2022 International Conference on Rehabilitation Robotics (ICORR)
25-29 July, 2022, Rotterdam, The Netherlands

Sharing Worlds: Design of a Real-Time Attention Classifier for Robotic Therapy of ASD Children*

Laura Santos[1,2], Bárbara Silva[2], Filippo Maddaloni[1], Alice Geminiani[1,3], Arianna Caglio[4], Silvia Annunziata[4], Ivana Olivieri[4], Catarina Barata[2], José Santos-Victor[2], Alessandra Pedrocchi[1]

2022 International Conference on Rehabilitation Robotics (ICORR) | 978-1-6654-8829-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICORR55369.2022.9896506

Abstract— **Joint attention is the capacity of sharing attention between two agents and an aspect of the environment, through the use of different cues, namely gaze. This capacity is of paramount importance for social skills. People with Autism Spectrum Disorder (ASD) present certain deficits in joint attention. Therefore, there is an increasing interest in finding therapies to improve this skill. Some of these therapies include robots since they are known to be attractive to people with autism due to their motivation ability and predictability when compared with humans. In this line, we have designed a real-time attention classifier for a triadic robotic therapy, using Gaze360 and geometrical considerations of the scene. We were able to classify the gaze of the therapist and the one of the child during the whole session, even in a highly unconstrained scenario with a single camera, achieving a mean accuracy of 59%. This classifier can be used for the measurement of joint attention, an important metric for the development of adaptive robotic therapies, where increasing levels of difficulty and engagement are provided dependent on the ASD children, who are characterised by high heterogeneity. Future work will pass by the calculation of this metric and integration on a robotic platform for ASD therapy to understand the impact of these robotic therapies in improving ASD symptoms, specifically on how ASD children share their attention with other people present in the rehabilitation scenarios.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterised by communication and social deficits and the presence of repetitive behaviours. Among ASD children, symptoms are very heterogeneous, therefore there is not a unique standardised treatment that can help all individuals. Recently, robots have been introduced in treatment protocols for children with autism thanks to their predictability and rule-based functioning, which make social interactions easier for ASD children [1].

One of the social skills targeted with robotic therapies is joint attention. Joint attention refers to the ability to share attention between a person and a social partner on an aspect of the environment (object or people) by acts of eye-gazing, pointing or other verbal or non-verbal indications [2]. Joint attention seems to be impaired in ASD children who tend to produce less declarative pointing gestures than typically developing children [3]. However, early intervention on joint attention can lead to positive outcomes in other developmental skills, such as communication and social skills [4].

In robotics therapies, either joint attention is trained specifically or it improves as an outcome of other trainings. In the first case, the child should look at target images when prompted by the robot [2], [5]. These scenarios tend to be very constrained with a fix child position and the use of multiple cameras. Such setup prevents the implementation of these therapies in clinical practice. In the second case, the joint attention is evaluated while the subjects are performing other therapies (imitation therapy [6], narrative skills training[3]). In these cases, the subject can move freely, in an unconstrained scenario, and the number of cameras is reduced (one or two) to not influence the therapy.

Regarding the measurement, joint attention is usually evaluated indirectly from gaze patterns of the participants. The measures can be obtained manually (e.g. one or two people analyse the frames recorded during the session [1], [2]) or automatically so that they are more objective, easier to obtain and can be included in the robot's control loop for adaptive therapies [5]. Gaze measures can be obtained using specific devices as eye-trackers [7] or algorithms for gaze estimation through RGB cameras. Methods for the indirect calculation of the gaze include estimation through the orientation of the head [5] or using facial landmarks [8], whose performance is limited by possible occlusions of the eyes.

Therefore, more advanced algorithms have been developed as Gaze360 [9], WHENet [10] and RT-Gene[11], allowing gaze estimation even if just part of the eye is visible. Gaze 360 consists in long short-term memory cells, where the output for one frame is dependent on the previous and following frames. Thus, even if the gaze is occluded, its estimation is still possible based on previous frames. Gaze360 receives as input a cropped image of the face and outputs a gaze direction estimation in terms of azimuth and elevation with respect to the camera reference frame. This algorithm was tested by [12] for the analysis of the eye-contact of ASD children during standard dyadic therapy. To our knowledge, this is the only work where the algorithm is tested in an unconstrained environment with a single camera.

In this work, our main goal is to design a real-time attention classification system. This classification system should be used in the future for the measurement of joint attention during a robotic therapy for ASD children, facilitating the

*This work was supported by the Portuguese Foundation of Science and Technology (projects SFRH/BD/145040/2019) and Associazione Paolo Zorzi (project *IOGIOCO*)

[1]NEAR-Lab, Department of Electronics, Information and Bioengineering, Politecnico di Milano, Via Ponzio, 34, 20133 Milano, Italy laurajoana.espinosa@polimi.it

[2] Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal.

[3]Department of Brain and Behavioral Sciences, University of Pavia, Pavia, Italy.

[4]CARELab, IRCCS Fondazione Don Carlo Gnocchi, Milan, Italy.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.

implementation of adaptive protocols, protocols in which the robot tasks are changed according to the attention of the children. The robotic therapy consists of several interactive turn-taking games between a humanoid robot, the child and the therapist (IOGIOCO protocol) for the training of gestures [13]. In our therapy, the subjects move freely and just one camera is used to reduce the setup times, which is required in clinical practice, and due to a space limitation (the therapy room dimensions: 5x3.5 m). We choose to apply Gaze 360 and we aim at bringing it into unconstrained robotic therapies scenarios, where targets can occlude other targets (e.g. therapist in front of the robot). Although Gaze360 method is prepared for unconstrained scenarios, it has long computational times, preventing real-time applications, specially due to the face detection process, provided by DensePose [14], an algorithm that fully reconstructs the whole body. Thus, our contributions in this work are mainly two:

- developing a method for choosing the best face detector to be associated with Gaze360
- implementing a full attention classification system for an unconstrained robotic therapy of ASD children (Figure 1).

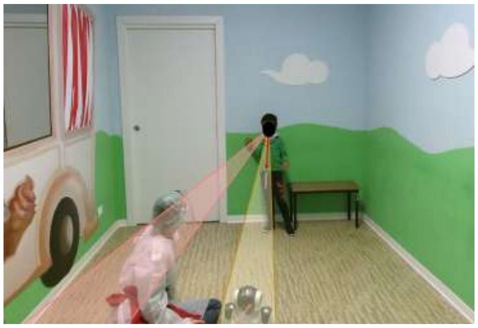

Fig. 1. Representation of the attention classification system, where the Therapist (pink cone) and NAO (yellow cone) areas of interest are compared with the subject gaze direction estimation (red arrow). In this case the system would classify as 'looking at the robot'.

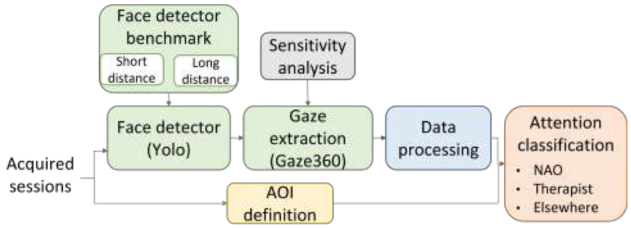

Fig. 2. Full overview of the attention classification system.

II. Methods

For achieving real-time gaze classification, first the best face detector was chosen considering the characteristics of the IOGIOCO protocol. Then, for specific acquisition sessions, the gaze was obtained in terms of an azimuth angle and was processed. Subsequently, this angle was compared with the areas of interest (AOI) of the different targets and classified according to them (patient looking at the robot or patient looking at the therapist). Finally, we performed a sensitivity analysis to understand the effect of the parameters of this gaze estimator (Figure 2).

A. Face Detector Benchmark

Gaze360 requires cropped images of the face to estimate the gaze. Therefore, we compared three face detection algorithms (YOLO [15], RT-Gene [11] and DensePose [14]) concerning the accuracy and the computational time to find the best match for our system. We expected the face detectors associated with Gaze360 to have different performances. Gaze360 is a neural network model with long short term units, thus dependent on the training images resolution and on the detection capability of the bounding boxes algorithms since it uses seven consecutive frames. RT-Gene and YOLO were chosen because they were already present in the literature: YOLO was associated with WHENet, and RT-Gene facial detector, based on Multi-Cascaded convolutional neural networks, was used in the whole architecture of RT-Gene. Both have been implemented for gaze estimation through the head pose. Here, we explored their behaviour in combination with Gaze360.

Since there are no standard protocols for the benchmarking of face detectors, we developed two validation procedures, which we tested on one subject: one at a short camera-subject distance and another at a long camera-subject distance. In the first, we compared the outcomes of the gaze estimation with the different face detectors with a gold-standard gaze estimation through an eye-tracker, Tobii T60, in a controlled scenario. In the latter, we analysed the algorithms in a condition closer to the therapeutic one. We needed the two validation scenarios because the gold-standard method could be only applied for distances lower than 80 cm, preventing its use in a long camera-subject distance setup.

During the first validation step procedure, the subject kept his head in a fixed position, putting the chin on the rest and performing only eye movements (Figure 3 (a)). The participant executed 5 validation acquisitions. In each validation sequence (Figure 3 (b)), 13 different points appeared one at a time for two seconds on the Tobii T60 screen, for a total of 14 steps (the central point was displayed twice). Knowing the dimensions of the Tobii screen, we converted the validation dots' coordinates to pixels, and their sequence constituted the expected signal. Then, we calculated the Root Mean Square Error (RMSE) between the gaze direction estimated by Tobii T60, the several algorithms (DensePose+Gaze360, YOLO+Gaze360, RT-Gene+Gaze360) and the expected signal, to evaluate their performance.

For the long camera-subject distance validation step, four points were fixed in different positions of a room (Figure 3 (c)). The subject looked at each of them, moving only the eyes, according to a sound played every 10 seconds. From the room geometry (Figure 3 (d)), an expected signal of azimuth was calculated to compare the algorithms' performance. A cross-correlation process was applied between this expected

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.

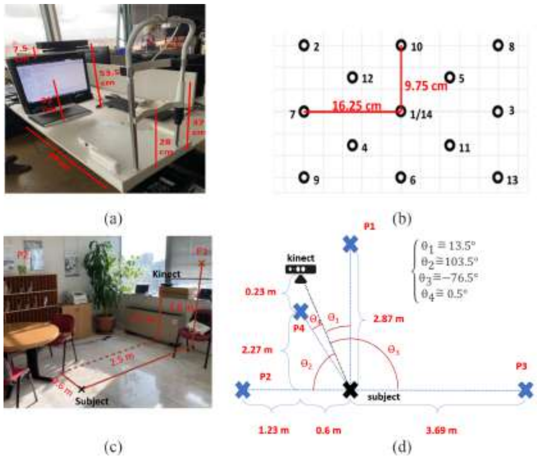

Fig. 3. Benchmark setups for the validation of the different face detection algorithms.

signal and the several estimated signals to adjust the delay between the sound and the actual point of interest switch. The algorithm outcomes were then evaluated, calculating the RMSE between the expected signal and the estimated azimuth.

After the two validation steps, YOLO was considered the best face detector method for our system due to a good balance of speed and accuracy.

B. IOGIOCO robotic therapy setup

In this study, attention was analysed indirectly through a robotic therapy done in Fondazione Don Gnocchi. The final objective of this therapy was to train semantic gestures (e.g. big, small, etc.). This study was approved by the Ethical Committee of Fondazione Don Gnocchi (date: 28/08/2019).

For this case study, the humanoid robot NAO (SoftBank Robotics) was chosen, since it has 25 degrees of freedom, being able to produce the different gestures. A Microsoft Kinect camera recorded the participant's movements. We placed the camera above the robot to capture both the subject with ASD and the therapist who were in front of the robot. Since it is a depth camera, it is able to estimate 3D joints' coordinates of the two people, also known as keypoints. These coordinates were used for controlling the robot to mirror each subject (further detail in [13]).

The therapist was tracked through a red t-shirt she wore during the session. A video recording was maintained throughout the therapy session using the same camera. Simultaneously, the 2D joint coordinates in pixels, calculated by the Kinect for each frame, were registered (called joint points from now on).

C. Clinical acquisitions

Two sessions of two children were analysed to study the impact of the gaze estimator. The numbers 6 and 7 were attributed randomly to identify them anonymously. Both were preschooler patients (under 6 years of age), with a clinical diagnosis of Autism Spectrum disorder, confirmed by experienced child neurologists according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition [16], and confirmed by Autism Diagnostic Observation Schedule – Second edition (ADOS) [17] and Autism Diagnostic Interview-Revised (ADI-R) [18]. The results for each patient are reported in detail in Table I. Patient 7 is a male, non-verbal patient with a global developmental delay who needs substantial support in everyday life. Difficult to engage in the robot activities, the functional play was absent. Patient 6 is a female with phrasal-speech language. This patient has a mild developmental delay with global better performances. She is more collaborative and more interested in the activities proposed by the therapist.

TABLE I
AUTISM ASSESSMENT OF THE CHILDREN CHOSEN. INT: RECIPROCAL SOCIAL INTERACTIONS, COM: LANGUAGE/COMMUNICATION, BEHAV: REPETITIVE BEHAVIORS/INTERESTS, SA: SOCIAL AFFECT, RRP: RESTRICTED AND REPETITIVE BEHAVIORS, CS: COMPARATIVE SCORE

	ADI-R			ADOS-2			
Child	Int	Com	Behav	Module	SA	RRP	CS
6	1.4	1.5	1.3	2	0.8	1	6
7	1.87	1.7	0.8	1	1.1	1.75	6

During all sessions, their regular therapist was present. The whole intervention had an increasing difficulty. In the first level, the ASD subject was familiarised with the robot functionalities(sounds, lights, movements). In level 2, the subject could move freely and be mirrored by the robot. In the third level, a triad was created between the robot, therapist and ASD child to specifically train gestures. The description of the full intervention (protocol and inclusion and exclusion criteria) can be found in [19].

Child 6 did the three levels during the analysed sessions, while child 7 performed the two initial levels. The duration of the sessions depended on the engagement of the child. These sessions happened in January of 2021 (child 6) and November of 2020 (child 7), during the pandemic period, thus, the therapist used personal protective equipment, including mask and visor, that influenced the perception of her gaze.

D. Gaze Extraction and Data Processing

From each video recorded during the therapy, the gazes of the participants were estimated. Each frame was passed to YOLO, which produced two bounding boxes. The images were cropped and provided to the Gaze360 algorithm that estimated the gaze.

Then, the gaze was attributed to each subject, using the centre of the bounding box and the joint point of the head tracked by the Kinect. If the joint point of the child was closer to the centre of the bounding box than the joint point of the therapist, the estimated gaze was considered to belong to the child. All the frames without the joint points of both the therapist and the child were discarded. In the end, the azimuth of both participants was filtered through a moving average filter with a window of 10 samples.

978-1-6654-8829-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.

E. Areas of Interest Definition

After obtaining the gaze orientation expressed by the azimuth angle (α), this angle should be compared with the different areas of interest (AOI), to identify which object/person the subject was looking at. For each participant, two targets were defined: robot and other person (therapist or patient). Each area of interest t was defined by two angles (a right angle, ϕ_{rt} and a left angle, ϕ_{lt}), depending on the width of the target and its position in space. Therefore, the participant was looking at the target if $\phi_{rt} < \alpha < \phi_{lt}$.

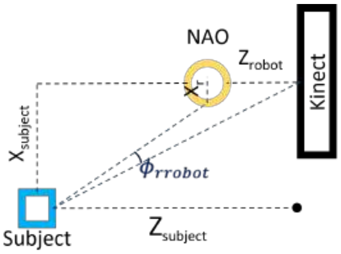

Fig. 4. Definition of the area of interest

The definition of these angles considered the geometry of the scene. For example, for the robot, through its position (X_{robot}, Z_{robot}), the ϕ_{rrobot} was given by Equation 1, where the $width$ was established by the physical dimensions of the robot (Figure 4) and $X_r = X_{robot} - \frac{width}{2}$.

$$\phi_{rrobot} = \arctan\left(\frac{X_{subject} - X_r}{Z_{subject} - Z_{robot}}\right) \quad (1)$$

X_{robot} and Z_{robot} changed at each session and were calculated through iterative process using the projection equation of the camera $c_{robot} = P * C_{robot}$, where P is the calibration parameters matrix, $C_{robot} = (X_{robot}, Y_{robot}, Z_{robot})$ the 3D coordinates vector of the robot and $c_{robot} = (u_{robot}, v_{robot}, 1)$ the corresponding 2D pixel coordinates vector. The P matrix was obtained through a Direct Linear Transformation using 6 keypoints (3 of the therapist and 3 of the child) in 3D and the corresponding jointpoints in 2D. Then, the value of Y_{robot} was fixed to -0.6m, since the camera was always at the same position and the height of the robot was known. Subsequently, for each session, different values of X_{robot} and Z_{robot} were tested until the correspondent projection in the image matched the head of the robot, which was verified by an operator. These values were then used in Equation 1 to obtain the AOI of the robot.

Regarding the other target (therapist and patient), the positions of each participant were used to establish the area of interest angles in relation to the Kinect camera. In this case, the positions of their heads were considered as the centre, and their $width$s were obtained by the difference between their shoulders positions. All these positions were given directly by the Kinect.

F. Classification performance

After establishing the areas of interest, the attention of each subject was classified and compared with the ground truth to evaluate the algorithm performance.

The ground truth was constructed by one operator who labelled the gaze of both participants in 700 frames of each therapy session. For each frame, the subject's gaze was classified as looking at the robot, looking at the therapist/patient, looking elsewhere or none of the above if the operator could not classify that frame. Then, we calculated the accuracy by comparing the estimated gaze prediction and the label from the operator. This accuracy was computed in relation to each target, so four performance indexes were obtained for the description of the algorithm (Patient looking at the therapist, Patient looking at the robot, Therapist looking at the Patient, Therapist looking at the robot).

Moreover, another independent rater labelled the participants' gaze in 350 of the 700 labelled frames to study the complexity of the acquisition scenario. Then, the Cohen's Kappa coefficient was computed to evaluate the agreement between the two raters.

G. Sensitivity analysis

Associated with the azimuth estimation (α), Gaze360 provides a confidence error angle σ such that $[\alpha - \sigma; \alpha + \sigma]$ covers the 10% quantile up to the 90% quantile range of the probability density function of the estimated angle α. Thus, higher values of σ lead to more uncertainty in the estimation.

To understand the impact of this parameter, we defined a threshold to discard all the estimations with a higher value of uncertainty ($\sigma > threshold$). The threshold was set in a progressively increasing range from 0 rad till 1 rad.

III. Results and Discussion

A. Face Detection Benchmark

The results of the short distance validation process comparing the expected signal with the signal estimated by Tobii T60 and the several algorithms are shown in Table II.

TABLE II

RMSE of short distance validation process [pixel]

	Mean ± Std. Dev. [px]
Tobii T60	154 ±33
DensePose + Gaze360	246 ±27
YOLO + Gaze360	420 ±80
RT-Gene + Gaze360	516 ±131

Although the results presented were just for one subject, Tobii T60, as expected, provided the most accurate performance for the gaze direction estimation. From this first validation experiment, RT-Gene provided the less accurate estimation, while DensePose drove the Gaze360 algorithm to the best accuracy among the other facial detectors.

For the long camera-subject distance procedure, the results of the RMSE after the cross-correlation process are reported in Table III, with the computational times of each face detector. From these results, RT-Gene still had the worse RMSE error, while DensePose and YOLO had similar errors. However, the YOLO algorithm takes five times less time to obtain a bounding box of the face. Therefore, YOLO was the chosen face detector for our attention system. Future work

978-1-6654-8829-7/22/$31.00 ©2022 IEEE

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.

should pass by testing with more subjects to increase the robustness of this conclusion.

TABLE III

RMSE and Face detection time of long distance acquisitions [deg].

	RMSE[deg]	Face detection time [s]
	Mean±Std. Dev.	Mean±Std. Dev.
DensePose + Gaze360	23.18 ± 2.65	0.335 ± 0.002
YOLO + Gaze360	24.24 ± 1.76	0.064 ± 0.008
RT-Gene + Gaze360	36.31 ± 16.27	0.069 ± 0.001

B. *Classification performance*

From the manual labelling performed by the two independent raters, Cohen's kappa coefficients of three sessions were calculated and are shown in Table IV. In general, the coefficients regarding the patients were lower than the coefficients of the therapist, highlighting the intrinsic complexity in the detection of the patient gaze direction during the therapy sessions. The lowest values refer to the child observations in the second session of subject 6 and in the first session of subject 7, reflecting a weak agreement [20]. Regarding the therapist, the agreement is classified as moderate and not as strong, which was expected for a healthy subject, due to the protective material (e.g. face mask) which makes the discrimination of the therapist's gaze harder.

TABLE IV

Inter-rater reliability - Cohen's kappa coefficient. NA represents the session which was not labelled by the second rater.

	Patient [%]	Therapist [%]
Child 6 - Session 1	81	91
Child 6 - Session 2	50	65
Child 7 - Session 1	44	67
Child 7 - Session 2	NA	NA

TABLE V

3D estimation accuracy overview for the two sessions (S1 and S2) of the two children. T represents the therapist, C , the child and R, the robot. The arrow indicates the object of interest, namely T ->C represents the Theraspist looking to the Child.

	Child 6[%]		Child 7[%]		
	S1	S2	S1	S2	Mean± Std. Dev. [%]
T ->R	62	77	53	59	63 ±9
T ->C	57	46	59	49	53 ±5
C ->R	66	63	65	53	62 ±5
C ->T	64	58	54	59	59 ±4

Table V shows the overall algorithm performance for each analysed session. The two worst outcomes were verified for the therapist looking at the child in Session 2 of both children. The proximity of the therapist to the camera in some frames of these acquisitions can justify these results. Thus, her face was completely covered and not detectable by the face detection algorithm every time she looked at the patient. In some frames, we verified that YOLO did not detect any face. Therefore, a deeper analysis of the facial detection algorithm may lead to a better estimation accuracy since the Gaze360 algorithm would work on a more continuous temporal sequence.

Moreover, in three of the four acquisitions the accuracy of the gaze of each participant towards the robot was higher than the one towards the 'other person', which is a reflex of the robot's static position. Contrary to the 'other person', who was always moving and their AOI depended on the Kinect detection, the robot had a fixed position, being the calculation of its AOI, probably, more precise. Consequently, the accuracy towards this target was higher in the majority of the acquisitions.

Comparing the results of Child 6 and of Child 7, the accuracies are higher for the former, which can be ascribed to the children's characteristics described in Section IIC. Child 6 has a mild developmental delay and was more collaborative than Child 7, namely, she was more still during the sessions, which influenced the Kinect detection algorithm, the face detection algorithm, and consequently the whole attention classifier system.

Furthermore, comparing Table V and Table IV for the first session of subject 6, the best Cohen's kappa coefficient values correspond to the best algorithm accuracy performances for what concerns the child gaze direction estimation. Thus, the algorithm outcomes are consistent with the manual labelling difficulty level.

C. *Sensitivity analysis*

Through the implementation of a threshold on the confidence error angle extracted from Gaze 360, the accuracy of the attention classification system changed (Figure 5 (a)). Overall, the accuracy decreases with threshold increment since the uncertainty is higher in the calculation of the gaze. Simultaneously the number of accepted frames should be considered since a low threshold is associated with a low number of accepted frames (Figure 5 (b)). Therefore, the tuning of this parameter is mandatory for evaluating the patient's attention during therapy.

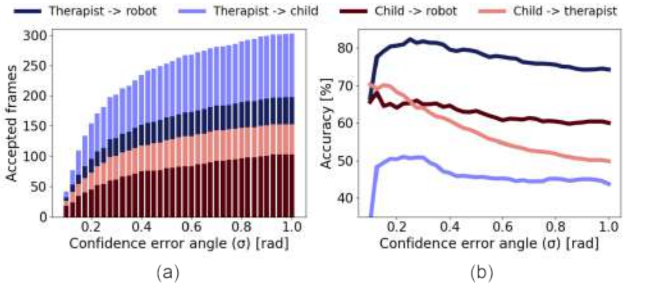

Fig. 5. Evolution of the performance of the algorithm with the variation of the threshold on the confidence error angle (a) and respective number of accepted frames (b).

IV. Conclusions and Future Work

We developed a full attention classification system for an unconstrained environment with a single camera setup. It

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.

achieved a minimum mean classification accuracy of 53% and a maximum of 63%. Differences between the child and the therapist and between the two children were found, probably reflecting their different levels of Autism. However, the number of children tested and sessions analysed was very small, preventing the statement of definitive conclusions. In addition, the definition of the ground truth demonstrated the challenges associated with the estimation of the child gaze but also compromised the robustness of the results obtained. Based on these preliminary results, we plan in future work to apply our system to a larger sample of children and sessions to monitor their progress. Moreover, a discussion phase should be included in the ground truth definition: the raters should establish an agreement regarding the frames harder to classify or decisively eliminate those frames from the analysis of the accuracy, to increase the results' robustness.

Compared to the literature, our work had to address additional challenges that justify in part its performance difference: it was developed for a triadic situation and not dyadic as the one of [12] and in an unconstrained scenario, contrary to [4]. Then, the therapist had to wear protection material due to the pandemic situation, adding complexity to the conditions of therapy scenarios. On the other hand, the main limitations of the algorithm could be ascribed to the face detection part and the robot position identification. As noticed by [12] and verified here, Gaze 360 is extremely influenced by a continuous identification of the face and more specifically by the proportion of the face size with respect to the whole input image. Possible solutions could be: (i) increasing the image resolution; (ii) applying a filter that uses previous cropped faces in case the new ones are not available; (iii) including the Kinect in the face detection, providing the cropped image correspondent to the detected skeleton to the face detectors, which would increase their performance. Regarding the robot position, an alternative could use the depth map of the Kinect and a mask for detecting the robot in the colour image(frame of the video) to obtain the precise robot location during the therapy.

Future work will consider these solutions to have a more precise quantitative metric of the attention and consequently calculate the joint attention. These automatic quantitative metrics can simultaneously facilitate monitoring rehabilitation therapies and allow a more standardised comparison between them.

We will integrate this measurement system into the IO-GIOCO platform in real-time to drive a more adaptive therapy. Autonomous and adaptive robots are essential for children with autism since they are characterised by very different symptoms which change during their development. An adaptive robot can not just deliver a more personalised therapy but also evaluate the child engagement. In this way, the robot can propose other exercises or augment the level of the prompts to increase the child attention. Therefore this system could bring robots to the clinical practice and eventually to more unconstrained environments such as home or school, permitting a continuous rehabilitation. Furthermore, this type of robotic therapy with a triadic interaction could allow the participation of ASD children in group activities, bringing their world closer to ours.

References

[1] A. P. Costa, L. Charpiot, F. J. 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," *2018 IEEE RO-MAN*, 2018.

[2] 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," *Mol Autism*, vol. 9, 2018.

[3] W.-C. So, C.-H. Cheng, W.-Y. Lam, Y. Huang, K.-C. Ng, H.-C. Tung, and W. Wong, "A robot-based play-drama intervention may improve the joint attention and functional play behaviors of chinese-speaking preschoolers with autism spectrum disorder: A pilot study," *J Autism Dev Disord*, 2020.

[4] 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 RO-MAN*, 2018.

[5] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with asd," *IEEE T Hum-Mach Syst*, vol. 48, 2018.

[6] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, and S. Di Nuovo, "Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability," *Robotics*, vol. 7, 2018.

[7] 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," *Front Psychiatry*, vol. 10, 2019.

[8] G. bin Wan, F. hao Deng, Z. jian Jiang, S. zhao Lin, C. lian Zhao, B. Li, G. Chen, S. hong Chen, X. hong Cai, H. bo Wang, L. ping Li, T. Yan, and J. ming Zhang, "Attention shifting during child—robot interaction: a preliminary clinical study for children with autism spectrum disorder," *Front Inform Tech El*, vol. 20, 2019.

[9] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," *2019 IEEE/CVF ICCV*, 2019.

[10] Y. Zhou and J. Gregson, "Whenet: Real-time fine-grained estimation for wide range head pose," in *31st BMVC*, 2020.

[11] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in *ECCV*, 2018.

[12] 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, 2021.

[13] 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. Rehabilitation Eng.*, vol. 29, 2021.

[14] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *IEEE CVPR*, 2018.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016.

[16] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-V*, 4th ed., 2013.

[17] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism diagnostic observation schedule 2nd edition," *WPS*, 2012.

[18] C. Lord, M. Rutter, and A. Le Couteur, "Autism diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders," *J. Autism Dev. Disord.*, vol. 24, 1994.

[19] A. S. Ivani, A. Giubergia, L. Santos, A. Geminiani, S. Annunziata, A. Caglio, I. Olivieri, and A. Pedrocchi, "A gesture recognition algorithm in a robot therapy for ASD children," *Biomed. Signal Process Control*, vol. 74, 2022.

[20] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, 2012.

Authorized licensed use limited to: Politecnico di Milano. Downloaded on December 20,2022 at 19:32:46 UTC from IEEE Xplore. Restrictions apply.