Attention Analysis in Robotic-Assistive Therapy for Children With Autism

Bárbara Silva, Laura Santos, *Graduate Student Member, IEEE*, Catarina Barata, Alice Geminiani, Gabriele Fassina, *Graduate Student Member, IEEE*, Ana Rita Gonzalez, Sara Ferreira, Bernardo Barahona-Corrêa, Ivana Olivieri, Alessandra Pedrocchi, *Senior Member, IEEE*, and José Santos-Victor, *Fellow, IEEE*

***Abstract*— Children with Autism Spectrum Disorder (ASD) show severe attention deficits, hindering their capacity to acquire new skills. The automatic assessment of their attention response would provide the therapists with an important biomarker to better quantify their behaviour and monitor their progress during therapy. This work aims to develop a quantitative model, to evaluate the attention response of children with ASD, during robotic-assistive therapeutic sessions. Previous attempts to quantify the attention response of autistic subjects during human-robot interaction tasks were limited to restrained child movements. Instead, we developed an accurate quantitative system to assess the attention of ASD children in unconstrained scenarios. Our approach combines gaze extraction (Gaze360 model) with the definition of angular Areas-of-Interest, to characterise periods of attention towards elements of interest in the therapy environment during the session. The methodology was tested with 12 ASD children, achieving a mean test accuracy of 79.5 %. Finally, the proposed attention index was consistent with the therapists' evaluation of patients, allowing a meaningful interpretation of the automatic evaluation. This encourages the future clinical use of the proposed system.**

***Index Terms*— Attention, Autism spectrum disorder, Gaze tracking, social-assistive robots.**

Manuscript received 5 March 2024; revised 18 April 2024; accepted 4 June 2024. Date of publication 7 June 2024; date of current version 19 June 2024. This work was supported in part by the Fundação para a Ciência e Tecnologia [Foundation of Science and Technology (FCT)] Portuguese FCT under Project SFRH/BD/145040/2019, Project CEECIND/00326/2017, and Project UID/50009/2020 LARSyS and in part by the MUSA-Multilayered Urban Sustainability Action-project, funded by the European Union-NextGenerationEU, under the National Recovery and Resilience Plan (NRRP) Mission 4 Component 2 Investment Line 1.5: Strenghtening of research structures and creation of R&D "innovation ecosystems", set up of "territorial leaders in R&D". *(Bárbara Silva and Laura Santos contributed equally to this work.) (Corresponding author: Laura Santos.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethical Board of Associação Portuguesa de Autismo and Centro de Apoio ao Desenvolvimento Infantil.

Bárbara Silva, Catarina Barata, and José Santos-Victor are with the Institute for Systems and Robotics and the Instituto Superior Técnico, Universidade de Lisboa, 1649-004 Lisbon, Portugal.

Laura Santos is with the Institute for Systems and Robotics and the Instituto Superior Técnico, Universidade de Lisboa, 1649-004 Lisbon, Portugal, and also with the NEARLab, Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20156 Milan, Italy (e-mail: laura.d.santos@tecnico.ulisboa.pt).

Alice Geminiani is with the NEARLab, Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20156 Milan, Italy, and also with the Champalimaud Foundation, 1400-038 Lisbon, Portugal.

Gabriele Fassina and Alessandra Pedrocchi are with the NEARLab, Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20156 Milan, Italy.

Ana Rita Gonzalez is with the CADIn-Centro de Apoio ao Desenvolvimento Infantil, 1050-215 Lisbon, Portugal.

Sara Ferreira is with Associação Portuguesa para as Perturbações de Desenvolvimento e Autismo, 1300-565 Lisbon, Portugal.

Bernardo Barahona-Corrêa is with the Champalimaud Foundation, 1400-038 Lisbon, Portugal, and also with the NOVA Medical School, Faculdade de Ciências Médicas, Universidade NOVA de Lisboa, 1099-085 Lisbon, Portugal.

Ivana Olivieri is with the IRCCS Fondazione Don Carlo Gnocchi, 20148 Milan, Italy.

This article has supplementary downloadable material available at https://doi.org/10.1109//TNSRE.2024.3411299, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2024.3411299

I. Introduction

AUTISM Spectrum Disorder (ASD) is a neurodevelopmental condition with an increasing prevalence in recent years, affecting 1 in every 36 8-year-old children, in the US [1]. It is characterised by impairments in the social and communication domains, along with the presence of repetitive stereotyped behaviours and interests. The deficits and their severity vary significantly across children. Without clear causes for this condition, a cure is still to be found [2]. In order to improve the social and motor abilities of ASD children, several therapeutic approaches have been used. The introduction of Social-Assistive Robots (SARs) in therapies for ASD was recently proposed, as SARs attract the children's interest, thanks to their stylised appearance and their simple repetitive movements [3]. Studies on robot-mediated intervention have demonstrated positive outcomes in different social skills, such as communication, attention, and imitation [4].

A crucial functionality would be the capability to assess the impact of therapies with SARs by evaluating the attention, often compromised in ASD children, namely the on-task attention. It represents the willingness to acquire and develop skills during a task, and it is a major prerequisite for a good performance in the therapy sessions [5]. Therefore, assessing the attention state of each child and monitoring the progress in longitudinal training would be crucial to evaluate therapy and the impact of novel therapy protocols.

Our primary goal is to create an accurate quantitative model, using data extracted from non-intrusive devices, for the evaluation of attention in ASD children during therapeutic sessions with SARs. The sessions are based on unconstrained triadic interactions between the ASD child, the therapist, and the robot [11] (Figure 1). In our specific case, the sessions are focused on gesture training, in which the robot presents

TABLE I
PROPOSED ATTENTION SYSTEMS IN PHYSICAL HUMAN-ROBOT INTERACTION WITH ASD SUBJECTS. CHARACTERISTICS FITTING OUR REQUIREMENTS ARE PRESENTED IN BOLD

Paper	Attention features	Features estimator	Areas-of-Interest	Nº of cameras	Nº of targets	Scenario
[6]	Head pose	Machine Learning	K-means	**1**	**3**	Constrained
[7]	Head pose	ALGazeAnalysis	Range of azimuth angles	**1**	1	Constrained
[8]	Head pose	Machine Learning	K-means	4	**3**	Constrained
[9]	Eye gaze & Head pose	OpenFace	Range of azimuth angles	3	**3**	Constrained
[10]	Eye gaze	Gaze360	Head bounding boxes	**1**	1	**Unconstrained**
Our	Eye gaze	Gaze360	Range of azimuth angles	**1**	**3**	**Unconstrained**

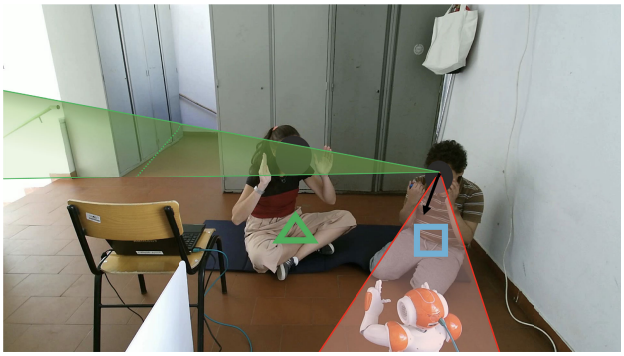


Fig. 1. Attention classification system representation, along with the Therapist (green cone) and NAO robot (red cone) Areas-of-Interest, and the ASD child gaze estimation (black arrow). The green triangle represents the Therapist and the blue square represents the child.

gestures that are then imitated by the child and therapist. This type of sessions implies complex requirements due to the children's and therapy's specific characteristics. First, the use of intrusive devices may disturb ASD children. Therefore, quantitative measures are only obtained with non-intrusive devices, such as cameras [12], posing technical challenges to the attention assessment task, given the distance between the subject face and the tracking devices, sometimes larger than 2 m. Second, the therapy needs to be very unstructured to adapt to the specific conditions of each ASD child. The therapist and the child cannot be constrained to keep specific positions, increasing even more the complexity of the tracking and attention assessment. Both actors can assume any kind of pose creating different types of occlusions.

A secondary goal of this work is to develop an attention analysis model that adheres to the principles of Explainable Artificial Intelligence (AI) [13]. Therefore, the system's outcomes should be interpretable and understandable by therapists, supporting their qualitative evaluation with quantitative data and, thus, enriching their own feedback about the session. Overall, even if the system was constructed for robotic therapy, it could be used for measuring attention in any kind of therapy.

In the rest of the paper, we start with the literature review detailing related works on attention assessment systems and our contributions to tackle their limitations (Section II). Then, we describe the clinical protocol and acquisitions (Section III) used to test our attention assessment system proposed in Section IV. Lastly, we present the obtained results, along with their discussion (Section V), as well as the conclusions and future work (Section VI).

II. STATE OF THE ART

Qualitative measures of attention have been extensively used in research and clinical practice in the ASD field. However, most of these measures are based on manual video processing [14], [15], which is prohibitively time-consuming, operator-dependent, and poorly reliable/accurate. Recent studies have focused on obtaining reliable, objective, and quantitative measures of attention based on the head orientation [6], [8], the detection of facial landmarks [16] and/or the eye gaze direction [9], [10]. To obtain these measures, since ASD children consider physical sensors uncomfortable [12], several non-intrusive sensors have been proposed, like cameras. For example, [7] profit from robot-integrated cameras to evaluate the child's attention during a triadic multi-robot interaction. Although the scenario was completely unconstrained, there was only one child relatively close to the two robots. Moreover, according to the documentation [17], the algorithm's performance decreases with more people in the scene and it does not work for distances larger than 2 m, distances that are common in our setup.

Alternatively, external cameras can be used as we did in our research. In general, in these works, the attention was quantified from the gaze and/or the head pose, following four steps: (i) acquisition of head/eyes videos; (ii) estimation of the head pose and/or the gaze, from the RGB video recordings, using image-based models and obtaining an attention angle; (iii) definition of Areas of Interest (AOIs) based on the objects of interest in the scenario (targets); (iv) attention classification towards each target, through the comparison of the attention angle with the AOIs. A summary of some of these works is presented in Table I and their details, in the next paragraphs.

Concerning the experimental setup, the works in [8] and [9] have the disadvantage of requiring multiple cameras to record the sessions, making the therapy environment more complex and distracting to the subjects during the sessions. The works [6], [8], [9] restrained the ASD subject movements during the experiments, since most gaze estimators, such as OpenFace, do not work/operate when the eyes are partially occluded. Our clinical partners considered these constrained scenarios unrealistic and impractical. Therefore, we decided to let the therapist and the child move freely in the room.

Under our experimental condition, we could leverage on other algorithms, such as the WHENet [18] and Gaze360 [19] models, developed for robust 3D head and gaze estimation, respectively. Gaze360 includes temporal information in the gaze direction predictions and, therefore, produces reliable gaze estimates even when the eyes are not fully visible. The authors apply the Densepose algorithm [20] to crop frames and obtain the bounding boxes around the subjects' heads, independently of their position in the space. Afterwards, the output of each frame passes through bidirectional Long Short-Term Memory cells. In the Gaze360 model, seven frames are used, corresponding to the current frame plus three previous and three subsequent frames. Consequently, even if the eyes of a person are occluded, this model can still estimate

the gaze angles. The Gaze360 model outputs full-range gaze angles, covering 360°, relative to the camera view, using spherical coordinates (azimuth and elevation). In this way, if the subject looks directly to the camera, the output is 0° for the azimuth and 0° for the elevation [19] independently of the subject's position. Given the robustness of this method, we decided to compare it with other gaze and head pose estimators.

Gaze360 has already been successfully applied in ASD children therapy, for example, in [10]. In this work, the authors designed a system (EYE-C) based on OpenPose [21] and Gaze360 for robust detection of eye-contact episodes between the therapist and the ASD child during unconstrained therapeutic sessions, using a single video camera. However, the work is limited to the detection of the gaze towards only one target, since it is based on the 2D images.

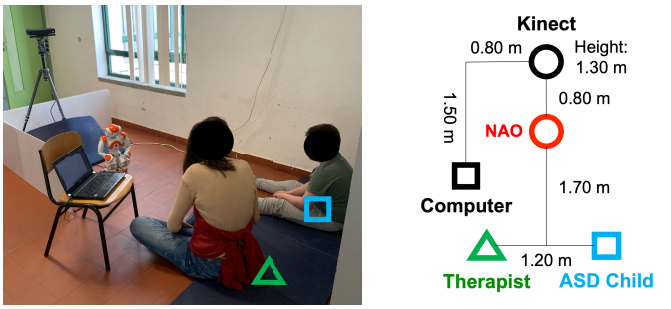
In our setup, we include multiple targets (the robot and the therapist) during the sessions. Therefore, our contributions in this work are:

- an accurate system to quantitatively assess attention. The system identifies the regions where the ASD children look at, during triadic unconstrained robotic therapy;
- attention indexes understandable by the therapists and consistent with their own opinions, aligned with the scope of Explainable AI tools.

III. Clinical Study and Data Collection

The main goal of the therapy in our clinical study was to train gestures during triadic interactions, involving the child, the therapist and the robot. A setup and a protocol for an imitation game were defined based on the clinical knowledge of Associação Portuguesa para as Perturbações do Desenvolvimento e Autismo (APPDA Lisboa), Centro de Apoio ao Desenvolvimento Infantil (CADIn) and Fondazione Don Gnocchi, in collaboration with Instituto Superior Técnico and Politecnico di Milano [11]. In this protocol our clinical partners established as main requirement the simultaneous presence of the robot and the therapist so that the therapist could be a role-model to the child and the child could immediately expand the tasks learned with the robot to the interaction with the therapist. To achieve this goal, the protocol consisted of an imitation game with 4 levels. While the first two levels were conceived to foster the familiarisation with the robot, levels 3 and 4 were intended for gesture training, where the robot performed a gesture while saying a simple sentence, and then the therapist and child mirrored it in a turn-taking exercise. The difference between levels 3 and 4 was the presence of daily life scenarios in the latter.

The clinical acquisitions took place in a school (Escola Básica Bernardim Ribeiro) between May and July 2021 in association with APPDA Lisboa. The participants were five ASD children, four males and one female, between 7 and 11 years old. Four children were diagnosed with level 3 of ASD, while one child was diagnosed with level 1 (less severe than the others), according to the Diagnostic and Statistical Manual of Mental Disorders V [22]. Numbers between 1 and 20 were randomly attributed to identify each of the five children anonymously. The study lasted seven weeks, with each child getting one session of 30 minutes each week. However, the number of sessions carried out by each child varied between 2 and 7, corresponding to their school attendance during the acquisition days.



(a) Therapy session (b) Top view scheme

Fig. 2. Setup representation with the therapist (green triangle), the ASD child (blue square), NAO (red circumference), the computer (black square) and Kinect (black circumference).

The study was revised and approved by APPDA's ethical committee. All participants' parents signed an informed written form, giving their consent for the participation of their children in the study.

A. Experimental Design and Data Collected

The therapy setup consisted of a triangle between the three actors, with the robot placed in the middle between the therapist and the child (Figure 2). The humanoid robot chosen was NAO since it attracts the children's attention, due to its toy appearance and simple and repetitive movements [23]. The therapist was responsible for the robot control using a computer placed near him/her. A non-intrusive Kinect sensor was placed behind NAO to record the sessions and to retrieve the position of 25 3D joints from the therapist and child skeletons. Contrarily to [7], we decided to use an external camera and not the robot camera due to its larger field of view (NAO: $60.9^o \times 47.6^o$ vs Kinect: $84.1^o \times 53.6^o$) and its information about the depth, which is required by our robotic therapy protocol (see [24]). The data acquired through the Kinect camera were saved, frame by frame, during the sessions. These data included the video, the estimated joints positions and the times related to each frame.

The therapy was carried out in the school atrium. Although therapist and child were in a certain part of the atrium, the scenario was considered unconstrained since they could move freely in that area. To label each of the skeletons, the therapist used a specific shirt which was tracked during the sessions.

Given that the first sessions were not recorded and the second ones consisted mainly of familiarisation levels, which were not the primary goal of the therapy, the on-task attention was only studied on Sessions 3 to 7.

IV. Attention Recognition System

To achieve the ultimate goal of evaluating the ASD children's attention in robot-mediated therapy, we developed an attention classification system based on the subject's gaze and the definition of Areas-of-Interest (AOIs) around each target within the therapy environment (Figure 3). Our methodology had five different steps.

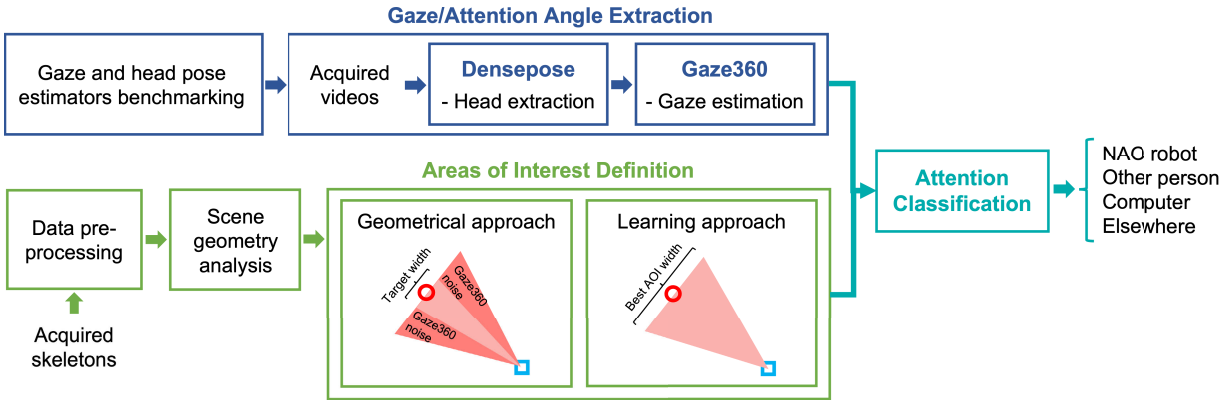


Fig. 3. Full overview of the attention classification system. The system is composed by two main blocks: the attention angle extraction and the Areas of Interest definition (Sections IV-D and IV-E). In the end, we compared the extracted angle with the defined AOIs to classify the attention (Section IV-F). For the choice of the attention estimator, we did an initial benchmark of gaze and head pose estimators (Section IV-A). For the definition of the AOIs, an initial data preprocessing (Section IV-B), followed by the analysis of the scene geometry (Section IV-C) was required.

We performed an initial benchmark in a constrained environment of gaze and head pose estimators to extract the attention angle (Section A). After, we analysed the data of our clinical acquisitions with the best estimator from the benchmark process. Contrarily to the benchmark, this was an unconstrained environment. The data analysis consisted of three parts: the skeleton preprocessing (Section B), the scene geometry analysis (Section C) which consisted on the calculation of the angular direction towards each target in each instant, and the definition of AOIs around each target (Section D). We defined the AOIs as a range of angles corresponding to looking at each target. The range for each target was established according to two approaches. In the geometrical approach, the widths of the AOIs were determined based on the geometry of the targets. In the learning approach, we trained the system, using a grid search approach with several widths for each target, and chose the ones that optimised the system performance. Having both the attention angle and the AOIs, we classified the gaze as "looking at each target (e.g. Robot, Other person)" or "looking elsewhere" being a proxy to the attention classification of each subject (Section F).

We only analysed the azimuth to discriminate the targets in our application, which were in separate horizontal directions during the sessions. Moreover, most head pose estimators are not accurate along the vertical axis (elevation), so this direction is disregarded for the attention assessment [18].

A. Benchmarking Gaze and Head Pose Estimators

We compared Gaze360 [19] (gaze) and WHENet [18] (head) pose estimators, to choose the most adequate one for the attention angle extraction. We selected these two methods as possibilities for our system due to their importance in the current literature. They are both full range (360°) estimators, able to estimate orientations even when the facial features are not visible in the video.

For this benchmarking controlled experiments were performed to collect ground truth data of the fixation points. Three experiments were designed to cover the full field of view and a range of scene distances, consistent with the therapy scenario. In these experiments, the subject was located at about 2.5m from the camera and several visual targets were positioned around the subject at known locations. The subject changed fixation points every 10 seconds after a sound cue.

TABLE II
AVERAGE RMSE OF WHENET AND GAZE360 AZIMUTH ESTIMATES [rad]

	Experiment 1	Experiment 2	Experiment 3
WHENet	**0.64**	0.42	0.43
Gaze360	1.02	0.46	**0.22**

The first experiment consisted of four fixation points around the subject (including one in the back) to incorporate different degrees of occlusion. In the second experiment, four fixation points were on the sides and in front of the subject, with one of the fixation points representing the robot position. In the third experiment, the subject could only move the eyes between three fixation points in front of him, keeping the head still (see Supplementary material for further details). In the end, we obtained the Root Mean Squared Errors (RMSEs) between the estimations and the ground truth, as shown in Table II.

Given the results for the first experiment, we concluded that the WHENet model is more accurate than Gaze360 in estimating the azimuth for the point in the back, when all facial features are occluded. However, when there is no fixation point behind the subject, both models performed similarly (Experiment 2). Experiment 3 proved that WHENet is not suited for our setup, since it can not follow the eye gaze. Overall, the Gaze360 model had a good performance (lowest average RMSE across the 3 experiments) and was chosen to be integrated into our attention system to extract the attention angles.

B. Skeletons' Preprocessing

Since the Kinect's skeleton estimation and tracking performance is noisy, we included a data preprocessing stage. Given the therapy environment, the detection of both (therapist and child) skeletons is extremely challenging, as the child and therapist bodies are frequently in overlap or partially occluded,

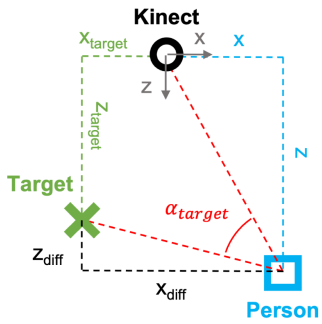


Fig. 4. Standard angle (α_{target}) representation. The green cross represents the target, while the blue square represents the person for which the standard angle is calculated. The origin of the reference frame is located in the centre of the Kinect.

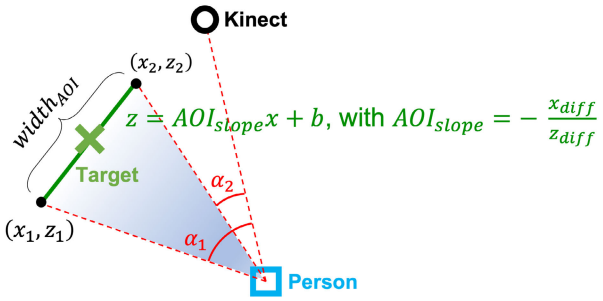


Fig. 5. Representative top view of an AOI. The green cross represents the target, while the blue square represents the person in analysis. The dark green line corresponds to the AOI width and the area in blue to the range of angles (receptive field) for looking at the target.

the participants are usually self-occluded (while sitting down), and the illumination conditions are far from perfect.

First, we discarded the frames in which the Kinect did not detect both skeletons. The skeletons in the remaining frames were filtered using a median filter to eliminate the outliers. After, we applied a linear interpolation to the therapist and child head joints (the only joint used in our work), to reconstruct the missing data.

Then, we compared the Kinect 2D head joints with the Densepose head boxes, outputted by the Gaze360 model (as defined in Section II), to assign the identity of each detected head bounding box (child/therapist). To compensate for the errors from the Kinect skeletons detection, we explored the effect of increasing the Densepose head bounding boxes size by $p \in \{25\%, 50\%, 75\%\}$. After, we checked which head joint (therapist or ASD child) was inside each Densepose bounding box, for each frame. We discarded the frames in which this correspondence was not possible. In this way, we kept only frames with both skeletons and the corresponding Densepose bounding boxes. As a collateral effect, this step also discarded frames in which the interpolation had a considerable error for the head joint.

C. Scene Geometry Analysis

The angular direction towards each target (called standard angles, from now on) was calculated to estimate where the child/therapist should be looking at during the therapy sessions (Figure 4). The targets corresponded to NAO, the Other Person (Therapist for the Child and Child for the Therapist), and the Computer. The computer attracted the ASD children's attention when the therapist chose the exercises and when the scenario images (in level 4) appeared on it. Thus, during levels 1, 2 and 3 the Computer was considered a distraction, while on level 4, it was an additional focus of attention. NAO and the Other Person were always considered focus of attention.

The standard angles (α_{target}) for looking at the different targets were calculated based on geometry and were relative to each person (therapist and ASD child). Since the angles varied according to the therapist/child's positions, they were calculated for every time instant t. The therapist and child 2D positions (x and z) were obtained using the head joints captured by Kinect. Then, two conditions were considered, according to the relative positions of the person and the target, where x_{diff} and z_{diff} are the difference between the 2D coordinates of the person and the target (Figure 4):

a) If the person and the target were at the same side of the camera and the person was closer to the camera both in the x and z axis [$x_{diff}(t)x(t) \leq 0 \wedge z_{diff}(t) < 0$]:

$$\begin{aligned}\alpha_{target}(t) &= n \arctan\left(\frac{x(t)}{z(t)}\right)^n + \\ &+ \arctan\left(\frac{z_{diff}(t)}{x_{diff}(t)}\right)^n + n\frac{x(t)}{|x(t)|}\frac{\pi}{2} \quad , \text{with} \\ &\begin{cases} n = 1, \text{if} \arctan\left(\frac{z(t)}{x(t)}\right) \geq \arctan\left(\frac{z_{target}(t)}{x_{target}(t)}\right) \\ n = -1, \text{if} \arctan\left(\frac{z(t)}{x(t)}\right) < \arctan\left(\frac{z_{target}(t)}{x_{target}(t)}\right). \end{cases}\end{aligned} \tag{1}$$

b) Other situations [$(x_{diff}(t)x(t) > 0) \vee (x_{diff}(t)x(t) \leq 0 \wedge z_{diff}(t) > 0)$]:

$$\alpha_{target}(t) = -\arctan\left(\frac{z(t)}{x(t)}\right) + \arctan\left(\frac{z_{diff}(t)}{x_{diff}(t)}\right). \tag{2}$$

D. Definition of Areas of Interest (AOIs)

After establishing the standard angle (the angular direction towards each target), as in [9], we proceeded to delineate Areas of Interest (AOIs) around each target along the horizontal direction (azimuth). The primary objective was to identify the range of angles indicative of looking at each target (Figure 5).

The AOIs were defined with a process involving five steps:

i) Each AOI was centred in the 2D coordinates of the target and set to be orthogonal to the line connecting the person and the target. The AOI slope was obtained using the equation in Figure 5.

ii) The width of the AOI [m] was defined using two approaches (geometrical and learning), as explained in the Section IV-E.

iii) After, the 2D coordinates of the borders of the AOI were calculated ((x_1, z_1) and (x_2, z_2)) [m], using Equation (3),

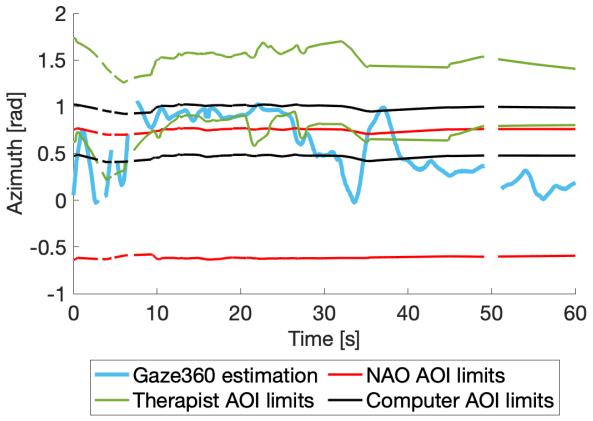


Fig. 6. Gaze360 estimates and AOIs limits before correcting the overlapping, for Child 19 during the first 60s of Session 4. The gaps across time correspond to the frames discarded during the data preprocessing.

with $r = \{1, 2\}$.

$$\begin{cases} x_r = x_{target} + \dfrac{width_{AOI}}{2}\sqrt{\dfrac{1}{1+AOI_{slope}^2}}(-1)^r \\ z_r = z_{target} + AOI_{slope}\dfrac{width_{AOI}}{2}\sqrt{\dfrac{1}{1+AOI_{slope}^2}}(-1)^r. \end{cases} \tag{3}$$

iv) The range of angles $[\alpha_1; \alpha_2]$, corresponding to looking at each AOI was computed $[rad]$. First, we calculated the relative positions between the person and the two extremities of the AOI. Then, we used Equations (1) or (2) to calculate the two angles, depending on the situation.

v) In the end, we corrected the overlapping AOIs, as shown in Figure 6. This process was done frame by frame and was composed of two parts.
First, if an AOI was completely overlapping another AOI, one of them was deleted, according to the scene geometry and the targets' priority (Other Person > NAO > Computer). For example, if the therapist or the child (Other Person) AOI was in the same gaze direction as NAO or the Computer, the AOI of the latter (NAO or Computer) was discarded. These priorities were established based on what was more frequent during the therapy sessions.
Then, whenever two AOIs partially overlapped, we calculated a decision threshold between the AOIs. For each instant, two Gaussian curves (one for each target) were estimated, $N_i(\mu_i, \sigma_i)$, with $i = \{1, 2\}$. For each target, we used Equation (4) to calculate the mean value of the AOI limits at that instant, μ_i. The standard deviation, σ_i, was calculated using an empirical rule. It was defined that half of the AOI width was equal to $k\sigma_i$, with $k \in \{1, 2, 3\}$ (Equation (5)). In this way, according to the empirical rule, 68%, 95%, and 99.7% of the values lie within k standard deviations of the mean.

$$\mu_i = \frac{\alpha_{1_i} + \alpha_{2_i}}{2}. \tag{4}$$

$$k\sigma_i = \frac{width_{AOI_i}}{2}. \tag{5}$$

We defined the limit between the AOIs as the x-value corresponding to the intersection of the two Gaussians distributions.

E. Estimation of AOIs Widths

Regarding step ii) of the AOIs definition, two approaches were studied: the geometrical and the learning approaches.

1) Geometrical Approach: In the geometrical approach, the AOIs' widths corresponded to the targets' dimensions increased by the Gaze360 noise. The widths of NAO and the participants were defined based on the literature. For the therapist and child, we tested the effect of using the same width for both groups or differentiate by groups. Thus, we used just the average shoulders' width (bideltoid) of a female adult (43.26 cm) [25], in the first case, or joined it with the average shoulders' width of the child (32.8 cm) [26] in the second case. The arms were not considered in our setup since they would totally occlude the Computer AOI with this assumption. For NAO, the shoulders' distance and arms' lengths established the AOI width since it used the arms to perform the gestures during the training levels. In this way, the NAO width was defined as $27.5 + 31.1 \times 2 = 89.7$ cm. The Gaze360 noise (0.069 rad) was obtained from the benchmarking experiments (see Section IV-A), corresponding to the standard deviation between the expected signal and the Gaze360 estimates.

2) Learning Approach: In the learning approach, the best widths of each target were automatically estimated using annotated data from a few therapy sessions. A group of annotators was asked to label the video frames, identifying the gaze direction for each subject. This information was used to empirically define the best combination of AOI widths (NAO, Other Person, and Computer), considering the trade-off between recall and false positive rate (FPR) scores. In particular, the best set of widths was defined as the one optimising the value of the Receiving Operating Characteristic (ROC) curve:

$$bestROCscore = min(\sqrt{(1 - Recall)^2 + FPR^2}). \tag{6}$$

The ranges of values tested for the NAO, Other Person and Computer widths varied between $[0.4, 3.0]m$, $[0.4, 2.0]m$, $[0.4, 1.0]m$, respectively, with increments of 0.2 m.

F. Attention Classification

For each instant, we compared the attention angle with the computed AOIs, creating a fixations' signal as attention metric. This signal reflected the attention towards each target or elsewhere. We considered that a person was looking at a target if the classification was constant for more than $400ms$, as indicated in [27]. We removed most of the spurious fixations, applying a median filter to the fixations' signal.

V. Attention Analysis and Results

This Section describes the experimental results obtained from our attention system with the data acquired during the clinical study described in Section III. We start by giving a description of the metrics used for performance evaluation (Section A) and report the results for the proposed attention classification system (Section B). After, we analyse the evolution of the on-task attention across sessions for both

the child and the therapist(Section C), and compare it with the therapist feedback of the same sessions (Section D). In the end we present the results of our proposed attention classification system in a more recent dataset with younger children with ASD, testing the generalizability of the model (Section E).

A. Data and Metrics

The sessions (from 3 to 7, as described in Section III-A) were randomly split into two or three sets depending on the approach used to calculate the targets' widths. In the geometrical approach, Session 6 represented the validation set, and the remaining sessions were the test set. In the learning approach, Session 3 represented the training set, Session 6 the validation set, and the remaining sessions were the test set. In the learning approach, we used the training set to define the best widths for each target. In both approaches, we chose the best model hyperparameters based on the validation set. In the end, we evaluated the model and obtained the model performance scores on the test set.

Two independent annotators labelled the videos from the therapy sessions. Since more than 25 videos were acquired, with some having more than 15,000 frames, acquiring labels for both the therapist and the patient is an extremely fastidious task. Therefore, the videos were only labelled every 3 seconds, reflecting the main changes in terms of fixations at the different targets. For each selected frame, the annotators determined where the therapist and the ASD child were looking at (NAO, Other Person, Computer or Elsewhere). In the end, we kept only the labels from frames where there was an agreement between the annotators. This corresponded to more than 75% of the labels for all the sessions, confirming the good inter-annotator agreement.

To assess the system performance, we compared the ground truth labels with the system estimates obtaining the recognition rate, as well as the false positive/negative rates. This calculation was done for the child and therapist altogether. From this point onward, we will use the recognition rate as the overall accuracy/performance indicator.

Since the ASD patients may have different behaviours from the therapist, the performance metrics were also obtained by group to study the effect of using the same AOI widths for all the people or by groups (Therapist and ASD children) in the learning approach.

B. Attention Recognition Rate

The classification system depends on a set of hyperparameters that were validated by computing the accuracy of the proposed system. The main hyperparameters were the Densepose bounding boxes ratio ($p \in \{25\%, 50\%, 75\%\}$) and variable $k \in \{1, 2, 3\}$ (recall Equation (5)), used to define the standard deviation of the Gaussian curves when two AOIs overlap. The results for the several hyperparameters configurations using the validation set of the APPDA study in the geometrical and learning approaches are reported in Table III, with the best configurations for each approach in bold.

Analysing the hyperparameters effect, the best value for k was 1 for the geometrical approach and 2 for the learning approach. Thus, we did not obtain a single choice of this hyperparameter to suit both (geometrical/learning) approaches.

TABLE III
ATTENTION CLASSIFICATION SYSTEM'S ACCURACY FOR THE DIFFERENT HYPERPARAMETERS CONFIGURATIONS IN THE VALIDATION SET [%]

		Approach					
		Geometrical			Learning		
Widths	$k\sigma \backslash p$	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
Same for both groups	$\mathbf{3}\sigma$	78.3	78.6	78.6	80.8	81.0	80.9
	$\mathbf{2}\sigma$	78.3	78.5	78.6	81.1	81.3	81.2
	$\mathbf{1}\sigma$	79.1	79.2	**79.2**	79.3	79.3	79.3
By group	$\mathbf{3}\sigma$	77.0	77.3	77.3	82.0	82.0	82.1
	$\mathbf{2}\sigma$	77.0	77.0	77.3	82.1	**82.2**	82.1
	$\mathbf{1}\sigma$	77.7	78.0	78.0	81.4	81.5	81.4

TABLE IV
ACCURACY OF OUR ATTENTION CLASSIFICATION SYSTEM USING THE BEST HYPERPARAMETERS CONFIGURATION [%]. SESSION 6 WAS THE VALIDATION SET FOR BOTH APPROACHES. SESSION 3 WAS THE TRAINING SET ON THE LEARNING APPROACH, BUT PART OF THE TEST SET ON THE GEOMETRICAL. ALL REMAINING SESSIONS WERE PART OF THE TEST SET FOR BOTH APPROACHES

	Approach	
	Geometrical	Learning
Session 3 (Test/Training)	78.6	**83.0**
Session 4 (Test)	79.5	**82.1**
Session 5 (Test)	78.2	**84.6**
Session 6 (Validation)	79.7	**82.2**
Session 7 (Test)	83.6	**89.1**

The parameter p that controls the size of Densepose bounding boxes, did not significantly affect the model in any of the approaches. However, since the results were slightly better for higher increases, useful and reliable keypoints were kept when augmenting the Densepose bounding boxes. Lastly, regarding the learning approach, the accuracy was higher when the widths were defined separately for each group (Therapist and ASD children), contrarily to what happened in the geometrical approach. While the best width in the learning approach for the therapist (0.4 m) is similar to the one chosen for the geometrical approach (0.43 m), the width of the child is very different in the two approaches (0.32 m in the geometrical approach versus 1.8 m in the learning approach). Therefore, in the geometrical approach an enlargement of child's width through the use of the therapist width was beneficial for the system's accuracy. Nevertheless, a separate definition of the widths boosts the system's performance, given the different angular directions towards each target for each group.

After setting the hyperparameters to the best configuration, the system performance scores were computed for all sessions. Table IV demonstrates that the proposed system generalises well, for the chosen hyperparameters, having high and consistent performance metrics for all the sessions, with accuracy values always equal or higher than 78% in the geometrical approach and 82% in the learning approach. Overall, the learning approach outperformed the geometrical approach.

Our results outperform the state-of-the-art approach in [8] (their accuracy: 73.5%), although they used a different dataset and protocol. Noteworthy, they rely on an head pose estimator, which is less effective in assessing attention than the gaze estimator we employed. Moreover, their study was conducted in a constrained scenario with more sensors than ours, quite different from the requirements of real therapeutic sessions as we considered in our approach. Comparison with the

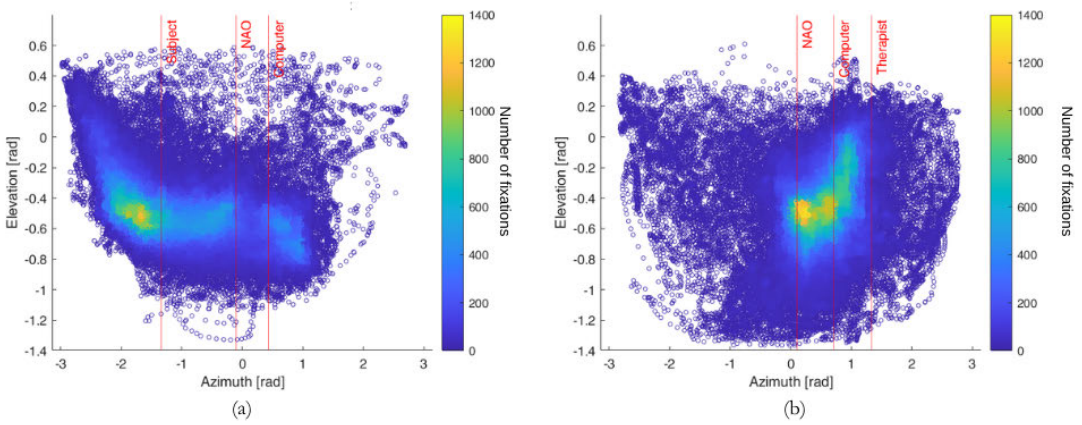


Fig. 7. Fixation maps of the therapist (a) and the children (b) gaze, considering the fixations of all sessions and all children. The red vertical lines represent the positions of the different targets.

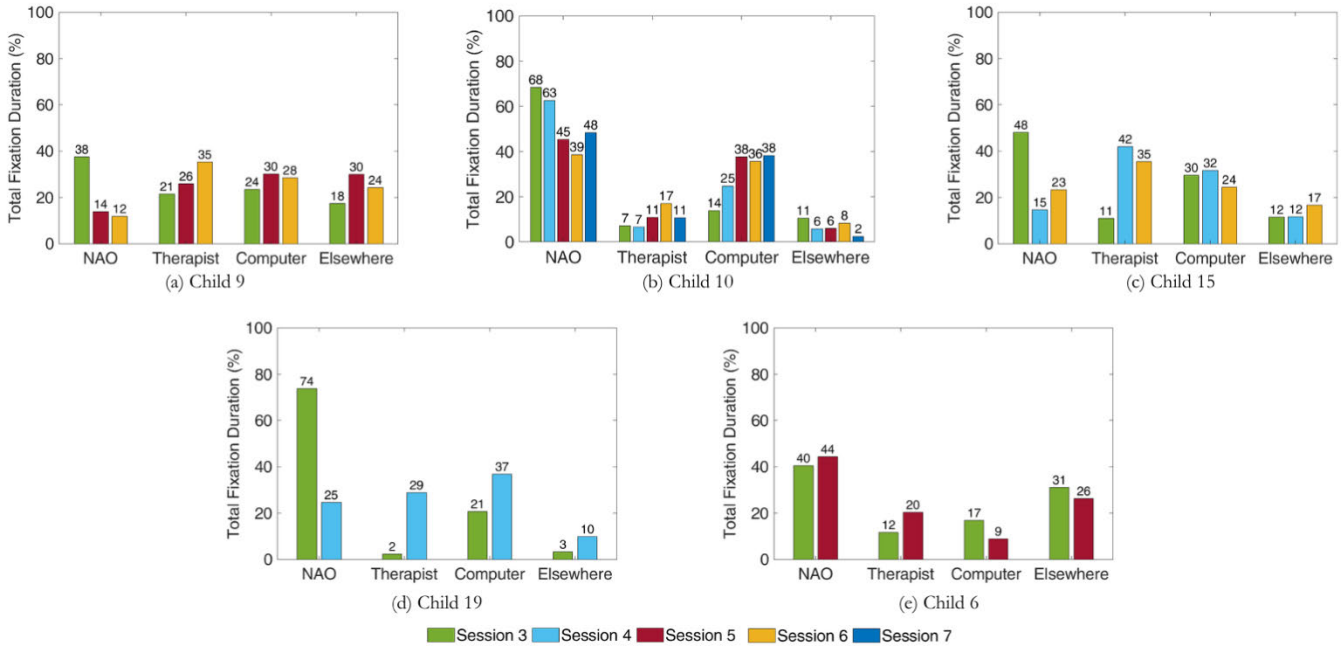


Fig. 8. Total Fixation Duration towards the targets and elsewhere along the sessions for Children (a) 9, (b) 10, (c) 15, (d) 19, (e) 6 [%]. In Session 3, Level 3 was performed. In the remaining sessions, Level 4 was performed by all children except Child 6 which repeated Level 3.

remaining state-of-the-art works (Table I) was not possible, due to the lack of system performance metrics in these papers.

C. Attention Analysis

To understand the on-task attention of each child, we did an attention analysis of the ASD subjects using the learning approach and its best hyperparameter configuration. From the fixation maps (Figure 7), we can observe the differences between the two groups: while the therapist seems to split the attention between the three targets, focusing principally on the child, the child divides its focus between the NAO and the computer, similarly to [6].

Then, we computed the Total Fixation Duration (TFD), for each child (see Figure 8). This expresses the overall time spent fixating each target during a session. Despite the different behaviours among children, there was a similar pattern for all children except Child 6: the interest in the NAO robot decreased with the sessions, while the attention towards the therapist increased. Regarding the Computer, the TFD increased always from session 3 to the consecutive session, even if it was not maintained in the following sessions. This behaviour was expected since the children performed Level 3 in session 3 and Level 4 in the other sessions, where the computer presented images related to the training scenarios. Child 6 did not have this behaviour (Figure 8 (e)) because in both sessions the Level 3 was performed. This observation is only possible due to the quantitative assessment of the children's attention during the sessions. The attention indexes can thus become an integral part of the clinical records, and be used in longitudinal studies to understand a child's evolution, as well as to serve as an input for dynamically modulating our protocol for each individual child.

D. Therapist Feedback

To investigate the agreement between the quantitative and qualitative analysis, we compared the attention system

TABLE V
OUR SYSTEM ATTENTION CLASSIFICATION ACCURACY FOR EACH CHILD IN EACH SESSION AND THERAPIST'S QUALITATIVE ANALYSIS. GREEN: GOOD SYSTEM ACCURACY (> 85%) OR POSITIVE THERAPIST FEEDBACK; YELLOW: AVERAGE SYSTEM ACCURACY (80% – 85%) OR NEUTRAL THERAPIST FEEDBACK; RED: LOW SYSTEM ACCURACY (< 80%) OR NEGATIVE THERAPIST FEEDBACK; P.: PERFORMANCE; T.: LIKES TO TOUCH ROBOT (FOR OUR ATTENTION SYSTEM, IT WAS A NEGATIVE CHARACTERISTIC, ALTHOUGH IT IS A POSITIVE PROTOCOL REMARK, SINCE THE CHILDREN SHOW INTEREST IN THE ROBOT)

	Child 6	Child 9	Child 10	Child 15	Child 19
Session 3	80 % Low P. T. NAO	76 % High P. T. NAO	84 % Low P.	84 % Low P.	93 % High P.
Session 4			76 % Avg P.	69 % Low P.	83 % High P.
Session 5	81 % Low P.	75 % High P. T. NAO	82 % High P.		
Session 6		78 % High P. T. NAO	82 % High P.	70 % Avg P.	
Session 7			87 % High P.		

outputs (Figure 8), with system accuracy and the therapist's qualitative feedback for each child and every session (Table V).

Observing the accuracy for each child, we realised that the model performed worse for Child 9 and 15, followed by Child 6. Comparing with the qualitative analysis of the therapist, presented in the same table, and the attention analysis, presented in Figure 8, we can draw the following remarks:

- Children who interact well with the robot (Child 10, 15, 19), improve their performance with the various sessions, and in general (Child 10 and 19) show higher system performance scores.
- According to the therapist's feedback, Child 10 is very interested in NAO, which is reflected in our system estimates, showing substantial attention towards NAO (Figure 8 (b)).
- Child 19 has the lowest level of ASD, which is consistent with the higher performance scores since this child tends to turn the head more towards the target when establishing eye contact.
- Children who like to touch the robot (Child 6 and 9), move a lot, and, consequently, cause detection problems on the Kinect data, deteriorating the system performance;

There is a good overall agreement between the therapist feedback, the system performance scores (accuracy of our system), and the classification of our attention system for each child (TFD). Moreover, when observing Figure 8, the therapist reported that the output of our system followed her expectations, especially for Child 10, whose level of attention towards the NAO robot was significantly higher when compared with the other children. In our opinion, these considerations demonstrate the possibility of using this system as an "explainable" AI tool, which is fundamental for the therapists to take ownership of the system and protocol, actively contributing to its co-development in a clinical environment.

TABLE VI
ACCURACY OF OUR ATTENTION CLASSIFICATION SYSTEM IN THE CADIN STUDY USING THE WIDTHS OF THE APPDA STUDY AND THE WIDTHS OBTAINED FROM THE LEARNING APPROACH

	APPDA widths	Learning widths
Session A (Test/Training)	73.2	**74.8**
Session B (Test/Validation)	69.2	**72.7**
Session C (Test)	73.0	**74.3**

E. Generalizability of the Model

To further test the attention system, we extended our sample using new data. The acquisitions happened in a specialized centre for neurodevelopment disorders, CADIn, between May and July 2022. The participants were seven male children, between 2 and 6 years old, without a definitive diagnosis, given their younger age, but with a prognosis of ASD. The study lasted nine weeks, with each child getting one session of 10 minutes each week. This study was also approved by CADIn ethical committee and all parents signed an informed consent.

The protocol and the setup were the same as the APPDA study, except the computer which was replaced by a tablet to reduce the attention of the child towards a distractor object. Therefore, two targets were used in this analysis: NAO and the Other Person. For each child, we selected three sessions that we named session A, B and C in chronological order.

During the study, the number of skeletons acquired by Kinect was extremely low since the children were smaller and the acquisitions room received more natural light. The interpolation step could not generate a sufficient number of skeletons, thus, we used a 3D pose reconstructor, described in [28] and applied the same processing steps presented before.

To start, we tested the parameters found in the APPDA study in the three selected sessions. Then we used the learning approach to find the best parameters, using sessions A, B, C as training, validation and testing set, respectively. Two annotators labelled selected frames from each session and 68% of the labels were kept for the final evaluation. The results are presented in Table VI. Although the learning approach provided better results than using the AOIs widths of the APPDA study, the performances were very close. Overall, we notice a decrease of the performance by 10%, but the algorithm still has a mean recognition rate of 74%, comparable with the state-of-the-art methods, in a scenario with more children with a larger age range and more challenging positions.

VI. CONCLUSION AND FUTURE WORK

We presented a quantitative attention analysis pipeline for ASD children, based on their gaze behaviour during unconstrained robot-assisted therapy sessions with multiple targets of interest. The need for strictly non-intrusive devices forced us to use a camera at a significant distance, greatly complicating the task of eye gaze estimation.

The proposed system combines a gaze estimation (Gaze360 model) and the definition of Areas of Interest (AOIs), followed by a gaze classification system to identify the different targets. Our main goal was achieved, reaching a total mean accuracy of **79.5%**, and outperforming the work proposed in [8], which was solely based on the head pose, and used a constrained

scenario with a higher number of cameras. Instead, our system was trained on primary school children and could generalise for preschool children. The therapist's qualitative observations are consistent with our quantitative results. This suggests that these metrics effectively capture the therapist's assessments and hold promise as clinical evaluation tools. Furthermore, therapists' ability to interpret these metrics bodes well for developing an explainable and transparent AI tool.

In addition to improving the system's modules (gaze estimation, AOI definition), future work will focus on:

- *Development of new clinical studies*, namely Randomized Controlled Trials to disentangle the effects of concomitant therapies and understand the impact of this robotic protocol; assess the importance of the robot in the triadic interaction, by designing studies of human-human interaction replacing the robot by a therapist, and comparing with our human-robot interactions.
- *Closed loop robot control* to react to the child's movements and attention/behaviour, creating protocols adapted to the ASD children's attention, and customised for each child, to enhance engagement and improve performance and learning. Complement the already mentioned TFD attention measurement method, with techniques from [7]. Moreover, study the effect of different robotic stimuli, as in [29], to verify which are most successful in capturing the child's attention.

We believe that our approach and results on attention/gaze evaluation are an encouraging step towards the clinical integration of robotic-assistive protocols, affording therapists with an AI-based, quantitative and understandable tool for monitoring the evolution of attention and social skills of ASD children.

REFERENCES

[1] M. J. Maenner et al., "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2020," *MMWR Surveillance Summaries*, vol. 72, no. 2, pp. 1–14, Mar. 2023.

[2] P. Pennisi et al., "Autism and social robotics: A systematic review," *Autism Res.*, vol. 9, no. 2, pp. 165–183, Feb. 2016.

[3] H. Kumazaki et al., "Optimal robot for intervention for individuals with autism spectrum disorders," *Psychiatry Clin. Neurosci.*, vol. 74, no. 11, pp. 581–586, Nov. 2020.

[4] A. Cerasa, L. Ruta, F. Marino, G. Biamonti, and G. Pioggia, "Brief report: Neuroimaging endophenotypes of social robotic applications in autism spectrum disorder," *J. Autism Develop. Disorders*, vol. 51, no. 7, pp. 2538–2542, Jul. 2021.

[5] B. Banire, D. Al-Thani, M. Qaraqe, K. Khowaja, and B. Mansoor, "The effects of visual stimuli on attention in children with autism spectrum disorder: An eye-tracking study," *IEEE Access*, vol. 8, pp. 225663–225674, 2020.

[6] S. M. Anzalone et al., "Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment," *Pattern Recognit. Lett.*, vol. 118, pp. 42–50, Feb. 2019.

[7] S. Ali et al., "An adaptive multi-robot therapy for improving joint attention and imitation of ASD children," *IEEE Access*, vol. 7, pp. 81808–81825, 2019.

[8] G. Nie et al., "Predicting response to joint attention performance in human-human interaction based on human–robot interaction for young children with autism spectrum disorder," in *Proc. 27th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 1–4.

[9] G.-B. Wan et al., "Attention shifting during child—Robot interaction: A preliminary clinical study for children with autism spectrum disorder," *Frontiers Inf. Technol. Electron. Eng.*, vol. 20, no. 3, pp. 374–387, Mar. 2019.

[10] G. Alvari, L. Coviello, and C. Furlanello, "EYE-C: Eye-contact robust detection and analysis during unconstrained child-therapist interactions in the clinical setting of autism spectrum disorders," *Brain Sci.*, vol. 11, no. 12, p. 1555, Nov. 2021.

[11] L. Santos, A. Geminiani, P. Schydlo, I. Olivieri, J. Santos-Victor, and A. Pedrocchi, "Design of a robotic coach for motor, social and cognitive skills training toward applications with ASD children," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1223–1232, 2021.

[12] Z. Zheng, E. M. Young, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Robot-mediated imitation skill training for children with autism," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 6, pp. 682–691, Jun. 2016.

[13] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdiscip. Rev., Data Mining*, vol. 11, no. 5, 2021, Art. no. e1424.

[14] A. P. Costa et al., "More attention and less repetitive and stereotyped behaviors using a robot with children with autism," in *Proc. 27th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 534–539.

[15] H. Kumazaki et al., "The impact of robotic intervention on joint attention in children with autism spectrum disorders," *Mol. Autism*, vol. 9, no. 1, p. 46, Dec. 2018.

[16] F. Alnajjar, M. Cappuccio, A. Renawi, O. Mubin, and C. K. Loo, "Personalized robot interventions for autistic children: An automated methodology for attention assessment," *Int. J. Social Robot.*, vol. 13, no. 1, pp. 67–82, Feb. 2021.

[17] *ALGazeAnalysis*. Accessed: Apr. 2024. [Online]. Available: http://doc.aldebaran.com/2-4/naoqi/peopleperception/algazeanalysis.html

[18] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *Proc. 31st BMVC*. BMVA Press, 2020, pp. 1–13.

[19] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "gaze360: Physically unconstrained gaze estimation in the wild," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6911–6920.

[20] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[21] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[22] America Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed. Arlington, VA, USA, 2013.

[23] N. I. Ishak, H. M. Yusof, S. N. Sidek, and N. Rusli, "Robot selection in robotic intervention for ASD children," in *Proc. IEEE-EMBS Conf. Biomed. Eng. Sci. (IECBES)*, Dec. 2018, pp. 156–160.

[24] A. S. Ivani et al., "A gesture recognition algorithm in a robot therapy for ASD children," *Biomed. Signal Process. Control*, vol. 74, Apr. 2022, Art. no. 103512.

[25] C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, and J. T. Mcconville, "Anthropometric survey of U.S. army personnel: Methods and summary statistics 1988," U.S. Army Natick RD&E Center, Natick, MA, USA, Tech. Rep. NATICK/TP.-89/027, 1988.

[26] M. Reed and K. D. Klinich, *A New Database of Child Anthropometry and Seated Posture for Automotive Safety Applications*, SAE Standard 2005-01-1837, 2010.

[27] C. Clifton et al., "Eye movements in reading and information processing: Keith Rayner's 40 year legacy," *J. Memory Lang.*, vol. 86, pp. 1–19, Jan. 2016.

[28] L. Santos, B. Carvalho, C. Barata, and J. Santos-Victor, "Extending 3D body pose estimation for robotic-assistive therapies of autistic children," in *Proc. 10th Biorob.*, Sep. 2024.

[29] S. Ali, F. Mehmood, Y. Ayaz, M. J. Khan, H. Sadia, and R. Nawaz, "Comparing the effectiveness of different reinforcement stimuli in a robotic therapy for children with ASD," *IEEE Access*, vol. 8, pp. 13128–13137, 2020.

Open Access funding provided by 'Politecnico di Milano' within the CRUI CARE Agreement