Gaze Analysis in Robotic Therapy for Autistic Children

Bárbara de Matos Águas Pereira da Silva

***Abstract*—This work aims to develop a quantitative model, using eye-gaze information, to evaluate the attention-response of children with Autism Spectrum Disorder (ASD), during therapeutic sessions with Social-Assistive Robots (SARs). ASD children show severe attention-deficits that hamper their ability to learn new skills. The automatic assessment of their attention-response would provide the therapists with an important biomarker to better quantify their behavior and monitor their progress/evolution. Previous attempts to quantify the attention-response of autistic subjects have focused on human-computer interactions tasks, with screen-based devices, that would distract the subject in therapeutical protocols with SARs. The work approach combines gaze extraction with the definition of context-dependent Areas-of-Interest (AOIs), to characterize periods of attention during the session. The methodology was tested with ASD children. Since extracting eye-gaze from optical-images is quite challenging, different methods were benchmarked. The Gaze360, which relies on image face-datasets and machine-learning, proved to be the most robust. For each target (therapist, subject, robot), the AOI (angular) horizontal/azimuth size was defined with two alternatives: a *geometrical-approach* combining the target's dimensions and the estimated Gaze360 noise, and a *learning-approach*. Once each target is associated to a range of fixation-angles, the eye-gaze estimates are used to classify the subject's focus-of-attention. Our experiments show that the *learning-approach* outperforms the *geometrical-approach*, achieving an accuracy above 82.0%. Finally, it is worth mentioning that the therapists understood the proposed attention-indices and found them aligned with their own evaluation of those subjects, an encouragement towards the future clinical use of the proposed system.**

***Index Terms*—Autism Spectrum Disorder, Social-Assistive Robots, Attention, Gaze tracking.**

I. Introduction

AUTISM Spectrum Disorder (ASD) is a neurodevelopmental condition with an increasing prevalence in the last years, affecting 1 in 64 children aged 4 years old, globally [1]. It is characterized by impairments in the social and communication domains, along with the presence of repetitive patterns of behaviors and interests. The symptoms and their severity vary significantly between 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 the ASD children, several therapeutic approaches have been used. Recently, it has been studied the introduction of Social-Assistive Robots (SAR) in the therapies [3]. SAR are usually able to attract the children's attention and interest, due to their simple and repetitive movements [4]. Studies on robot-mediated intervention have demonstrated positive outcomes in different social skills, such as communication, attention and imitation [5].

[1]B. Silva is with Instituto Superior Técnico, Lisboa, Portugal.

An important factor to assess the quality of the therapies with SAR is the engagement of the participants [6]. Assessing the children engagement during the therapies is important, since it provides a more complete and clear notion of the therapy sessions, supplementing the therapist feedback. According to the children engagement, the protocols can be updated and adapted, in order to improve them and achieve better outcomes [7]. One of the main engagement features is the attention, which is often compromise in ASD children, namely the on-task attention. This type of attention represents the willingness to acquire and to develop new skills during a task, being a major prerequisite for a good performance in the therapy sessions [8]. Therefore, it is important to develop assessment tools for this capability.

To assess the children behavior and evaluate both the social and motor training, qualitative measures of the attention have been used extensively in research and clinical practice, mainly through manual video coding [9], [10]. However, this process is time consuming. Therefore, recent studies have focused on obtaining reliable, objective and quantitative measures of the attention based on the head orientation [11], the detection of facial landmarks [12] and/or the eye gaze [13]. Nevertheless, this is a hard task given the requirement of using non-intrusive devices, since ASD children can find physical sensors uncomfortable and distracting [14]. Recently, multiple methods for non-intrusive quantitative analysis of the attention of ASD children have been proposed. Some works used screen-based eye trackers, such as Tobii, in human-computer interactions, which is not feasible in robotic therapy. Although in a sparse number, some studies focused on assessing the autistic subjects' attention in 3D spaces, during human-robot interaction, by using non-intrusive cameras, able to record the procedures.

In [13], the attention was quantified through the eye gaze estimation, based on a geometrical approach to define the Areas of Interest (AOIs) around the targets. Each teenager performed one trial, which consisted of two consecutive conversations: with a female human and a female-type android robot. To obtain the gaze estimation, a small Tobii eye-tracker device, similar to the Microsoft Kinect (referred as Kinect from now on) was used. This device was able to calculate when each subject was looking to the interlocutor face, by manually defining AOIs around their faces, with an augmented elliptic form to overcome the eye-tracker noise. In this study, the ASD group showed lower attention towards social targets than the Typically Developed (TD) group and both groups showed more interest in the robots' faces than in the human faces. However, the use of Tobii is difficult to implement in

unconstrained environments, therefore other approaches were studied.

In [11], [15], the attention was quantified through the head pose, based on a machine learning approach. NAO was placed on top of a chair in front of the child and two stimulus were placed on each side of the child. To capture the trials, in [15], 4 web cameras were placed around the child, while in [11], 1 Kinect camera was used. Since the targets were placed in locations that required head movements when changing the focus of attention, the attention was assessed based on the head pose, which was estimated using a supervised machine learning method. To classify the attention, the k-Means algorithm was used, to find n clusters in the 2D head pose estimations. In [15], the model achieved an accuracy of 73.5% and, in [11], the ASD group showed less Joint Attention (JA) than the TD group.

Lastly, in [16], the attention was quantified through the eye gaze and the head pose estimations. The study was conducted in a therapy environment using 3 robots, side by side, in front of the child and 3 cameras positioned around the room. Using the cameras' outputs, the eye gaze and the head pose were estimated by the OpenFace [17]. It uses Convolutional Neural Networks to detect several facial landmarks. It can also identify landmarks in the eyes area, using a Constrained Local Neural Field, and thus, calculate the associated gaze angles (azimuth and elevation). In [16], the attention angle was obtained by doing the average between the gaze and head pose estimations. To obtain an attention classification, the angle was compared with the manually defined AOIs, corresponding to 3 ranges of azimuth angles, one for each robot. However, the OpenFace is not able to estimate the gaze when a part of the eyes is occluded. Therefore, other algorithms were developed, such as Gaze360 [18].

The Gaze360 model [18] is a method for robust 3D gaze estimation able to predict the gaze without visible eyes, through the inclusion of temporal information in the gaze estimations. The proposed model receives, as input, multiple cropped head frames, which are passed through a backbone network. To obtain the cropped head frames, the Densepose [19] facial detector algorithm is suggested by the authors. Afterwards, the output of each frame passes through bidirectional Long Short-Term Memory cells, which are neural networks that model sequences where the output for one frame is dependent on past and future frames. In the Gaze360 model, 7 frames are used, corresponding to the current frame, the 3 previous frames and the 3 following frames. Consequently, even if the gaze is occluded, it is possible to calculate the gaze angles (azimuth and elevation) based on the previous and following frames. This allows to estimate the gaze, even if a person turns his/her back to the camera. Thus, it is a full range gaze estimator, covering 360° [18]. The Gaze360 model outputs gaze angles relative to the camera view, using spherical coordinates (azimuth and elevation). This means that if the subject looks directly to the camera, independently of the subject's position, the output is $0rad$ for the azimuth and $0rad$ for the elevation [18].

Considering the proposed quantitative attention estimation models in a physical human-robot interaction environment, the number of studies focused in unconstrained spaces is still sparse. Regarding the setups, NAO is the most used robot. For some studies [15], [16], multiple cameras were used in order to estimate the gaze and head pose, which is not ideal, since it means a more complex setup and consequently, more distractions for the ASD subjects. In [11], the authors showed the feasibility of using only one Kinect to measure the attention. From the different AOIs definitions, it is possible to conclude that both geometrical and learning approaches have been proposed when analysing the attention based on the gaze or head pose estimations. In the geometrical approaches, the AOIs have an increased shape of the target or they are defined as a range of azimuth angles. In the learning approach, the method used was the k-Means algorithm to cluster the data.

Overall, considering the specificities of an ASD therapy, the best method corresponds to the one capable of estimating the gaze, since it is the main source of attention, using only one camera, while adopting either a geometrical or a learning approach for the AOIs definition.

Therefore, this work focuses on the development of an accurate quantitative model, able to evaluate the ASD children's attention during therapeutic sessions. This problem is complex especially due to the children and therapy intrinsic characteristics. The use on non-intrusive devices can make the attention assessment harder, since the distance between a subject face and the tracking device is higher. Furthermore, considering the therapeutic environments, which are unconstrained, the people are able to move freely, making tracking and attention assessment extremely challenging. Using the data extracted from these devices, a secondary goal of this work is that the attention assessment model should comply with the Explainable Artificial Intelligence (AI) concept, thus, producing results which can be interpreted and understood by therapists and reflect their own opinions.

In the rest of the paper, first, the clinical study is described, afterwards, it is presented the gaze and head pose estimators benchmarking, followed by the proposed attention assessment framework. The framework is composed by a data pre-processing, a scene geometry analysis and the definition of AOIs. Lastly, the results obtained for the proposed attention assessment framework are presented, along with their discussion.

II. Methodology

A. Clinical study

In order to evaluate the ASD children's gaze in robotic therapy, one clinical study was done and analysed. The main goal of the therapy was to train gestures during triadic interaction sessions between the therapist, the ASD patient and the robot NAO. To instigate this kind of interaction, a setup and a protocol for an imitation game were defined based on clinical knowledge [20].

The clinical study corresponds to a school study done between May and July of 2021 in Escola Básica Bernardim Ribeiro in association with APPDA Lisboa, the main Portuguese association of autism. The participants were six ASD children, 5 males and 1 female, with ages between 7 and 11

years old. Five children were diagnosed with level 3 of ASD, while one child was diagnosed with level 1, according to the Diagnostic and Statistical Manual of Mental Disorders V [21]. Level 1 is the less severe, indicating that the child requires relatively little support, while level 3 is the most severe, meaning that the child requires very substantial support. The study lasted 7 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 depending on their school attendance during the acquisition days.

The protocol consisted of an imitation game with several levels to train gestures [20]. The setup consisted of a triangle between the three entities, with NAO robot placed in the middle of the therapist and the child. A non-intrusive Kinect sensor was placed behind NAO to extract 25 3D joints from the therapist and child skeletons and to record the sessions. The therapist was responsible for the robot control, through a computer placed near him/her. In this way, the therapist was able to choose the best exercises for each child, personalising the therapy.

The data acquired through the Kinect camera was saved, frame by frame, during the sessions. This data is constituted by the video, the calculated skeletons and the times of each frame expressed in the Unix timestamp.

Since the therapy was done in the school atrium, multiple people passed through the space during the sessions. These people have to be considered in the analysis, not only because they are a distraction for the therapist and children attention, but also because some of them passed inside the Kinect camera range, being their skeletons detected by it. To keep only the data of interest to the study, the two skeletons with a higher x coordinate of the left shoulder are kept. To distinguish between the therapist and the child skeleton, the therapist used a red scarf which was tracked during the sessions.

B. Benchmarking Gaze and Head Pose Estimators

To estimate where a person was looking at, a model that outputs the people's gaze or head pose is needed. Thus, gaze and head estimators were studied and compared. Only the azimuth is analysed in this work, since it is sufficient to discriminate the targets in our application, which are positioned in different horizontal directions during the sessions.

To validate and compare the Gaze360 [18] (gaze) and WHENet [22] (head) estimators, 3 controlled experiments at long distance were done. They are both full range (360°) estimators, able to perform even when the face features are not visible in the video.

The first experiment consisted of 4 fixation points around the subject (including one in the back). The second experiment consisted of 4 fixation points located on the sides or front of the subject. The third experiment consisted of 3 fixation points located in front of the subject and only eye movements were performed, keeping the head pose still. The obtained Root Mean Squared Errors (RMSEs) between the estimations and the expected signal are shown in Table I.

Given the results for the first experiment, it was concluded that the WHENet model is more accurate than Gaze360 estimating the azimuth for the point in the back, when all facial features are occluded. However, when there is no fixation point in the back, both models performed similarly (Experiment 2). The results for Experiment 3 prove that the WHENet model is not suited for our setup, since it is not able to follow the eye gaze when the head pose is the same. Concluding, the Gaze360 model had a good performance which mainly depended on the visibility of facial features, usually visible during the therapy sessions, since the targets were located in front or to the sides of the people. Therefore, the Gaze360 model was implemented in the proposed framework.

TABLE I
AVERAGE RMSE [PX] OF THE AZIMUTH ESTIMATIONS OF THE WHENET AND GAZE360 MODELS FOR THE 3 LONG DISTANCE EXPERIMENTS, WHERE THERE IS THE MOVEMENT OF: EXPERIMENT 1- WHOLE BODY; EXPERIMENT 2 - HEAD AND EYES; EXPERIMENT 3 - JUST EYES

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

C. Data Analysis

In this work, it is proposed a system that estimates where the people were looking at during the sessions. This system is composed by a gaze extractor (Gaze360) and it is based on the definition of AOIs to obtain the range of angles that correspond to looking at the different targets (Figure 1). To define the AOIs, 2 approaches (geometrical and learning) were studied. Having both the gaze estimation and the AOIs, the gaze was classified and the attention-indexes obtained for the clinical study described before.

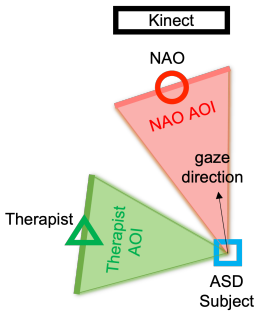


Fig. 1. Representative setup, along with the Therapist (green) and NAO robot (red) AOIs and the subject gaze estimation (black arrow)

Data Preprocessing

The Kinect camera was not able to detect both skeletons correctly for all the frames. There were moments where the child and therapist bodies crossed and overlapped. Furthermore, people often sat and in some sessions the space had a poor illumination. In this way, the Kinect decreased its skeleton computation capability and tracking performance. Thus, the data needed to be preprocessed.

First, a symmetry relative to yz was applied to the reference frame, $(x, y, z) \rightarrow (-x, y, z)$, since the Kinect obtained symmetrical images. After, the frames in which the Kinect did not detect both skeletons were discarded. The remaining

data was filtered using a median filter with a window of seven frames to extract the outliers.

Since the percentage of data lost from the Kinect restrained the analysis of some sessions, an interpolation of the remaining data was explored. The skeletons were reconstructed by doing a linear interpolation to the therapist and child keypoints. Assuming that the participants did not move considerably during a session, if the first and last keypoints missed, a constant interpolation was done to set them to the closest value, corresponding to the first and last detected keypoints, respectively.

To relate the Gaze360 output with each person, the Kinect 2D head joints were compared with the Densepose head boxes, outputted by the Gaze360 model. Both were represented in pixels and in the same coordinate system. However, they were obtained in different scales: $1920 \times 1080px$ for the Kinect and $960 \times 720px$ for the Densepose. Thus, the Densepose bounding boxes were scaled to the Kinect scale. Then, they were increased to compensate the errors from the Kinect skeletons detection by $size * (1 + p)$, with $p = \{0.25, 0.50, 0.75\}$.

After, it was checked which head joint (therapist or ASD child) was inside each Densepose bounding box, for each frame. If it was impossible to have both head joints inside two different Densepose head boxes, the frame was discarded. In this way, only frames with both skeletons and the corresponding Densepose bounding boxes were kept. This step also ensured that the frames in which the interpolation had a considerable error were discarded.

Given the higher percentage of lost data without interpolation (Table II), it was decided to incorporate the interpolation in the data preprocessing.

TABLE II
PERCENTAGE OF LOST DATA IN SESSION 4: WITH AND WITHOUT DATA INTERPOLATION. THE RED CELLS REPRESENT THE SESSION IN WHICH MORE THAN 2/3 OF THE DATA WAS LOST [%]

	Child 10	**Child 15**	**Child 19**
No Interpolation	28	81	82
Interpolation	1	20	17

Scene Geometry Analysis

To estimate where the people were looking at during the therapy sessions, the angles corresponding to looking at the different targets were calculated. From now on, these angles will be called standard angles.

The targets corresponded to NAO, the Other Person (Therapist for the Child and Child for the Therapist) and the Computer. The Computer was considered, since it attracted the ASD children's attention, when the therapist interacted with it to choose which exercises to perform and when the scenarios, from level 4, appeared on it. Thus, during levels 1, 2 and 3 the computer was considered a distraction, while in level 4 was a focus of attention.

The standard angles 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 people's positions, they were calculated for each frame. The therapist and child positions were obtained using the head joints extracted by the Kinect camera.

Considering that NAO was the closest entity to the camera, only one equation was needed to calculate the standard angles for looking at it. The angle was given by Equation 1, where x and z are the 2D positions of the therapist or the child, in the Kinect coordinate system. x_{diff} and z_{diff} are the difference between the 2D coordinates of the person and the target, given by Equations 2. An example is presented in Figure 2.

$$\alpha_{NAO} = \arctan\left(\frac{x}{z}\right) - \arctan\left(\frac{x_{diff}}{z_{diff}}\right) \tag{1}$$

$$\begin{aligned} x_{diff} &= x - x_{target} \\ z_{diff} &= z - z_{target} \end{aligned} \tag{2}$$

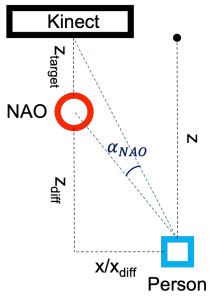


Fig. 2. NAO standard angle (α_{NAO}) representation. The red circumference represents NAO, while the blue square represents the person from which the standard angle is calculated. The referential is located in the center of the Kinect

To obtain the standard angles for looking to the other person and at the computer, four different conditions were used, according to the person and the target positions:

a) $x_{diff} \times x > 0$

If the person and the target were on different sides of the camera ($x \times x_{target} < 0$) or if they were at the same side, but the target was closer to the camera in the x axis ($x \times x_{target} > 0$ and $|x| > |x_{target}|$), the angle corresponding to looking at the target was given by Equation 3.

$$\alpha_{target} = -\arctan\left(\frac{z}{x}\right) + \arctan\left(\frac{z_{diff}}{x_{diff}}\right) \tag{3}$$

b) $x_{diff} \times x \leq 0 \wedge z_{diff} > 0$

If the person and the target were at the same side of the camera and the person was closer to the camera in the x axis and further in the z axis ($x \times x_{target} \geq 0$ and $|x| \leq |x_{target}|$ and $|z| > |z_{target}|$), Equation 4 was used.

$$\alpha_{target} = \arctan\left(\frac{x}{z}\right) - \arctan\left(\frac{x_{diff}}{z_{diff}}\right) \tag{4}$$

c) $x_{diff} \times x \leq 0 \wedge z_{diff} < 0$

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 ($x \times x_{target} \geq 0$ and $|x| \leq |x_{target}|$ and $|z| < |z_{target}|$), the angle for looking at the target was given by Equation 5, using $n = 1$ if $\arctan\left(\frac{z}{x}\right) \geq \arctan\left(\frac{z_{target}}{x_{target}}\right)$, and $n = -1$ if $\arctan\left(\frac{z}{x}\right) < \arctan\left(\frac{z_{target}}{x_{target}}\right)$.

$$\alpha_{target} = n \arctan\left(\frac{x}{z}\right)^n + \arctan\left(\frac{z_{diff}}{x_{diff}}\right)^n + n\frac{x}{|x|}\frac{\pi}{2} \quad (5)$$

Analysis of Gaze(360) Angles Distribution

The gaze estimations, calculated by the Gaze360, were filtered to reduce the noise, using a mean filter with a window of 7 frames. Then, the Gaze360 estimations were centralized to each target, according to Equation 6, using the calculated standard angles (α_{target}).

$$\alpha_{centralized} = \alpha_{Gaze360} - \alpha_{target} \quad (6)$$

After, the histograms of the angles distribution were obtained for each centralization and for each person in each session. Analysing the angles distribution, it was possible to observe that the maximum of the histogram was deviated from the expected target angles. Therefore, the effect of correcting these offsets for looking at each target was evaluated.

According to the histogram bar widths, the number of maximums and their positions varied. In this way, several bar widths were tested, using one of the sessions. Since the targets had different azimuth locations for the children and the therapist, the best bin widths were computed separately for each group (ASD and Therapist). Since there were 3 targets, the best widths were the ones where most histograms from each group had 3 maximums. The obtained results were $0.35rad$ for the Therapist and $0.25rad$ for the Children.

To obtain the offsets, an automatized extraction of the histogram maximums was studied. For each centralized histogram, the position, in the x-axis, of the closest maximum to the center ($0rad$) is used as the offset of that target. The obtained results are shown in Figure 3.

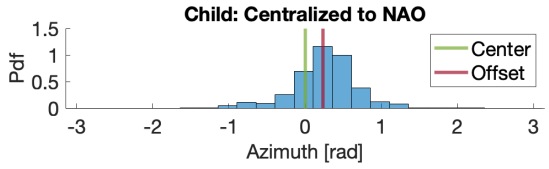


Fig. 3. Histogram of the Gaze360 estimations centralized to NAO for Child 10 in Session 2. The green line represents the center of the histograms ($0rad$) and the red line the computed offset. Pdf: Probability distribution function

The obtained offsets were later added to the calculated AOIs.

Areas of Interest Definition

According to the literature review, it was decided to define AOIs around each target (NAO, other person and computer), in the horizontal direction (azimuth), as previously explained, with the final goal of finding the range of angles corresponding to looking at each target. To reach this goal there were 7 stages:

1. Each AOI was centered in the 2D coordinates of the target and defined in the normal to the line connecting the person and the target. The AOI slope was obtained using the equation in Figure 4, where x_{diff} and z_{diff} represented the difference between the 2D coordinates of the person and the target.

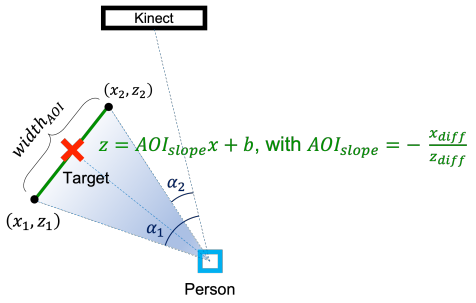


Fig. 4. Representative top view of an AOI. The red cross represents the target, while the blue square represents the person in analysis. The green line corresponds to the AOI and the area in blue to the range of angles for looking at that target AOI

2. The width of the AOI was defined $[m]$, using two approaches (geometrical and learning), as explained below.

3. The 2D coordinates of the borders of the AOI were calculated $[m]$. Having the slope, the width and the center of the AOI, two limits ((x_1, z_1) and (x_2, z_2)), one on each side of the target, were obtained using Equations 7, where r=1,2.

$$\begin{cases} x_r = x_{target} + (-1)^r \times \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \\ z_r = z_{target} + (-1)^r \times AOI_{slope} \times \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \end{cases} \quad (7)$$

4. The range of angles corresponding to looking at each AOI were computed $[rad]$. To obtain the relative positions between the person and the two extremities of the AOI, Equation 2 was used, with $(x_{target}, z_{target}) = (x_1, z_1)$ and $(x_{target}, z_{target}) = (x_2, z_2)$. Afterwards, Equation 1 was applied when NAO was the target, while Equations 3, 4 and 5 were chosen for the remaining targets. At the end, two angles were obtained, α_1 and α_2, one for looking at each side of the AOI.

5. The Gaze360 offsets $[rad]$, obtained for each target through the centralized histograms, were added to the calculated angles, α_1 and α_2.

6. The obtained angles were associated with the left and right limits of the AOI, obtaining $[\beta_{right}; \beta_{left}]$. To do this association, the angles θ_k were corrected to belong to the interval $[-\pi; \pi]$. Moreover, it was checked if the absolute difference between the 2 values of θ_k was lower than π (Equation 8).

$$\begin{cases} [\beta_{right}; \beta_{left}] = [\theta_{min}; \theta_{max}] \text{ , if } |\theta_1 - \theta_2| < \pi \\ [\beta_{right}; \beta_{left}] = [\theta_{max}; \theta_{min}] \text{ , if } |\theta_1 - \theta_2| > \pi \end{cases} \quad (8)$$

7. At the end, the AOIs which were overlapping, as shown in Figure 5, were corrected. This process was done frame by frame and was composed by 2 parts.

First, it was checked if an AOI was totally overlapping other. In these cases, one of them was deleted, according to the scene geometry and the targets priority. When the therapist or the child (other person) AOI were in the same gaze direction as NAO or the Computer, the AOI from the last target (NAO or

Computer) was discarded. This happened because the scenario was fixed and both the therapist and child were inside the scene. Thus, the person was always the one blocking the view to the other targets. If the AOI of NAO was covering or being totally covered by the computer AOI, the computer AOI was deleted, since NAO was a main target in the protocol.

Then, for the instants in which two AOIs partially overlapped, a limit between the AOIs was calculated. For each instant, two Gaussian curves (one for each target) were created, $N(\mu, \sigma)$. For each target, the mean, μ, was defined as the mean value of the AOI limits at that instant (Equation 9). The standard deviation, σ, was calculated using an empirical rule. It was defined that half of the AOI width was equal to $k\sigma$, with $k = \{1, 2, 3\}$ (Equation 10). In this way, according to the empirical rule, 68%, 95% and 99.7% of the values lie within k standard deviations of the mean.

$$\mu = \frac{\beta_{left} + \beta_{right}}{2} \tag{9}$$

$$k\sigma = \frac{width_{AOI}}{2} \tag{10}$$

The x-value in which the Gaussians intersected was defined as the limit between the AOIs.

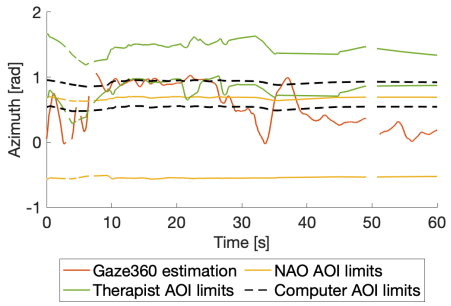


Fig. 5. Gaze360 estimation and AOIs limits before correcting the AOIs overlapping, for the Child 19 during the first $60s$ of Session 4. The gaps across time correspond to the frames discarded during the data preprocessing

Geometrical Approach

In the geometrical approach, referred in the step 2 of the AOIs definition, the widths of the AOIs were decided based on the targets dimensions and the Gaze360 noise, which was added to each side of the target limits to obtain the AOIs final widths. The widths of NAO and the participants were decided based on the literature. For the therapist and child it was considered the average shoulders' width (bideltoid) of a male adult, which is $47.6cm$, according to [23]. For NAO, the shoulders' width, as well as the arms length were taking into account, since NAO is using the arms to make the gesture training. In this way, the NAO width was defined as $27.5 + 31.1 \times 2 = 89.7cm$.

The Gaze360 noise was obtained from the controlled experiments for the estimators benchmarking. The obtained signals were segmented in order to keep only the segments where there were no changes of gaze direction. Then, the signal segments corresponding to looking at the same fixation point were concatenated and the standard deviation between the expected signal and the Gaze360 estimations were obtained for each fixation point. At the end, the highest standard deviation value ($0.069rad$) was assumed to be the Gaze360 noise.

After step 6 of the AOIs definition, the Gaze360 noise $[rad]$ was added to the obtained range of angles, according to Equations 11.

$$\begin{aligned} \beta_{right} &= \beta_{right} - noise \\ \beta_{left} &= \beta_{left} + noise \end{aligned} \tag{11}$$

Learning Approach

In the learning approach, referred in step 2 of the AOIs definition, the best widths for each target were obtained through ground truth comparison. The data from the 7 sessions was divided in training, validation and testing sets. The training set was used to find the best widths for each target. The validation set was used to make decisions relative to the model hyperparameters and validate the choice of the best width. The test set was used to test the model and obtain the model performance scores.

The proposed model was trained, using the training set, for multiple configurations of the hyperparameters, thus, for each configuration a best set of AOI widths was obtained. Each set of widths consisted of one width for each AOI (NAO, Other Person and Computer). The values tested for the NAO width varied between 0.4m and 3.0m, with increments of 0.2m. For the other person, varied between 0.4m and 2.0m, also with increments of 0.2m. While the values tested for the computer width varied between 0.4m and 1.0 with increments of 0.2m. This means, the model was run for 14*9*4=504 combinations of widths.

To compute the best combination of widths, the Receiving Operating Characteristic (ROC) curve was computed for each set of hyperparameters. The best set of widths corresponded to the one with the ROC curve value closer to the upper left corner of the graph, as given by Equation 12.

$$bestROCscore = min(\sqrt{(1 - Recall)^2 + FPR^2}) \tag{12}$$

Then, the best combination of AOI widths, for each hyperparameters configuration, was implemented in the proposed model. In order to choose the best configuration, the model was run in the validation set for each configuration and the performance scores were compared. At the end, the best configuration of hyperparameters was implemented in the model and it was applied in the test set. To evaluate the generalization model capacity, the performance scores were obtained.

Fixation Signal

The attention metrics were obtained based on the fixations. To obtain the fixations, the Gaze360 estimations were quantified and a binary signal for looking at each target was generated (NAO, Other Person and Computer). For each frame, the binary signal was set to 1, if the Gaze360 estimation was inside the AOI range of angles, and 0, otherwise. To facilitate the computation of the attention metrics, the final Gaze360

signal was obtained by summing the 3 signals, using different factorization values for each target.

Then, it was considered that a person was looking at a target if a fixation occurs. According to literature, a fixation was defined as at least $400ms$ looking to an AOI [24]. In this way, to remove most of the non-fixations, the final gaze signal was filtered with a median filter with a window of $800ms$.

III. Experimental Results and Discussion

A. Data and Metrics

Given that some sessions were not recorded, the data to analyse had to be selected. There were 7 sessions, from which the first one was not recorded and the second one consisted mainly of familiarization levels. Since the familiarization was not the main goal of the Protocol, the attention was only analysed for Sessions 3 to 7. Moreover, since one Child was only present in Session 1 and 2, he/she was excluded from the analysis.

To evaluate proposed model performance, the data was split randomly in three sets. Session 4 was used as train set, Session 6 as validation set and the remaining sessions as test set.

To obtain the ground truth, the videos from the therapy sessions were labelled by two annotators. The labelling was done by selecting where the therapist and the ASD child were looking at in the selected frame (NAO, Other Person, Computer or Elsewhere). Since more than 25 videos were acquired, with some having more than 15000 frames, their label for both the therapist and the patient is a cumbersome task. Therefore, the videos were labelled every 3 seconds, a period which reflects the main changes in terms of fixations at the different targets. At the end, only the labels from frames, with both annotators totally agreeing were kept, resulting in more than 75% of the labels being kept for all the sessions, confirming the good inter-annotator agreement.

To evaluate the proposed model, a confusion matrix for each session was obtained to compute the evaluation metrics. Therefore, a 2×2 confusion matrix was calculated for each person (Therapist and Subject) looking at each target (NAO, Other Person and Computer) and Elsewhere. This was done by comparing the ground truth classification with the classification estimated by the proposed model. For each session, $8 \times n_{Subjects}$ confusion matrices were computed (4 for each person), where $n_{Subjects}$ is the number of subjects that were present in that session day. At the end, all the confusion matrices from each session were summed in order to obtain a final one.

Since the ASD patients may have different behaviors 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 group), in the learning approach.

B. Proposed System

After establishing the whole framework, the system hyperparameters were validated by computing the accuracy of the proposed model through the comparison of the gaze classification estimations with the ground truth classification. The main hyperparameters were the Densepose bounding boxes ratio ($p = \{0.25, 0.50, 0.75\}$), the Gaze360 offsets and variable $k = \{1, 2, 3\}$, from Equation 10, used to define the standard deviation of the Gaussians when 2 AOIs overlap.

Geometrical Approach

The results for the several hyperparameters configurations using the validation set in the geometrical approach are shown in Table III, with the best configuration in bolt. Observing the results, the accuracy is higher in all the tests without the offsets correction than in the tests with the offsets correction. In general, the best increase of the Densepose bounding boxes is $p = 50\%$ and the best value of k depends on the correction of the Gaze360 offsets.

TABLE III
GEOMETRICAL APPROACH ACCURACY OF THE PROPOSED MODEL CLASSIFYING THE GAZE AS LOOKING AT THE DIFFERENT TARGETS AND ELSEWHERE FOR THE DIFFERENT HYPERPARAMETERS CONFIGURATIONS USING THE VALIDATION SET [%]

	Without Offsets Correction			**With Offsets Correction**		
$k\sigma \setminus p$	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
3σ	78.5	78.7	78.6	76.6	76.6	76.4
2σ	79.0	79.1	79.1	76.4	76.5	76.4
1σ	80.0	**80.1**	80.1	76.2	76.1	76.1

After implementing the best hyperparameters configuration, the model performance scores were computed for the test set, as shown in Table IV. The model has a good performance for all the sessions, with high and consistent scores for all the metrics, proving that it generalizes well for the chosen hyperparameters.

TABLE IV
GEOMETRICAL APPROACH MODEL PERFORMANCE SCORES, CLASSIFYING THE GAZE, USING THE CHOSEN HYPERPARAMETERS CONFIGURATION [%]

	Session				
	3	4	5	6 (Validation)	7
Accuracy	81.1	80.8	79.2	80.1	83.9

Learning Approach

The results for the several hyperparameters configurations using the validation set in the learning approach are shown in Table V, with the best configuration in bold. Observing the results, the accuracy increases for all the hyperparameters configurations when the Gaze360 offsets are not corrected and the widths are used by group.

TABLE V
LEARNING APPROACH ACCURACY OF THE PROPOSED MODEL CLASSIFYING THE GAZE AS LOOKING AT THE DIFFERENT TARGETS AND ELSEWHERE FOR THE DIFFERENT HYPERPARAMETERS CONFIGURATIONS USING THE VALIDATION SET [%]

		Offets Correction					
		Without			**With**		
Widths	$k\sigma \setminus p$	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
Group	**3**σ	82.0	82.0	82.1	79.1	78.7	79.3
	2σ	82.1	**82.2**	82.1	78.5	78.6	78.5
	1σ	81.4	81.5	81.4	79.3	79.4	79.4
Total	**3**σ	80.8	81.0	80.9	78.0	78.3	78.3
	2σ	81.1	81.3	81.2	77.9	78.2	78.2
	1σ	79.3	79.3	79.3	78.4	77.9	78.5

Analysing the hyperparameters effect, the Gaze 360 offsets correction is the main source of change of the performance scores. The accuracy is higher without the offsets correction, for all the other hyperparameters configurations. This occurs because the Therapist and the Computer targets (for the Child) are shifted to the same position when correcting the offsets, due to their close location, which imply the existence of only one maximum in the centralized histograms. Thus, one of the following two situations happens frequently: (1) the AOIs have a small range of angles or (2) one of them is totally overlapping the other. The last situation results in the exclusion of the Computer AOI. Consequently, the performance decreases for these two targets, leading to lower performance scores.

The parameter k, from Equation 10, seems to affect the model differently depending whether the Gaze360 offsets are corrected or not. Without the offsets correction, the best value is always $k = 2$, while using the offsets correction, this is usually the value with the worst performance. Thus, no conclusion can be taken about the effect of this hyperparameter.

The parameter p, regarding the Densepose bounding boxes increase, does not affect the model significantly. Without doing the offsets correction, the results seem better for $p = 50\%$, while using the offsets correction, the performance is better for $p = 75\%$. Since the results are better for higher increases, it is proved that augmenting the Densepose boxes, allows to keep useful and reliable keypoints that improve the system performance.

Analysing the model performance scores, shown in Table VI, it is concluded that the proposed framework is generalizing well, having high and consistent performance metrics for all the sessions, including accuracy values always above 80%. Overall, the scores are very good, given the study conditions and the wrong keypoints detection from the Kinect.

TABLE VI
PERFORMANCE SCORES OF THE LEARNING APPROACH MODEL USING THE CHOSEN HYPERPARAMETERS CONFIGURATION [%]

	Session				
	3 (Training)	4	5	6 (Validation)	7
Accuracy	83.8	82.0	84.6	82.2	89.1

Comparing the model performance scores (Tables IV and VI), it is visible that the learning approach outperforms the geometrical approach.

C. Attention Analysis

To understand the on-task attention of each child, an attention analysis of the ASD subjects was done using the best approach and hyperparameters. The Total Fixation Duration (TFD), expressing the time fixating each target along sessions, was obtained for each child (Figure 6), as well as the individual accuracies in each session (Table VII), in order to relate with the therapist qualitative feedback of the sessions (Table VIII). The attention was only studied for children with more than 2 analysed sessions (Child 9, 10 and 15).

For all children, in all sessions except for the green one, Level 4 was performed. Since in this level the computer is considered a focus of attention, it is expected that the interest in the Computer is higher, which is verified for Child 10 (Figure 6b). For Children 9 and 15 (Figures 6a and 6c), the gaze towards the computer increases from the green to the light blue session, however, it is not kept. Thus, these children only present a slightly higher interest in the computer when Level 4 is presented for the first time.

Despite the different behaviors between children, for all of them, the interest in the NAO robot decreases along the sessions, while the attention towards the therapist increases. This proves that the protocol has to be updated/adapted in order to keep the children engaged and stimulated.

Concerning the TFD towards elsewhere, it is higher than the expected for some of the children, which can be partially justified by the people passing in the scene during the sessions. These people attracted the children's attention and thus, the study should be done in a private space.

D. Further Insights

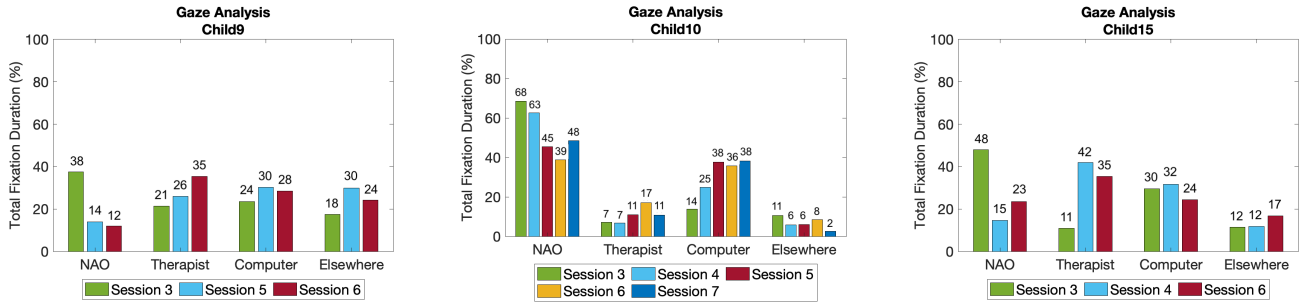
To take conclusions about the agreement between a quantitative and a qualitative analysis, the proposed system results and the qualitative therapist feedback were compared. Observing the model accuracy for each child, the model performs worse for Child 9, followed by Child 6 and 15 (Table VII).

TABLE VII
MODEL ACCURACY FOR EACH CHILD IN EACH SESSION

	Child 6	Child 9	Child 10	Child 15	Child 19
Session 3	80	76	84	84	93
Session 4	—	—	76	69	83
Session 5	81	75	82	—	—
Session 6	—	78	82	70	—
Session 7	—	—	87	—	—

Comparing with the qualitative analysis of the therapist for each session, presented in Table VIII, and the attention analysis, presented in Figure 6, some conclusions can be taken:

- Child 6 likes to touch the robot and offers resistance to the work. Consequently, he/she moves a lot, which causes Kinect detection problems and deteriorates the proposed framework performance;
- Child 9 interacts well with the robot, however likes to touch it and uses the scenarios to play. Thus, he/she also moves considerably, which causes Kinect detection problems and deteriorates the performance of the proposed framework. Due to the toys, the duration of looking elsewhere is also higher than for the other children, as shown in Figure 6a.
- Child 10 interacts well with the robot, being focused in the tasks, which justifies the higher performance scores. According to the therapist feedback, he/she is very interest in NAO, improving his/her performance along sessions, which is reflected in the model attention estimations showing a high attention towards NAO (Figure 6b).
- For Child 15, the reasons for the lower performance scores are not clear. However, he/she interacts with the robot despite his/her difficulties while performing the tasks. His/Her performance increases with therapist instructions and encouragements.



(a) Child 9 (b) Child 10 (c) Child 15

Fig. 6. Total Fixation Duration towards the targets and elsewhere along the sessions for Children (a) 9, (b) 15 and (c) 10 [%]

TABLE VIII
THERAPIST QUALITATIVE ANALYSE OF THE SESSIONS. T.: LIKES TO TOUCH; R.: OFFERS RESISTANCE TO; P.: PERFORMANCE; I.: INTERACTION

	Session 3	**Session 4**	**Session 5**	**Session 6**
Child 6	T. NAO R. work	—	R. work	—
Child 9	T. NAO Good P.	—	T. NAO Good P.	T. NAO Good I.
Child 10	Good I. Low P.	Avg P.	Good I. High P.	Good I. High P.
Child 15	Good I. Low P.	Low P.	—	Avg P.
Child 19	High P.	High P.	—	—

- Child 19 pays attention and does every task correctly. He/She is the one with the lowest level of ASD, justifying the higher performance scores.

Therefore, there is an agreement between the therapist feedback, the model performance scores and the model estimations of attention for each child. Moreover, after showing the Figures 6 to the therapist, she reported that the results obtained were in accordance with her expectations, specially for the Child 10, which had a level of attention towards NAO much higher than the other children. Overall, these considerations are a demonstration of the possibility of using this framework as an explainable AI tool.

IV. CONCLUSIONS AND FUTURE WORK

ASD children show deficits in attention, which can influence their ability to learn new skills. Assessing their attention during triadic therapy sessions with SAR and the therapist is a major prerequisite to provide a more complete overview of the therapy and supplement the therapist feedback.

This work presented a pipeline for a quantitative attention analysis of ASD children based on their gaze during a robotic therapy. The complexity of this task was dictated by the characteristics of the environment and of the children participating in these therapies. Specifically, the need of using only non-intrusive devices with ASD lead to the selection of a camera which was placed at a certain distance from the children, hardening the gaze estimation. The proposed system was composed by the estimation of the gaze and definition of the AOIs, followed by a gaze classification into the different targets. To extract the gaze the Gaze360 model was used. To define the AOIs, two approaches were analysed: a geometrical approach and a learning approach. Given the performance metrics of the model, the learning approach was chosen as the best, achieving the main goal of this work, with the proposed model reaching a total accuracy higher than **82.0%** for all the sessions. Comparing with the state of the art, our system performed better than the one proposed in [15], which was based on the head pose and achieved an accuracy of 73.5%. Moreover, for all the children in all the sessions, the model had accuracies between 69.0% and 93.0%. On the other hand, the qualitative analysis of the therapist was similar to the quantitative results, demonstrating that these quantitative measures captured the therapist assessment and could be used as therapeutic evaluation measures. Furthermore, the understandability of these metrics by the therapist proved the capability of this framework to be an explainable AI tool.

Future work could pass first by the development of new clinical studies, which should include more participants and sessions, for a better understanding of the children's attention patterns. Regarding the proposed model, there are some problems, mainly concerning the Densepose head bounding boxes and the Kinect skeletons detection. For the Densepose bounding boxes, there are two main problems. The model has a slow processing taking too long to extract the gaze from the videos, making it impossible to analyse the gaze online. Moreover, the head detection has errors, outputting bounding boxes that cover the whole body instead of just the head, resulting in noisy estimations of the Gaze360 model. To solve both problems, other head detectors should be studied. The wrong detection of the skeletons by the Kinect is related with the incorrect distinction of the red scarf used by the therapist and with the lower Kinect rate calculating the skeletons. The first condition can be surpassed by using a brighter and unusual color to have an easier detection and by saving all the detected skeletons in each frame, only distinguishing them offline, comparing the skeletons and the detected object positions. The Kinect lower rate computing the skeletons could be due to simultaneously programs running, namely the NAO control system. Therefore, the performance of other acquisition cameras should be studied, as well as offline skeleton detection methods. Moreover, the therapist and

child should be standing during the sessions to improve the computation of the skeletons by the Kinect.

The model hyperparameters showed to be very important to increase the model performance, mainly when the Gaze360 estimations present intrinsic errors resulting in offsets for looking at each target. To solve this problem and improve the Gaze360 estimations, the camera should be placed at the eyes level. The attention towards the computer is also higher than the expected, even in the sessions without Level 4. Therefore, the computer is considered a distraction and it should be removed from the scene. Consequently, the scenarios should be projected behind the camera.

Lastly, the adjustment of the quantitative attention model to a real-time scenario could be explored for the creation of protocols that adapt according to the ASD children attention. These protocols would be customized for each child, engaging him/her more, consequently, improving his/her performance and learning process. On the other hand, the cognitive behavior, expressed by the attention, could be related with the affective and behavior engagements. To achieve this goal, new quantitative models could be created for the evaluation of the children facial expressions (affective) and of their performance in the imitation tasks (behavioral) [7]. In this way, it would be possible to have a complete overview of the effect of these therapies in the ASD children.

REFERENCES

[1] K. A. Shaw, M. J. Maenner, J. Baio, EdS1, A. Washington, D. L. Christensen, L. D. Wiggins, S. Pettygrove, J. G. Andrews, T. White, C. R. Rosenberg, J. N. Constantino, R. T. Fitzgerald, W. Zahorodny, J. Shenouda, J. L. Daniels, A. Salinas, M. S. Durkin, and P. M. Dietz, "Early identification of autism spectrum disorder among children aged 4 years - Early autism and developmental disabilities monitoring network, six sites, united states, 2016," *MMWR Surveillance Summaries*, vol. 69, no. 3, pp. 1–11, 2020.

[2] P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia, "Autism and social robotics: A systematic review," *Autism Research: official journal of the International Society for Autism Research*, vol. 9, no. 2, pp. 165–183, 2016.

[3] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research," *Annual review of biomedical engineering*, vol. 14, no. 1, pp. 275–294, 2012.

[4] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, H. Ishiguro, M. Kikuchi, T. Sumiyoshi, and M. Mimura, "Optimal robot for intervention for individuals with autism spectrum disorders," *Psychiatry and Clinical Neurosciences*, vol. 74, no. 11, pp. 581–586, 2020.

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

[6] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder - A personalized computational modeling approach," *Frontiers in Robotics and AI*, vol. 7, p. 43, 2020.

[7] O. Rudovic, J. Lee, L. Mascarell Maricic, B. Schuller, and R. Picard, "Measuring engagement in robot-assisted autism therapy: A cross-cultural study," *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017.

[8] 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, 2020.

[9] A. P. Costa, L. Charpiot, F. R. Lera, P. Ziafati, A. Nazarikhorram, L. Van Der Torre, and G. Steffgen, "More attention and less repetitive and stereotyped behaviors using a robot with children with autism," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 534–539.

[10] H. Kumazaki, Y. Yoshikawa, Y. Yoshimura, T. Ikeda, C. Hasegawa, D. N. Saito, S. Tomiyama, K.-m. An, J. Shimaya, H. Ishiguro, Y. Matsumoto, Y. Minabe, and M. Kikuchi, "The impact of robotic intervention on joint attention in children with autism spectrum disorders," *Molecular Autism*, vol. 9, no. 1, p. 46, 2018.

[11] S. M. Anzalone, J. Xavier, S. Boucenna, L. Billeci, A. Narzisi, F. Muratori, D. Cohen, and M. Chetouani, "Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment," *Pattern Recognition Letters*, vol. 118, pp. 42–50, 2019, cooperative and Social Robots: Understanding Human Activities and Intentions.

[12] F. S. Alnajjar, M. L. Cappuccio, A. M. Renawi, O. Mubin, and C. K. Loo, "Personalized robot interventions for autistic children: An automated methodology for attention assessment," *International Journal of Social Robotics*, vol. 13, pp. 67–82, 2021.

[13] Y. Yoshikawa, H. Kumazaki, Y. Matsumoto, M. Miyao, M. Kikuchi, and H. Ishiguro, "Relaxing gaze aversion of adolescents with autism spectrum disorder in consecutive conversations with human and android robot - A preliminary study," *Frontiers in Psychiatry*, vol. 10, p. 370, 2019.

[14] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 6, pp. 682–691, 2016.

[15] G. Nie, Z. Zheng, J. Johnson, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 1–4.

[16] G. bin Wan, F. hao Deng, Z. Jiang, S. Lin, C. lian Zhao, B. Li, G. Chen, S. hong Chen, X. hong Cai, H. bo Wang, L. ping Li, T. Yan, and J. Zhang, "Attention shifting during child-robot interaction: A preliminary clinical study for children with autism spectrum disorder," *Frontiers of Information Technology & Electronic Engineering*, vol. 20, pp. 374–387, 2019.

[17] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018.

[18] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6911–6920, 2019.

[19] I. K. Riza Alp Güler, Natalia Neverova, "DensePose: Dense human pose estimation in the wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[20] 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 Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 1223–1232, 2021.

[21] A. P. Association, *Diagnostic and statistical manual of mental disorders, Fifth Edition*. Arlington, VA, 2013.

[22] Y. Zhou and J. Gregson, "WHENet: Real-time fine-grained estimation for wide range head pose," in *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

[23] L. Hanson, L. L. Sperling, G. Gard, S. Ipsen, and C. O. Vergara, "Swedish anthropometrics for product and workplace design." *Applied ergonomics*, vol. 40, no. 4, pp. 797–806, 2009.

[24] C. Clifton, F. Ferreira, J. M. Henderson, A. W. Inhoff, S. P. Liversedge, E. D. Reichle, and E. R. Schotter, "Eye movements in reading and information processing: Keith Rayner's 40 year legacy," *Journal of Memory and Language*, vol. 86, pp. 1–19, 2016.