategies to ease the autism screening process, gain a better
understanding of ASD phenotypes, and scale the solutions to less
controlled environments where they can make a tangible impact.

Our work so far has focused on several key areas. First, collecting
the necessary data to enable any further analyses and modeling,
be it datasets annotated with relevant behavioral cues (e.g. gaze,
gestures), or recordings of screening sessions annotated accord-
ing to the ADOS-2 protocol. Second, preliminary attempts at the
end-to-end one-stage prediction of ASD based on raw inputs, and
specifically, pose skeletons. Third, modeling of gaze as an interme-
diate step for ASD recognition. In particular, we have focused on
the gaze-following task and its possible extensions to localize the
gaze target and introduce informative semantic components (e.g.
eye-contact). Lastly, we explored the use of inertial and localization
sensors embedded in toys to study play activities in children from a
sensory-motor perspective.

There are several remaining steps that we plan to tackle next: (i)
finalise data collection and annotations; (ii) on the computational
side, incorporate the temporal component in our gaze methods,
move to gestures and interactions modeling, and leverage IoT data
coupled with visual streams to develop multimodal activity recog-
nition methods; (iii) on the autism side, study ASD predictions from
intermediate cues, and once we have a large enough sample of
ADOS-2 annotations, study how these behavioral quantities shape
patients’ profiles and assess whether distinct clusters emerge. Fi-
nally, we plan to integrate our methods together into a solution to
be tested in a different and more challenging environment.

**Acknowledgement.** This research is supported by the AI4Autism
project (Digital Phenotyping of Autism Spectrum Disorders in chil-
dren, grant agreement no. CR- SII5 202235 / 1) of the Sinergia
interdisciplinary program of the SNSF.


A APPENDIX: ADOS ANNOTATION

Figure 9: A sample annotation sheet covering 5 minutes of an ADOS session. We can see the duration of different coded behaviors (e.g. request using eye contact and vocalization) and ADOS activities (e.g. free play).