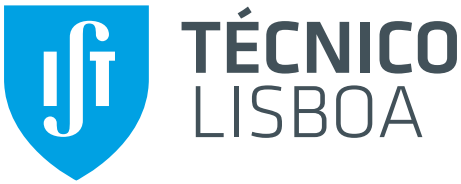


Gaze Analysis in Robotic Therapy for Autistic Children

Bárbara de Matos Águas Pereira da Silva

Thesis to obtain the Master of Science Degree in

Electrical and Computer Engineering

Supervisors: Prof. José Alberto Rosado dos Santos-Victor
Prof. Ana Catarina Fidalgo Barata

Examination Committee

Chairperson: Prof. João Fernando Cardoso Silva Sequeira
Supervisor: Prof. José Alberto Rosado dos Santos-Victor
Member of the Committee: Prof. João Miguel Raposo Sanches

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to thank my supervisors Prof. Catarina Barata and Prof. José Santos-Victor for giving me the opportunity of developing this thesis and for all their support and help. Special thanks to Laura Santos, for all her availability and immense help during the last months, answering all my questions, independent of where she was or the time she had.

A huge thanks to APPDA Lisboa, in particular to Sara Ferreira, for their collaboration, giving me the opportunity of performing the therapeutic sessions with the robot.

Lastly, I would like to thank my family and friends for supporting me at all moments.

Abstract

This thesis 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 thesis 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.

Keywords: Autism Spectrum Disorder (ASD), Social-Assistive Robots (SAR), Attention, Gaze tracking

Resumo

Esta tese tem como objetivo desenvolver um modelo quantitativo para avaliar a atenção, através do olhar, de crianças com autismo, durante sessões terapêuticas com Robôs-Socialmente-Assistivos (SARs).

Crianças com austimo apresentam graves défices de atenção que dificultam a sua capacidade de aprender novas competências. A avaliação automática da sua atenção forneceria aos terapeutas um bi-omarcador importante para quantificar melhor o seu comportamento e monitorar o progresso/evolução. Estudos anteriores focaram-se em tarefas de interação humano-computador, com dispositivos baseados em ecrãs, que distrairiam o sujeito em protocolos terapêuticos com SARs.

A abordagem da tese combina a extração do olhar com a definição de Áreas-de-Interesse (AOIs), para caracterizar os períodos de atenção durante a sessão. A metodologia foi testada em crianças com autimo. Visto que extrair o olhar de imagens ópticas é bastante desafiante, vários métodos foram avaliados. O Gaze360, que depende de imagens da cara e aprendizagem-automática, provou ser o mais robusto. Para cada alvo (terapeuta, sujeito, robô), o tamanho horizontal/azimute da AOI (angular) foi definido usando duas alternativas: uma *abordagem-geométrica* que combina as dimensões dos alvos com o ruído estimado do Gaze360, e uma *abordagem-de-aprendizagem*. Tendo cada alvo associado a uma gama de ângulos-de-fixação, as estimativas do olhar são utilizadas para classificar o foco-de-atenção do sujeito. As nossas experiências mostram que a *abordagem-de-aprendizagem* supera a *abordagem-geométrica*, alcançando uma exatidão acima de 82.0%.

Por fim, vale mencionar que os terapeutas compreenderam os indices-de-atenção propostos e consideraram-nos alinhados com a sua própria avaliação, um estímulo para o futuro uso clínico do método proposto.

Keywords: Perturbações do Espectro do Autismo (PEA), Robôs-Socialmente-Assistivos (SAR), Atenção, Detecção do Olhar

Contents

List of Tables

List of Figures

Acronyms

AFD Average Fixation Duration.

AI Artificial Intelligence.

AOI Area of Interest.

ASD Autism Spectrum Disorder.

CNNs Convolutional Neural Networks.

JA Joint Attention.

RAT Robot-Assisted Therapy.

RJA Responsive Joint Attention.

RMSE Root Mean Squared Error.

ROC Receiving Operating Characteristic.

SAR Social-Assistive Robots.

SFC Sum of Fixation Count.

TD Typically Developed.

TFD Total Fixation Duration.

Chapter 1

Introduction

1.1 Motivation

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 [10]. 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 [11]. In order to improve the social and motor abilities of the ASD children, several therapeutic approaches have been used. Recently, the introduction of Social-Assistive Robots (SAR) has been studied to improve the social and motor capacities of these children [12]. SAR are usually able to attract the children's attention and interest, due to their simple and repetitive movements. Multiple robots have been designed, with different appearances and functionalities, being tailored to the goals and conditions of each therapy [13].

Assessing the children's engagement during the therapies is an important task, 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 achieve better outcomes [14]. One of the main engagement indicators is the attention, which is often compromised 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 [15]. Therefore, it is important to develop assessment tools for this capability.

1.2 Problem Statement and Goals

This thesis focuses on the development of an accurate quantitative system, able to evaluate the attention of ASD children during therapeutic sessions.

This problem is complex especially due to the children and therapy intrinsic characteristics. In general, ASD children are sensitive to intrusive devices, since they are uncomfortable and distracting. Thus, these sensors decrease their focus during the therapy and, consequently, their performance. Therefore, non-intrusive devices, like cameras, should be preferred. However, such devices can make the attention assessment harder, since the distance between a subjects face and these devices is considerable, in order to capture the whole scene. Furthermore, considering the therapeutic environments, which are unconstrained, participants are able to move freely, making tracking and attention assessment extremely challenging.

Using the data extracted from these devices, a secondary goal of this thesis is that the attention as-

sessment system 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.

1.3 Approach Overview

To achieve the ultimate goal of evaluating the ASD children's attention in robotic therapy, two clinical studies were analysed (Pilot study and School study) and a system based on the eye gaze estimations and the definition of an Area of Interest (AOI) for each target was proposed (Figure 1.1).

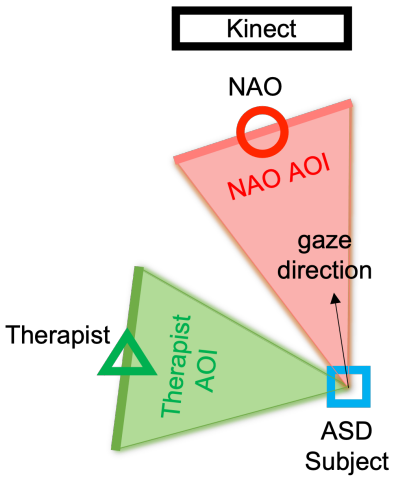


Figure 1.1: Representative setup, along with the Therapist (green) and NAO robot (red) AOIs and the subject gaze estimation (black arrow). The camera corresponds to the Kinect

Both clinical studies aimed to train gestures during triadic therapy sessions between the therapist, the ASD patient and the selected robot, NAO. The final setup, protocol and robotic system used for the therapeutic sessions resulted from a collaboration between two universities (Instituto Superior Técnico and Politecnico di Milano) and two clinical institutions (Associação Portuguesa para as Perturbações do Desenvolvimento e Autismo and Fondazione Don Gnocchi). The setup consisted of a triangle connecting the three entities (NAO, ASD child and therapist). A non-intrusive Microsoft Kinect camera was placed behind NAO to extract the therapist and ASD subject 3D joints coordinates and record the sessions. For the school study, an additional Computer was placed near the therapist to control the robot, while, in the pilot study, this was done by an operator. The protocol consisted of an imitation game with several levels to train gestures.

To analyse the attention during triadic sessions, the gaze direction given by an estimator was compared to the AOI. If the gaze was inside the AOI it was considered that the subject was looking to that target. Thus, an initial benchmarking of gaze and head pose estimators was done, followed by the development of a gaze analysis system. The benchmarking was done for both short and long distances. In the short distance tests, the gaze was estimated by the Tobii eye-tracker [16] and the OpenFace [5] and Gaze360 [6] models. In the long distance, the gaze estimation from Gaze360 was compared with the WHENet head pose estimation. At the end, the Gaze360, which relies on image face-datasets, was the chosen model [6].

Afterwards, the gaze analysis system started to be developed. The data acquired during the therapeutic sessions was preprocessed, with the missing data from the school study joints being reconstructed using interpolation. Then, the angles corresponding to fixating each target in each instant were calculated (standard angles), based on the scene geometry and the position of the participants, detected

by the Microsoft Kinect camera.

Finally, the Areas of Interest were defined around each target. They were determined in the horizontal direction (azimuth) only, given the scene geometry and the fact that most estimators are not accurate enough in the vertical direction (elevation). These AOIs were constructed with the goal of finding the range of angles corresponding to looking at each target. To define them, 2 approaches (geometrical and learning) were analysed. In the geometrical approach the widths of the AOIs were based on the geometry of the targets, while in the learning approach, the widths for the AOIs were defined by training the system to find the best width for the AOI of each target.

Having the proposed framework, the system was tested with data acquired for both studies. Given the high-accuracy obtained by the system in unconstrained environments, it was concluded that this thesis resulted in a successful system with a good quantitative analysis of the attention. Moreover, some attention indices, which can be easily understood by the therapist and external people, were obtained, proving the adequacy of this system as an explainable AI tool.

1.4 Thesis Outline

This thesis is organized in four additional Chapters.

Chapter 2 provides a literature review, detailing other research works related with this thesis. The existing works on quantitative attention measures are reviewed and compared. The keywords used in the literature search involved the concepts of autism, social robots, attention and gaze tracking.

Chapter 3 describes the therapy studies, the gaze estimators benchmarking and the proposed attention assessment system. The methods used for the data pre-processing, the scene geometry analysis and the AOIs definition are presented.

Chapter 4 includes the results obtained for the gaze estimators benchmarking and the proposed attention assessment system, along with their discussion.

The last chapter (Chapter 5) contains the conclusions taken from the work done in this thesis, along with the possible future directions.

Chapter 2

State of the Art

2.1 Social Robotics for Autism

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder, influenced by genetic and environmental factors, with unknown specific causes and biomarkers. It is characterized by persistent social and communication deficits, and the presence of repetitive and stereotyped behaviors and interests [17]. The Autism and Developmental Disabilities Monitoring (ADDM) Network estimated the prevalence of ASD to be 1 in 54 children aged 8 years old [17] and 1 in 64 children aged 4 years old [10], worldwide, in 2016, showing an increase in the last years. The ASD prevalence appears to be 4.3 times higher in boys than in girls [18]. Without no clear causes and a high heterogeneity in severity and nature of the symptoms, the diagnosis of early developmental symptoms and an early intervention is critical for positive long-term outcomes [11].

ASD can not be cured, but there are therapeutic approaches that help to improve abilities in individuals with autism. However, due to the clinical heterogeneity characterizing individuals with ASD, there is no best therapeutic approach, as one approach might work for one person and not work for another. There are behavioral treatments which can improve the quality of life. They are focused on improving social and communication skills, and promote independence. During these treatments, Applied Behavior Analysis (ABA) is included, in which positive behaviors are encouraged and negative behaviors are discouraged. Other therapeutic approaches employ technology to facilitate human-human social interaction, such as computer-assisted, virtual reality and robot-assisted [12].

Robots have been used in rehabilitation and therapy for physical deficits. However, individuals with autism require social assistance. In the last years, there has been a focus on developing robots to provide assistance to humans through social interaction. This field of research is called Social-Assistive Robots (SAR) [12]. One of the implementations of SAR is Robot-Assisted Therapy (RAT). RAT has gained attraction in the last two decades to become a promising application area in the autism [19]. Robots make social interactions easier for the children, due to their rule-based and predictable systems [20]. As consequence, individuals with ASD tend to achieve a higher degree of task engagement when interacting with robots than with human trainees [13]. Furthermore, while in general the affinity of humans to another person is stronger than to artificial objects, for ASD individuals this is not verified. ASD individuals sometimes show behaviors towards robots that individuals without ASD have towards humans [13]. Studies on robot-mediated intervention have demonstrated positive outcome in different social skills, such as communication, attention and imitation [21].

The recent studies have been mainly focused on developing robotic models able to reproduce human behaviors and stimulate pro-social behaviors and abilities for social engagement in ASD people. Multiple

appearances, features and functional capabilities have been developed, using engineering and clinical psychology [19]. Regarding the physical appearance, there are three main types of robots: humanoid, animal-like and machine-like. Although, the best robot for ASD individuals is still to be found [13], the humanoid robots are the most used. Due to their anthropomorphic physical shape, resembling a human but with simpler expressions, they have the capability of engaging children and interacting with them [12].

One of the humanoid robots widely used in research and therapy for autism is NAO [22], developed by Aldebaran-Robotics [23]. A comprehensive review of robotic technology for autistic children showed that among 68 robots, NAO was used in 30.5% of 208 studies [24]. NAO is a small and lightweight biped commercial humanoid, with $57cm$ height and a weight of $4.5Kg$. It has 11 degrees of freedom (DOF) for its lower limbs and 14 more for its upper parts (Figure 2.1). Its 25 DOF offer a great mobility, enabling to adapt to the environment [22]. NAO is capable of responding to the inputs in a fast way and performing a variety of tasks, such as walking, grabbing objects and standing-up by itself. It is also capable of producing light and sound stimulus, through its several LEDS in its eyes and ears and its loudspeakers in its ears [22]. It has two incorporated cameras, that provide an image resolution of $640 \times 480px$ at 30 frames per second, each. Given its simple appearance, it looks more approachable to children with ASD than other robots. Due to its success in imitation and attention tasks, it is the most preferable robot for Autistic children [23].

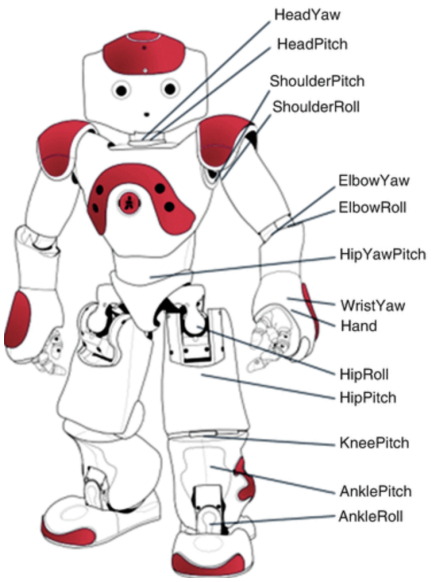


Figure 2.1: NAO robot Degrees Of Freedom (DOF) (extracted from [2])

Several studies using Social-Assistive Robots for ASD, have been focusing on a dyadic interaction, between the robot and the ASD subject [25, 20, 26, 27], which is a limitation for the translation to the daily life, since robot-human interaction can not be easily generalized to human-human interaction. Therefore, triadic interactions, between the robot, the ASD subject and another person should be considered [28].

Lately, it has been explored the use of humanoid robots to train not only social, but also motor capabilities [28]. To assess the children behavior and evaluate both the social and motor training, it is very important that non-intrusive systems are used, since ASD children can find physical sensors uncomfortable and distracting [25]. One of the most used non-intrusive systems is the Microsoft Kinect, referred as Kinect from now on [25, 29, 30, 27, 28].

The second version of Kinect is a 3D RGB-D sensor, composed by a RGB camera with a resolution of $1920 \times 1080px$ and an infrared camera with a resolution of $512 \times 424px$. It provides synchronized color and depth images at a $30Hz$ frequency. It is based on Time of Flight (ToF) technology to estimate the

distance to each point of the scene using active light pulses. Kinect is able to track markerless human-motion, up to 6 human bodies simultaneously. The skeletons are defined by 25 3D joints, as shown in Figure 2.2, according to the reference system in Figure 2.3 [31].

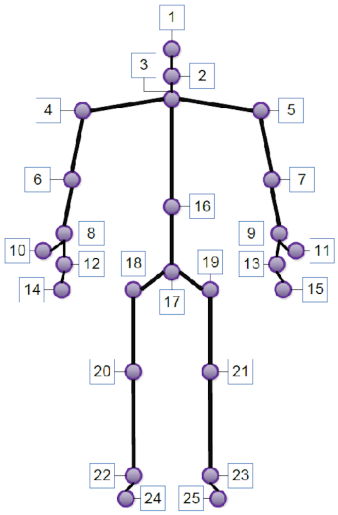


Figure 2.2: Skeleton joints calculated by the Kinect camera (extracted from [3])

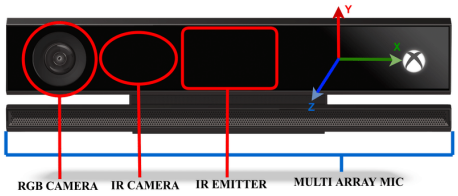


Figure 2.3: Kinect coordinate system and its components (extracted from [4])

2.2 Attention

An important factor to assess the quality of a robotic therapy is the engagement of the participants [26]. Engagement is described as the interaction and connection between a person and the environment in a developmentally appropriate manner, consisting of three forms: affective, behavioral and cognitive. Affective engagement is defined as the interest expressed through emotions. Behavioral engagement refers to involvement in the activities and on-task behaviors. Finally, cognitive engagement is usually described by the attention and is defined as the willingness to acquire and perform new skills [14, 32].

Children with ASD usually have deficits in social, joint and on-task attention. Social attention is determined by the focus given to social stimulus. Joint Attention (JA) is defined by the attention shared between two people towards visual stimulus. Usually it is divided into responding to other's attention (RJA) or initiating JA episodes (IJA). On-task attention is specified by the focus on a target object, while ignoring the distractions. Thus, the on-task attention is an important feature to analyse during the therapies, since it is a major prerequisite to the development of learning skills [15]. Until now, qualitative measures of attention have been used extensively in research and clinical practice, mainly through manual video coding [20, 33]. However, this process is time consuming. Thus, recent studies have focused on obtaining reliable, objective and quantitative measures of the attention based on the eye gaze [34], the head orientation [35] or the detection of facial landmarks [36].

To obtain a quantitative analysis of the gaze, there are mainly two parts: eye tracking and gaze estimation. The eye tracking consists in detecting the eyes and following their position frame by frame. The gaze estimation comprises the calculation and tracking of the gaze in 3D (gaze direction) and/or 2D (point of regard (PoR)). The gaze direction is usually expressed by 2 coordinates: azimuth and elevation. The PoR is determined by the intersection of the gaze direction with the closest object, on the image plane [37, 38]. To obtain a quantitative analysis of the head pose, the head tracking and the head pose estimation are required. Similar to the eye tracking, the head tracking consists in detecting the head and following its position frame by frame. The head pose estimation is usually expressed by 3 coordinates: pitch, yaw and roll. Finally, a quantitative analysis of the attention based on the facial landmarks uses just the facial features recognition to understand if a subject is interested in an object or not.

2.2.1 Gaze, Head Pose and Facial Landmarks Estimation Models

To analyse the attention, the first developed eye-tracking devices involved intrusive procedures, such as requiring the subjects to wear head-mounted devices [39]. Recently non-intrusive ways have been studied, with several image-based gaze estimators, head pose estimators and facial landmarks detectors being proposed [38]. While some of them are able to compute multiple features (OpenFace [5] and RT-Gene [7]), others are only able to compute the gaze (Tobii T120 Eye Tracker [16] and Gaze360 model [6]) or the head pose (WHENet model [8]). Tobii T120 is also an eye tracker, whilst the remaining referred models are only estimators, accepting RGB images as input.

The Tobii T120 Eye Tracker [16] is a real-time eye tracking system, integrated into a 17" TFT monitor with a resolution of $1280 \times 1024px$ (Figure 2.4). It can be used for several applications of eye tracking studies, as long as the stimuli can be presented on a screen. The eye-tracking system is completely non-invasive and does not constrain head or body movements. It is based on Pupil Centre Corneal Reflection and is composed by projectors and cameras located in the lower part of the screen and image processing algorithms. While the projectors send near-infrared light on the eyes, the cameras extract high resolution images of the user's eyes with a rated accuracy of $0.5°$ at a sampling rate of $60Hz$. In the end, the eyes' position and gaze point are calculated using image processing algorithms [16]. However, it is only effective at a distance of $55cm$ to $70cm$ from the participant. Therefore, other algorithms for long distance gaze estimation were developed.

Figure 2.4: Tobii T120 Eye Tracker

The OpenFace 2.0 [5] is an accurate open source tool capable of automatic facial behavior analysis. It can detect facial landmarks, estimate the 3D head pose and eye-gaze, and recognize facial action units, both in real-time and offline. It uses Convolutional Neural Networks (CNNs) to detect several facial landmarks. It can also identify landmarks in the eyes area connected with the iris and pupil, using a

Constrained Local Neural Field, and thus, calculate the associated gaze angles (azimuth and elevation), as shown in Figure 2.5. However, this estimation is not possible if some part of the eyes is occluded, covering only a range of approximately $180°$ [5]. Therefore, other algorithms were developed in the Computer Vision field, such as Gaze360.

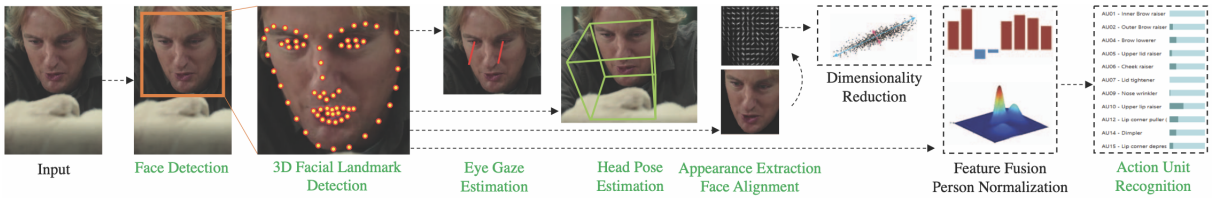
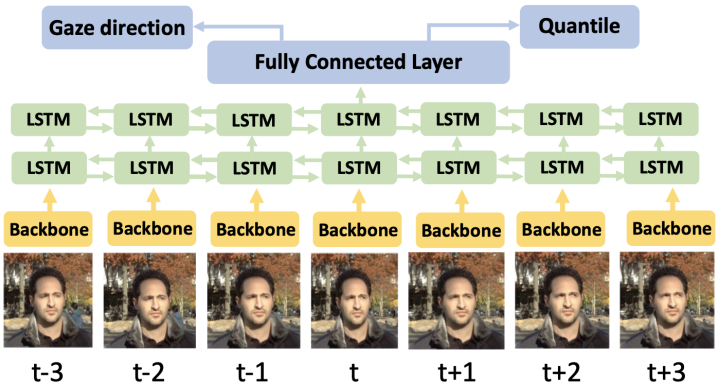
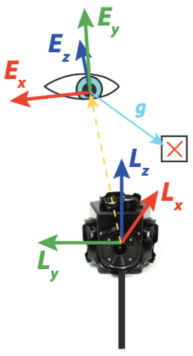


Figure 2.5: OpenFace 2.0 facial behavior analysis pipeline, including the eye gaze estimation. The outputs of the system are indicated in green (extracted from [5])

The Gaze360 model [6] is a method for robust 3D gaze estimation in unconstrained images. The Gaze360 model extends the previous models, being able to predict the gaze without visible eyes, through the inclusion of temporal information in the gaze estimations and the estimation of the gaze uncertainty. The proposed model, presented in Figure 2.6a, receives, as input, multiple cropped head frames, which are passed through a backbone network. To obtain the cropped head frames, the Densepose [40] 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 (Figure 2.7). This allows to calculate the gaze, even if the person turns his/her back to the camera. Thus, it is a full range gaze estimator, covering $360°$ [6].



(a) Model architecture



(b) Gaze estimation coordinate system

Figure 2.6: Gaze360 (a) model architecture and (b) gaze estimation transform between the subject's eye coordinate system (E) and the camera coordinate system (L). Positive E_z is pointing away from the camera (extracted from [6])

Figure 2.7: Ground truth gaze (yellow) and Gaze360 estimations (red) (extracted from [6])

Both the OpenFace 2.0 and the Gaze360 model output 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 (Figure 2.6b) [5, 6].

Another image-based estimator is the RT-GENE (Real-Time Eye Gaze Estimation in Natural Environments) model [7]. It is a method for 3D robust gaze and head estimation in unconstrained images. The eye gaze estimation only works when the eyes of the subject are visible, therefore, it is not a full range estimation as the Gaze360 model. To estimate the head pose, several networks are used, as shown in Figure 2.8. First, the facial landmarks, in the input RGB image, are detected using Multi-Task Cascaded Convolutional Networks (MTCNN) [7]. Secondly, the head pose is obtained by adopting the state-of-the-art method presented in [41]. This approach relies on CNNs and adaptive gradient methods to estimate the head pose in unconstrained inputs [41]. However, this method only performs well for frontal views. Thus, other algorithms were developed, such as the WHENet.

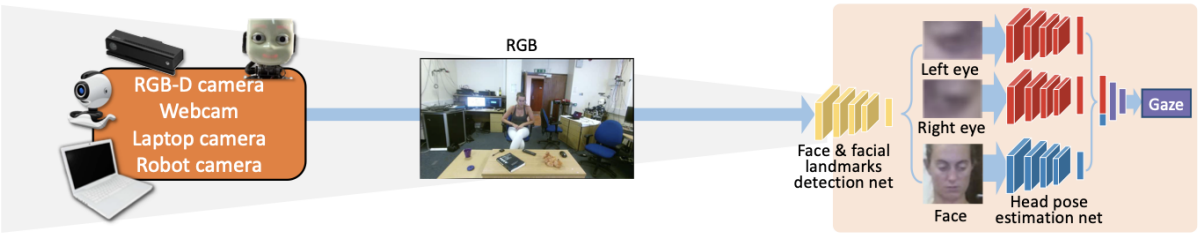


Figure 2.8: RT-Gene model architecture (extracted from [7])

The WHENet (Wide Headpose Estimation Network) model [8] extends the head pose estimation to the full-range of head-yaws, while keeping the accuracy for the frontal head pose estimations. The network receives as input the cropped head images, which can be obtained by the YOLO_v3 [42] facial detector algorithm, as suggested by the authors. The network is an adaptation of the multi-loss framework presented in [43] to the wide range estimation. This framework relies on multi-loss deep CNNs to estimate the intrinsic Euler angles (yaw, pitch and roll) accurately and robustly. It uses three separate losses, one for each angle. Each loss is composed by a classification and a regression loss. The WHENet estimator adjusts the method to the full range by replacing the multi-loss ResNet50 networks, used in [43], with an EfficientNet characterized by modified loss functions for both classification and regression losses [44]. To stabilize the network for large-yaws, a new wrapped-loss is introduced in the model. A key factor of the WHENet model is its robustness, maintaining its accuracy in conditions such as occluded face features and poor quality images [8].

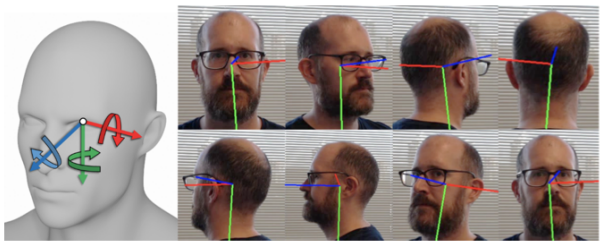


Figure 2.9: Head pose estimated coordinates, parameterized by pitch (red-axis), yaw (green-axis) and roll (blue-axis) (extracted from [8])

Similar to the gaze estimators, both the RT-GENE and the WHENet estimators, output the head pose in spherical coordinates and relative to the camera view coordinate system.

2.2.2 Attention Indices

The most common indices that have been explored for the attention evaluation of children with ASD are based on fixations and/or saccades. Fixations are represented by series of Points of Regard very close in time and space, this is, a period of time in which the eyes are locked towards an object. Saccades are the rapid eye movements between fixations [15, 38].

To analyse the fixations towards each region separately, Areas of Interest are defined around each social and non-social visual stimuli, as shown in Figure 2.10. Various methods have been suggested to define the AOIs. The most used approach is constructing shapes around the targets of interest. These shapes vary between studies and can be geometrical or have a free form according to the target [45].

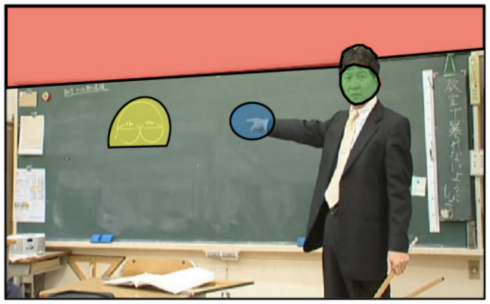


Figure 2.10: Representative visual stimulus (AOIs). The green AOI represents the teacher's face, the blue the teacher's fingers, the yellow the figures at which the teacher is pointing, and the red the wall with no objects (extracted from [9])

To measure fixations, some of the most used indices are the Total Fixation Duration (TFD), the Average Fixation Duration (AFD) and the Sum of Fixation Count (SFC). The TFD, also referred to as dwell time, corresponds to the total time of fixations towards a specific AOI. The AFD is the average duration of each fixation towards a specific AOI. The SFC is the total number of fixations towards a specific AOI [15, 38].

2.2.3 Attention Analysis

Recently, multiple models for a quantitative attention analysis of the subjects with ASD have been proposed. As referred before, several features can be used, such as the eye gaze, the head pose and the facial landmarks. These features are chosen based on the environment and conditions. When the attention analysis study happens in a screen-based environment, the attention is quantified based on the gaze, which is detected using a screen-based eye-tracker, such as Tobii. When the attention is analysed in physical human-robot interactions, multiple features are used based on the study conditions. To analyse the attention through the gaze, eye-trackers without integrated screens, such as Gaze360, are used.

Screen-based Environment

Regarding quantitative attention analysis of ASD subjects, most of the studies found during the literature review were performed in constrained environments, using screen-based eye-trackers to present the visual stimulus, as shown in Table 2.1 [9, 15, 46, 47, 48]. Among these 5 studies, Tobii was the most common screen-based eye-tracker and the presented visual stimulus were usually unrelated with social robots and ASD therapy, going from virtual classes to social movies, as in [9, 15, 46, 47]. However, in [48], two experimental conditions were designed to study the effect of the robots in joint attention. In the first experiment, the human condition, a video of a man trying to induce joint attention towards the objects

was displayed. In the other experiment, the robot condition, a video of NAO was shown on the screen, following the same procedure performed by the man in the human condition. Regardless the presented visual stimulus, for all of these studies, the attention was quantified by relating the short distance 2D eye gaze estimation (PoR), outputted by the eye tracker, with the AOIs defined around the targets presented on the screen. Most of the studies defined the AOIs manually, with the shapes depending on the study. One study defined them as rectangular boxes around the targets [15]. Others used free forms according to the targets' shapes [9, 48], with at least one study adding a margin between the AOI and the target shape [46]. Finally, in [47], the AOIs were defined with different geometrical shapes, depending on the target form.

In 4 out of the 5 presented studies, the participants were children. For all the studies, the procedures were done with a control group of Typically Developed (TD) subjects, beyond the group of subjects with ASD, being the number of participants higher or equal to 20 for each group in each study. Therefore, conclusions about the different behaviors between TD and ASD subjects were taken. For all the studies, the ASD group showed lower attention towards social targets than the TD group. Moreover, in [15], the ASD group showed higher attention towards nonsocial targets than the TD group and lower performance scores in the attention task, which highlights their on-task attention deficits. In both [15, 46], the results indicated shorter fixations for the ASD group, when comparing with the TD. Regarding [48], the ASD group demonstrated less JA than the TD group. However, both groups showed more interests in the robots' faces than in the human faces.

Table 2.1: Literature review of quantitative attention analysis of ASD subjects in screen-based environments. TFD: Total Fixation Duration; JA: Joint Attention; AFD: Average Fixation Duration; SFC: Sum of Fixation Count

Paper	**Visual stimulus**	**Participants**	**Eye Tracking Device**	**AOI shape**	**Attention Indices**
[9]	Virtual class	53 children: ASD = 26 TD = 27	iView X RED	Target	TFD JA
[15]	Simulated virtual class	46 children: ASD = 20 TD = 26	Tobii X2-60	Rectangular	TFD AFD SFC
[46]	CB paradigm Social movies	44 young adults: ASD = 22 TD = 22	Tobii X-60	Target + Margin	TFD
[47]	Movies	390 children: ASD = 83 TD = 307	Gazefinder	Geometrical	TFD
[48]	**IJA with: Human and Robot (NAO)**	**30 children: ASD = 30 TD = 30**	**Tobii X3-120**	**Target**	**TFD JA**

Physical (Human-Robot Interaction) Environment

Although in a sparse amount, there are some studies focused on assessing the autistic subjects' attention in 3D spaces, during human-robot interaction (Table 2.2). These studies use non-intrusive cameras, able to record the procedures. The cameras can have a low-resolution, such as the NAO robot camera and a mobile phone camera, or high-resolution, being able to extract RGB-D images, such as the Kinect. The studies are divided in four groups, according to the features used for the attention measurement: eye gaze, head pose, facial landmarks and gaze and head pose simultaneously. For all the studies, at least one camera and one robot were used, with NAO being the most common choice.

Table 2.2: Literature review of quantitative attention methods for ASD subjects in physical environments. TFD: Total Fixation Duration; JA: Joint Attention; AFD: Average Fixation Duration; SFC: Sum of Fixation Count

Paper	Attention Features	Setup	Participants	Feature Extraction	AOI shape	Attention Indices	Accuracy [%]
[34]	Eye Gaze	Conversation with Human and Robot: Tobii X2-60	10 teenagers: ASD = 4 TD = 6	Tobii X2-60	1.5 $\times$ ellipses	TFD	——
[49]	Eye Gaze + Head Pose	Therapy: 3 Robots 3 Cameras	74 children: ASD = 52 DD = 18 NYD = 3 TD = 1	OpenFace	Range of azimuth angles for each robot	TFD	——
[35]	Head Pose	NAO Kinect	58 children: ASD = 42 TD = 16	Machine Learning	K-means	TFD JA	——
[50]	Head Pose	NAO 4 Cameras	34 children: ASD = 32	Machine Learning	K-means	TFD AFD SFC	73.5
[36]	Facial Landmarks	NAO Mobile Phone	11 children: ASD = 11	Haar Classifiers	——	JA	——

In [34], the attention was quantified through the eye gaze estimation, based on a geometrical approach to define the Areas of Interest. It is a preliminary study, done with the goal of investigating the attention behaviors of ASD individuals in human-human and human-robot communication. It had the participation of 10 adolescents with ages between 15 and 18 years old, including 4 with ASD and 6 TD. Each teen performed one trial, which consisted of two consecutive scripted conversations: with a female human and a female-type android robot, called Actroid. The same appearance of a real individual was given to the android robot, which had 11 facial degrees of freedom. To obtain the gaze estimation, a small Tobii eye-tracker device without an integrated screen, with a similar size to the Kinect, was used. In the beginning of the trials, the eye-tracker was calibrated to output the fixation points on a virtual screen located in the position of the interlocutors faces (human and android robot). The AOIs were defined manually around their faces, with an elliptic form. To overcome the eye-tracker noise, the ellipses were augmented by a factor of 1.5. In this study, the ASD group showed lower attention towards social targets than the TD group and both groups demonstrated more interest in the robots' faces than in the human faces.

In [49], the attention was quantified through the eye gaze and the head pose estimations. This preliminary study investigated the impact of the physical design of the robots in therapy-like settings with ASD children. It had the participation of 74 children with ages between 5 and 8 years old. Among the children, 52 were diagnosed with ASD, 18 had developmental delay (DD), 3 were not diagnosed (NYD) and 1 was TD. The study was conducted in a therapy environment using 3 robots with different appearances and functionalities to interact with the children. At least one operator familiar with the robots and a facilitator able to instruct the child were present in the room. The intervention consisted of 2 parts. In the first part, the children only watched the robots performing different tasks, such as greeting, singing and telling a story, while in the second part, they were encouraged to play with the robots. The setup consisted of the 3 robots, side by side, in front of the child and 3 cameras positioned around the room in different locations. Using the cameras' outputs, the eye gaze and the head pose were estimated by the OpenFace model. Doing the average between the previous two outputs, the attention angle was obtained. Afterwards, the attention angle was compared with the defined AOIs. These AOIs corresponded to 3 ranges of azimuth angles, one for each robot, manually defined.

In [35, 50], the attention was quantified through the head pose, based on a machine learning approach. The first study explored JA in ASD children. It had the participation of 42 ASD and 16 TD children with an average age of 8 years old. The setup consisted of NAO placed on top of a desk in front of the child and a Kinect sensor placed on its feet, on top of the desk, in order to capture the full body and mainly the face of the child. During the sessions, the robot induced JA towards two stimulus, placed on each side of the room: an image of a cat and an image of a dog [35]. The second paper explored RJA in robot-mediated therapy of ASD children. This paper was divided in two parts: the first with 14 ASD participants and the second with 20. The ages of the participants ranged from 1 to 5 years old. The setup consisted of NAO placed on top of a chair in front of the child and 4 web cameras, around the child. During the experiment, NAO tried to induce JA towards two monitors placed on each side of the child [50]. For both [35, 50], since the targets were placed in locations that required head movements when changing the focus of attention, the attention was estimated based on the head movements, with the head pose estimation and classification being done in similar ways between studies. The head pose was estimated using the supervised machine learning method applied in [51]. To classify the attention, the k-Means algorithm was employed, to find k clusters in the 2D head pose estimations. In [35], 3 clusters were found, $k = 3$: one corresponding to looking at NAO, other to the cat image and other to the dog image. In [50], several number of clusters were experimented, with k being set as $\{8, 16, 32, 48, 64, 80, 96, 112, 128, 144, 160\}$. For the best value of k, the model achieved an accuracy of 73.5%. In [35], the ASD group showed less JA than the TD group.

Contrary to the previous presented studies, in [36], the attention was quantified based on the facial landmarks detection. The authors proposed a robot-mediated therapy to improve the attention of ASD children and an assessment system for ASD children. The study had the participation of 11 ASD children, with ages under 16 years, being the average age 9 years old. The setup consisted of the NAO robot with an additional mobile phone mounted on its chest to display emotions. The mobile phone camera was used to detect the facial landmarks of the child, using Haar classifiers, which determine light and dark areas in an image, based on the contrast values between adjacent rectangular groups of pixels [52]. Since there was only one target, and the camera was placed on it, the attention was classified based on the face detection. If the face is not detected, it is assumed that the child is not paying attention to the robot. Otherwise, it is assumed that the child is paying attention to the robot.

All of the presented studies were performed with children or teenagers. In 3 out of the 5, the procedures were done with a control group of TD children, beyond the group of ASD children. However, one of the studies only had one TD participant [49] and other only had a total of 10 children (4 ASD and 6 TD) [34], which represents a low number for the comparison between the groups.

Limitations and Contributions

Regarding the attention indices used in the previous studies, the most common metric is the Total Fixation Duration, computed in 9 out of the 10 studies, followed by the JA, measured in 4 studies. 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, with some of the presented works corresponding only to preliminary studies. The number of participants and sessions in each study is very limited and thus, it is not possible to generalize the results. Regarding the setups, NAO is the most used robot. For some studies, 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 [35], the authors showed the feasibility of using only one Kinect camera 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. Regarding the method chosen for quantifying the attention using the facial landmarks, it is not feasible with multiple targets, unless, there is a camera implemented on each target.

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.

Chapter 3

Methods

In this chapter, it is presented a description of the clinical acquisitions analysed during this thesis, as well as the proposed system to assess the ASD children's gaze during robotic therapy.

To assess the attention during the therapy, an initial benchmarking of gaze and head pose estimators was performed, followed by the development of an attention analysis system (Figure 3.1) that estimates where the people are looking at during the sessions. This system is based on the gaze extraction and the definition of AOIs to obtain the range of angles that correspond to looking at the different targets. To define the AOIs, two approaches are analysed. Having the gaze and the AOIs, a gaze classification is done. After the gaze classification, the attention indices are defined and calculated for both clinical studies.

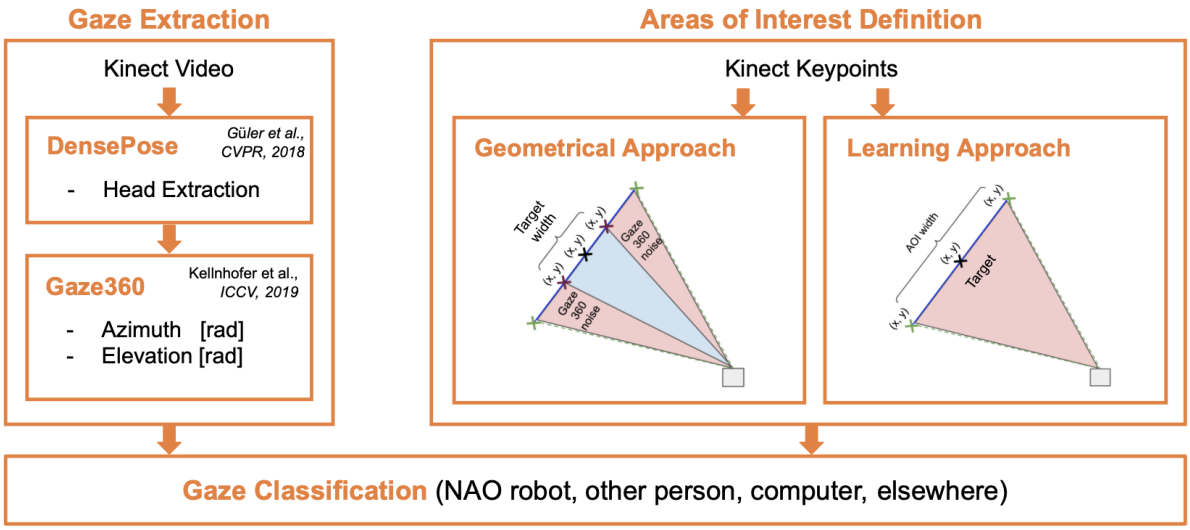


Figure 3.1: Proposed gaze analysis system, consisting of a gaze extraction and the definition of the AOIs, followed by a classification of the gaze according to the different targets AOIs

3.1 Clinical Acquisitions

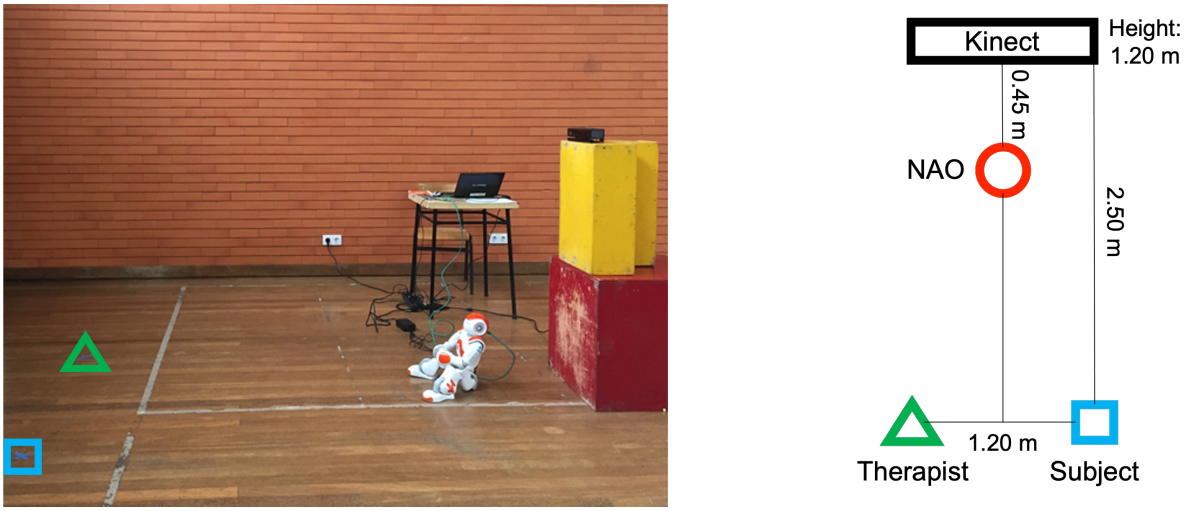
In order to evaluate the ASD children's gaze in robotic therapy, two clinical studies were analysed. The main goal of both studies was to train gestures during triadic interaction sessions between the therapist, the ASD patient and the robot NAO. To instigate this kind of interaction and achieve the goal, a setup and a protocol for an imitation game were defined based on research and clinical knowledge [28]. Both

the setup and the protocol were adjusted between the first and second studies, according to the new study conditions.

3.1.1 Pilot Study

The first clinical study corresponds to a pilot done before the starting of this thesis, between November and December of 2019, in Associação Portuguesa para as Perturbações do Desenvolvimento e Autismo (APPDA Lisboa). APPDA Lisboa is a portuguese organization dedicated to autism. The study had the participation of 3 ASD male adults and consisted of five therapy sessions. However, one of the subjects performed only 4 sessions due to medical reasons.

As shown in Figure 3.2, the setup consisted of a triangle between the three entities, with NAO placed in the middle of the therapist and the subject. In order to perform live mirroring, a Kinect camera was placed behind the robot. The therapist and the ASD subject were standing during the sessions.



(a) Therapy session

(b) Top view scheme (not scaled)

Figure 3.2: Pilot study setup. The therapist is represented by a green triangle, the ASD subject is represented by a blue square and NAO is represented by a red circumference

The Kinect is able to calculate 25 keypoints in 2D (px) and 3D coordinates (m) of the therapist and the subject (skeletons), as shown in Figure 2.2 of Section 2.1. The 3D joints, calculated using the coordinate system in Figure 2.3 (Section 2.1), are then used to replicate the movement of the participants on the robot [28].

During these sessions, there was also a computer operator present, who stayed at the left side of the therapist, outside the Kinect range. This operator has to be considered in the study analysis, since the therapist gave instructions to the operator on what exercises to perform. Therefore, he/she attracted the therapist's and subject's attention.

To distinguish between the therapist and the subject, a red shirt was used by the therapist. For each frame, the skeleton closer to the red shirt was saved as the therapist skeleton, while the other one was saved as the subject skeleton.

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

The protocol for the imitation game was composed by 3 training levels and an initial (IE) and final evaluation (FE) of the ASD subjects, as shown in Figure 3.3

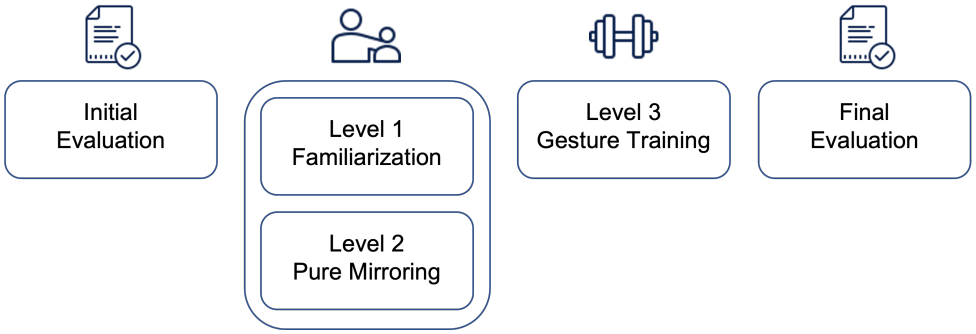


Figure 3.3: Pilot study protocol, consisting of a initial and final evaluation, two levels for the familiarization with NAO and one level of gesture training

The evaluation was done in the first (IE) and last sessions (FE) to assess the subject evolution. In this evaluation, there were 12 gestures, divided into 2 levels of difficulty. The ASD subject was asked to perform the gestures from level 1, first with a verbal instruction from the therapist and in case of failure, with a visual instruction (with the help of a card with the gesture illustration). At the end, it was registered whether the subject was able to perform the gesture after the verbal instruction, after the visual instruction or was not able to perform it. If the subject successfully executed at least 3 gestures from level 1, the evaluation passed to the level 2 gestures, using the same methodology.

After the evaluation, the first level (L1) was the familiarization with the robot, where, firstly, the robot was presented to the ASD adult and its capabilities were explored through 4 different types of stimuli: light stimuli created with its LEDs, sound stimuli through talking, visual stimuli through its movements and finally a simultaneous sound and visual stimuli through speaking and moving at the same time.

The second level (L2) was pure mirroring. In this level, the Kinect camera detected the subject's skeleton and the robot performed a live imitation/mirroring of the subject's upper body movements.

The third level (L3) was the gesture training of the subject. This was the main level of this study thus, the one that the subjects executed during most of the sessions. The goal was for the ASD subject to imitate semantic gestures in a turn-taking exercise. The 18 gestures presented in Figure 3.4 were performed and each one of them was associated with a simple and brief sentence. In this level, there were two different coaching modes, distinguished by the order of who performs the gestures first. In the Robot Coach mode, the gesture was first performed by the robot, then by the therapist and in last by the ASD subject. While in the Therapist Coach, the gesture was first performed by the therapist, then by the robot and lastly by the subject. More details are presented in [28].

The study was composed by 5 therapy sessions of approximately 20 minutes each. The Table 3.1 describes the levels performed on each session by each subject.

Table 3.1: Pilot study sessions. *: Session not recorded; L: Level; IE: Initial Evaluation; F: Final Evaluation

	Subject 8	Subject 10	Subject 21
Session 1	IE*	IE*	IE*
Session 2	L1-L3	L1-L3	L1-L3
Session 3	L2-L3	L1-L2	L2-L3
Session 4	L2-L3	L2-L3	L2-L3
Session 5	FE*	FE*	FE*

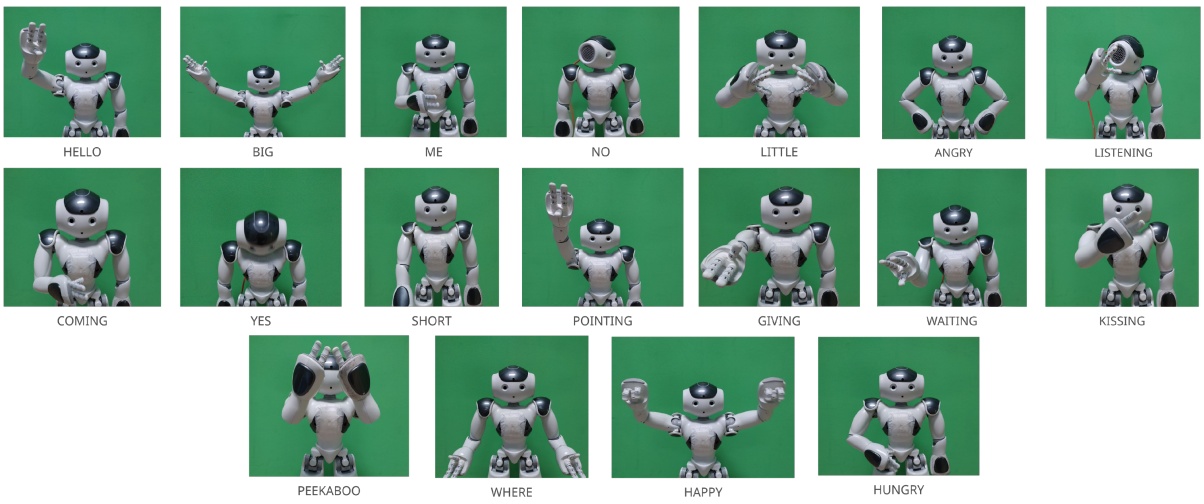


Figure 3.4: Representation of the 18 gestures included in the Protocol

3.1.2 School Study

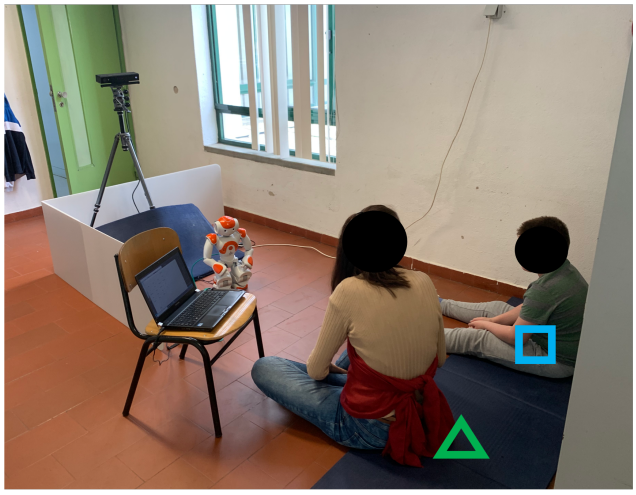
The second clinical study corresponds to a school study done between May and July of 2021 in Escola Básica Bernardim Ribeiro in colaboration with APPDA Lisboa. The participants were six ASD children, 5 males and 1 female, with ages between 7 and 11 years old. Five children were diagnosed with level 3 of ASD, while one child was diagnosed with level 1, according to the Diagnostic and Statistical Manual of Mental Disorders V [53]. 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 the presence of the children in the school during the acquisition days. As mentioned before, the setup was adjusted to the new scene geometry and the protocol was improved.

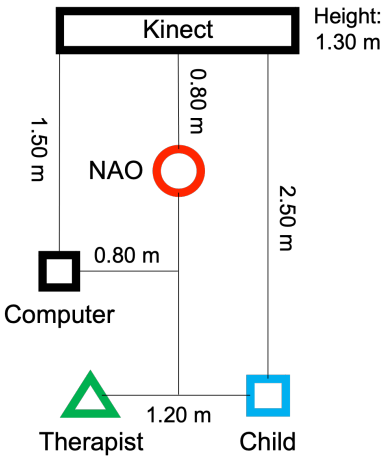
The main geometry of the setup was the same of the previous Pilot Study, with a triangle connecting the three entities. However, since the school space was smaller, the distances between the three agents were adapted. In order to facilitate the work of the therapist and improve the children's attention to the protocol, the computer operator was removed from the setup and the therapist was responsible for the robot control. Therefore, the computer was placed near the therapist as shown in Figure 3.5. In this way, the therapist was able to choose the best exercises for each child, personalising the therapy. The therapist and the ASD child were sat during the sessions.

The data from the sessions was acquired using the Kinect camera. The acquired data was the same of the pilot study, except for the times of the exercises, which were not saved in this study, due to an error during the acquisitions.

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 were 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 more to the right side were kept. To distinguish between the therapist and the child skeleton, the therapist used a red scarf during the sessions. For each frame, the closest skeleton to the red scarf was saved as the therapist skeleton, while the other was saved as the child skeleton.



(a) Therapy session



(b) Top view scheme (not scaled)

Figure 3.5: School study setup. The therapist is represented by a green triangle, the ASD child is represented by a blue square, NAO is represented by a red circumference and the computer is represented by a black square

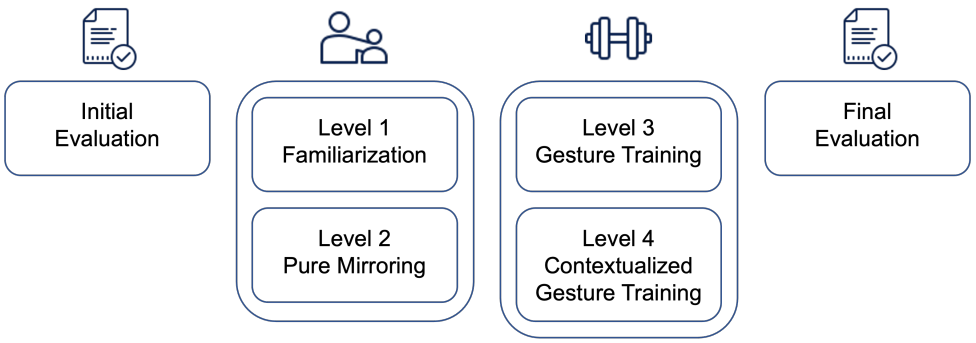


Figure 3.6: School study protocol, consisting of a initial and final evaluation, two levels for the familiarization with NAO and two levels of gesture training

The protocol was updated, with the addition of a fourth training level, as shown in Figure 3.6. It expanded the gesture training towards situations of the daily life. The gesture training was made while the ASD child observed an image of a common place like kitchen, bedroom, school, train or beach. This level was similar to level 3 in multiple aspects, being each gesture associated with a simple and brief sentence. However, in this case, both the gesture and the sentence were also associated with the chosen scenario. In Figure 3.7 it is possible to see the three entities (NAO, Therapist and Child) performing a gesture. The Table 3.2 describes the levels performed by each child in each session. For data anonymization purposes, a random number between 0 and 20 was attributed to each child.

3.2 Benchmarking Gaze and Head Pose Estimators

To estimate where a person is looking at, a model that outputs the people's gaze or head pose is needed. In this section, some gaze and head estimators were studied and compared. Since most estimators present noisy estimations of the elevation angle and the targets were positioned in different horizontal directions during the sessions, it was decided that the azimuth angle was sufficient to distinguish the targets. Thus, only the azimuth was analysed in this thesis.

First, the Gaze360 [6] and the OpenFace [5] models were validated in a short distance experiment,

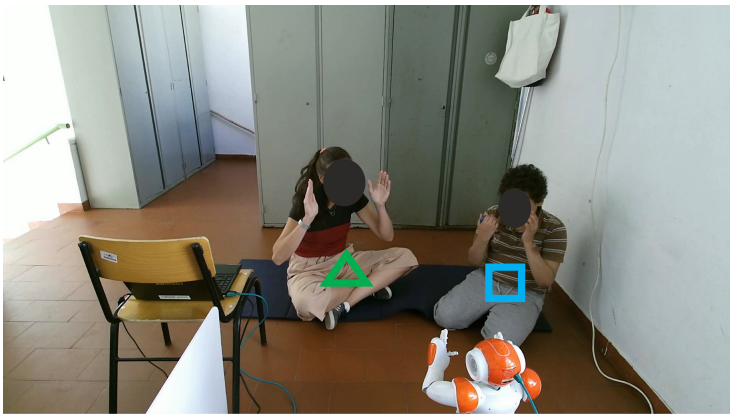


Figure 3.7: Representation of a gesture performed by NAO, the Therapist and the ASD Child, simultaneously

Table 3.2: School study sessions. *: Session not recorded; L: Level; IE: Initial Evaluation; F: Final Evaluation

	Child 6	Child 9	Child 10	Child 15	Child 19	Child 20
Session 1			IE*	IE*	IE*	IE*
Session 2		IE, L1-L3	L1-L2	L1-L2	L1-L2	L1
Session 3	IE*, L1-L2	L3	L3	L3	L3	
Session 4			L4	L4	L4	
Session 5	L3	L4	L3 - L4	L3 - L4		
Session 6		L3-L4	L3 - L4	L3 - L4		
Session 7		FE*	L3 - L4, FE			

using Tobii as standard attention measure system. After, 3 controlled experiments at long distance were done to validate and compare the Gaze360 (gaze) and WHENet [8] (head) estimators. At the end, the most adequate estimator was chosen and inserted in the proposed system.

3.2.1 Short Distance Tests

To validate the Gaze360 and OpenFace models at a short distance, a controlled experiment using Tobii Eye Tracker T120 was done. In addition to the Tobii Eye Tracker, the Kinect camera and a chin rest were also used. The Kinect camera obtained the video to extract the Gaze360 and OpenFace outputs (gaze angles). The chin rest was used to prevent head movements and minimize the noise. Both devices (Kinect and Tobii) were synchronized through a computer.

As shown in Figure 3.8, the Tobii Eye Tracker was placed at a distance of $59cm$ from the eyes and the Kinect was placed on top of Tobii. Both cameras (from Tobii and the Kinect) were aligned with the center of the chin rest. Beyond the Kinect video, the times of each frame and the Tobii outputs were also saved in the computer.

During this experiment, 14 fixation points (Figure 3.9) appeared on the Tobii screen, with the resolution set to $1024 \times 768px$, for 2 seconds each. The subject was instructed to follow and fixate the points, without moving the head. The experiment was performed by 2 subjects with each subject repeating it 3 times.

To compare the 3 gaze estimators, the pixels from Tobii screen where the subject was looking at, at each instant, were calculated, for each estimator. For the Tobii Eye Tracker the pixels were automatically computed, just a translation of the reference frame was performed to have its origin in the center of the screen and facilitate the calculations.

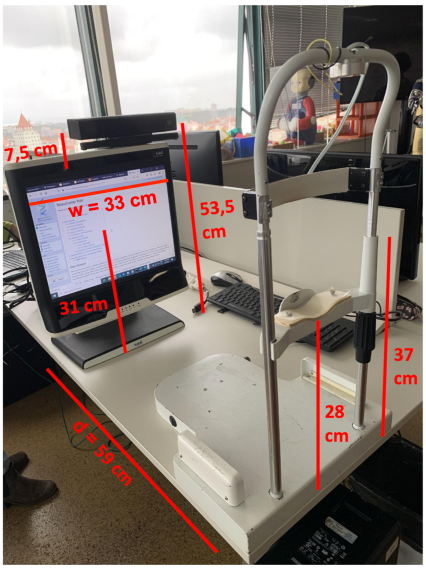


Figure 3.8: Short distance benchmarking setup, consisting of Tobii Eye Tracker T120, with a Kinect camera on top, and a chin rest

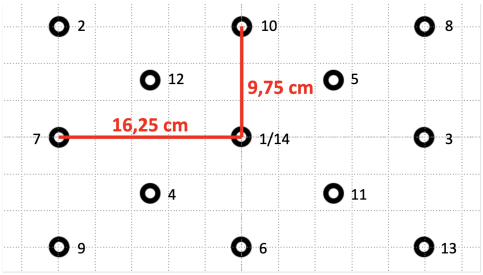


Figure 3.9: Fixation points shown on the Tobii screen during the short distance benchmarking experiments (not scaled)

For OpenFace, the videos were run on the model and the azimuth gaze estimations were obtained. For Gaze360, since the chin rest fixed the head of the subject, instead of using its official face detector, Densepose [40], the bounding boxes for the head were manually defined with resolutions of $275 \times 275px$ and $448 \times 448px$. The two resolutions were used to study the effect of the bounding box size on the Gaze360 estimation. Afterwards, the cut videos were run on the Gaze360 model and the gaze estimations were obtained.

The gaze estimations from the 3 different models were all filtered with a mean filter using a window of 6 frames, in order to reduce the noise. Then, both the OpenFace and Gaze360 estimations in radians (α) were converted into the pixels on the Tobii screen, according to the scene geometry, as given by Equations 3.1 and 3.2. The variables d and w represent the distance to Tobii and the width of the screen, respectively, as shown in Figure 3.8. The value $1024px$ is the horizontal resolution of the Tobii screen.

$$x_{cm} = -d \tan(\alpha) \tag{3.1}$$

$$x_{pixel} = \frac{1024 \times x_{cm}}{w} \tag{3.2}$$

In order to compare quantitatively the gaze estimators, the expected signal was calculated using Equation 3.2 and the x position of the fixation points, obtained through Figure 3.9. After, the signals

were shifted in the temporal axis manually to adjust the different delays coming from the acquisitions and the data processing. Lastly, the Root Mean Squared Error (RMSE) between the estimations and the expected signal was obtained according to Equation 3.3, where n is the number of frames acquired by Tobii or the Kinect camera depending on the estimator being evaluated.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2} \tag{3.3}$$

Given the obtained results, presented in Section 4.1.1 and summarized in Table 3.3, it was decided that the most adequate estimator was Gaze360.

Table 3.3: Average RMSE of the Gaze points estimations relative to the expected signal [px]

	RMSE [px]
Tobii Eye Tracker	190.5
OpenFace	252.4
Gaze360 (275px)	344.0
Gaze360 (448px)	**242.6**

3.2.2 Long Distance Tests

To validate the Gaze360 model at a long distance, 3 controlled experiments were done, comparing Gaze360 (gaze) and WHENet (head) estimators.

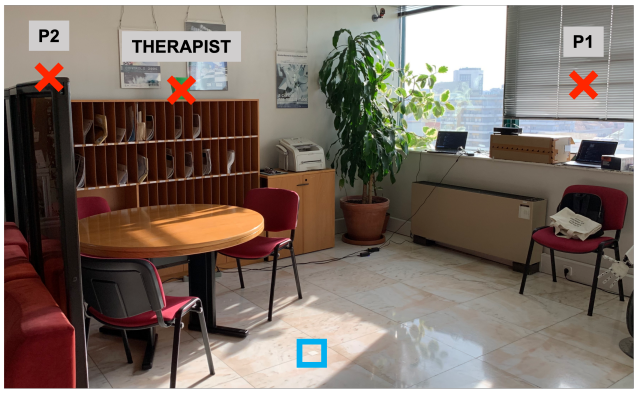
The WHENet estimator was chosen for the long distance validation, since it is a full range ($360°$) estimator, just like Gaze360. In this way, it allowed to validate the model at moments in which the face features were not visible in the video.

The experiments and setups were defined to simulate some of the pilot study conditions. For the 3 experiments, the Kinect camera and the subject were in the same positions as for the pilot study (Figure 3.2b).

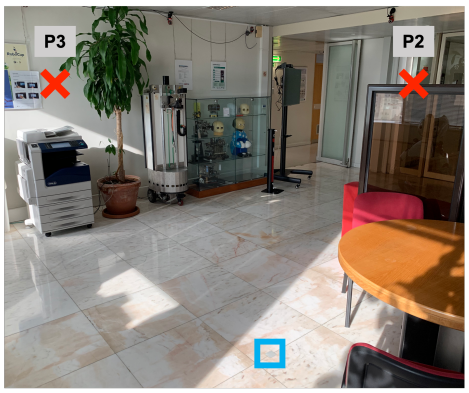
Since WHENet is a head estimator, the first and second experiments were done with head movements accompanying the eye movements. In this way, it is expected that the output of both Gaze360 and WHENet are similar. The last experiment was done only with eye movements, without moving the head, to validate the Gaze360 model and observe the WHENet model estimations.

1) First Experiment

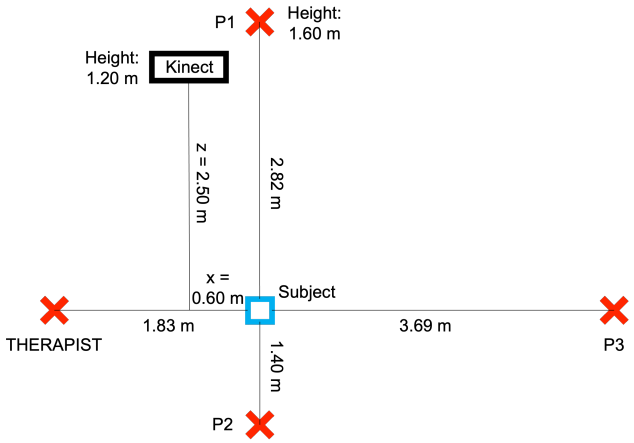
The first experiment consisted of 4 fixation points, placed around the person, as shown in Figure 3.10, with fixation point THERAPIST simulating the therapist during the sessions. To minimize the noise, the fixation points were all put at the eyes' level. The subject was instructed to look at each point for 10 seconds, in the order P1-THERAPIST-P2-P3-P1, having the body, head and eyes turned towards it, as shown in Figure 3.11a. This experiment was performed 4 times.



(a) Fixation points P1, P2 and Therapist



(b) Fixation points P2 and P3



(c) Scheme view from above (not scaled)

Figure 3.10: Long Distance Benchmarking Setup: Experiment 1. The red crosses represent the 4 fixation points and the blue square represents the subject

(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

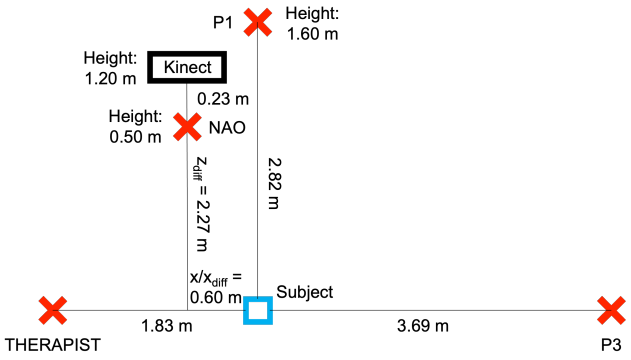
Figure 3.11: Long Distance Benchmarking: Looking to the side turning (a) the body, head and eyes, (b) the head and eyes and (c) only the eyes towards the target

2) Second Experiment

The second experiment consisted also of 4 fixation points, however the fixation point behind the subject (P2) was replaced by a point simulating the NAO robot (NAO), as shown in Figure 3.12. Similar to the previous experiment, the subject was instructed to look at each point for 10 seconds, in the order P1-THERAPIST-P3-NAO-P1. However, instead of having the body, head and eyes turned in the same direction, this experiment was done without body movements. Meaning that the eyes and the head were both turned in the direction of each fixation point, but not the body, as shown in Figure 3.11b. This

experiment was performed 3 times.

(a) Fixation points 1, Therapist and NAO



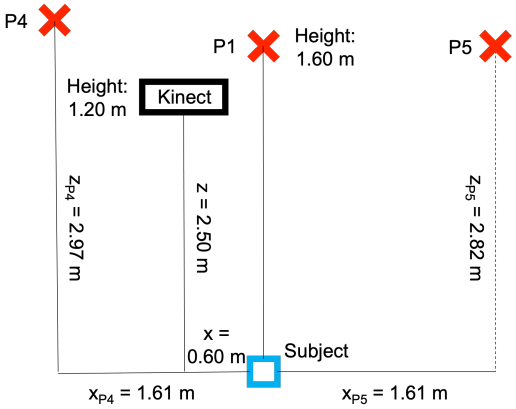
(b) Scheme view from above (not scaled)

Figure 3.12: Long Distance Benchmarking Setup: Experiment 2. The red crosses represent the 4 fixation points and the blue square represents the subject

3) Third Experiment

The third experiment consisted of 3 fixation points, all at the eyes' level, as shown in Figure 3.13. The goal was to validate Gaze360 when there are no head movements. The subject was instructed to look at each point for 12 seconds, in the order P1-P4-P5, moving only the eyes and keeping the body and head turned to fixation point P1, as shown in Figure 3.11c. This experiment was performed 3 times.

(a) Fixation points 1, 4 and 5



(b) Scheme view from above (not scaled)

Figure 3.13: Long Distance Benchmarking Setup: Experiment 3. The red crosses represent the 3 fixation points and the blue square represents the subject

In all the experiments, the Kinect videos, the detected skeletons and the frame times were saved. To compare and validate the estimators, the expected signal and the estimations were obtained for each experiment. Each estimator is implemented with a face detector model to cut the head for each frame. The WHENet model uses YOLOv3 [42], while the Gaze360 employs Densepose [40]. Thus, to calculate the azimuth angle estimations, the videos were run in Densepose + Gaze360 and YOLOv3 + WHENet models.

The expected signals were calculated in degrees, based on the scene geometry, using Equations 3.4, where α_{fp} is the angle towards the fixation point fp, with $fp = \{P1, P2, P3, THERAPIST, NAO, P4, P5\}$. The angles were calculated relative to the camera, using the same reference system of Gaze360. Thus, if the subject was looking to the camera, α_{fp} is $0rad$.

$$
\begin{cases}
\alpha_{P1} = -\arctan\left(\frac{x}{z}\right) \times \frac{180}{\pi} \\
\alpha_{THERAPIST} = \alpha_{P1} + 90 \\
\alpha_{P2} = \alpha_{P1} + 180 \\
\alpha_{P3} = \alpha_{P1} - 90 \\
\alpha_{NAO} = \arctan\left(\frac{x_{diff}}{z_{diff}}\right) \times \frac{180}{\pi} + \alpha_{P1} \\
\alpha_{P4} = \arctan\left(\frac{x_{P4}}{z_{P4}}\right) \times \frac{180}{\pi} + \alpha_{P1} \\
\alpha_{P5} = -\arctan\left(\frac{x_{P5}}{z_{P5}}\right) \times \frac{180}{\pi} + \alpha_{P1}
\end{cases}
\tag{3.4}
$$

Similar to the short distance benchmarking, the signals were shifted manually in the temporal axis to adjust the delays coming from the acquisitions and the data processing. The RMSEs between the estimations and the expected signal were also obtained according to Equation 3.3.

Given the obtained results, presented in Section 4.1.2 and summarized in Table 3.4, it was decided that the most adequate set of face detector and estimator was Densepose and Gaze360.

Table 3.4: Average RMSE of the Azimuth estimations (relative to the expected angles) of the WHENet and Gaze360 models for the 3 long distance experiments [px]

	Experiment 1	**Experiment 2**	**Experiment 3**
Yolov3 + WHENet	**0.64**	**0.42**	0.43
Densepose + Gaze360	1.02	0.46	**0.22**

3.3 Proposed System

In this section, it is presented a system (Figure 3.1) that estimates where the people are looking at during the sessions. This system is based on the definition of Areas of Interest to obtain the range of angles that correspond to looking at the different targets.

The system was first developed for the pilot study and then implemented in the school study, with some adjustments. As shown in the scheme presented in Figure 3.1, it is composed by a gaze extraction and the definition of AOIs.

For the gaze extraction, the Gaze360 estimator, validated in the previous section 3.2, was used. As explained before, only the azimuth was analysed to estimate the location where people were looking at. 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-indices obtained for both clinical studies.

3.3.1 Data Preprocessing and Curation

The data preprocessing was different between studies, due to the different study conditions. For the pilot study, the distinction between skeletons was not working properly in the beginning of all the sessions, while for the school it was confirmed in the beginning of all the sessions that the distinction was correct. For the school study, the bodies crossed and overlapped multiple times, due to the children moving more, which did not happen in the school study. Thus, the data processing used for the pilot study could not be implemented in the school study and had to be adjusted.

However, two data processing steps were equal for both studies. Firstly, a symmetry relative to the plane yz of the reference frame was applied, $(x, y, z) \rightarrow (-x, y, z)$, since the Kinect obtained symmetrical

images. The gaze estimations, calculated by the Gaze360 model, were filtered to reduce the noise, using a moving average filter with a window size of 7 frames.

Pilot Study

For the pilot study, the videos, the respective times and the Kinect keypoints were cut using the exercises times, in order to keep only the most relevant data for the analysis.

During the pilot study, no people appeared inside the Kinect range, and only 2 skeletons were detected for each frame.

The Kinect camera was not able to detect the therapist skeleton correctly in the beginning of each session, thus, a therapist recognition was performed. The first keypoints of both skeletons were discarded until the first therapist head keypoint with a positive x location was detected. After, a manual filter of the keypoints was performed by eliminating the keypoints with a difference higher than $0.5m$ between frames.

To distinguish which Gaze360 head bounding box corresponded to each skeleton, in the first frame, the one more to the left in the image was related with the therapist and the one more to the right is associated with the ASD subject. Then, it was checked for each frame which Gaze360 bounding box was closer to the previous therapist one and the same for the subject.

At the end, the frames without the detection of the 2 Gaze360 bounding boxes and the 2 skeletons were discarded.

School Study

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, due to their movements and the tracking of the Kinect decreased its performance. Furthermore, the participants were sat and in some sessions the space had a poor illumination. Due to this reason, the data was processed, according to Figure 3.14.

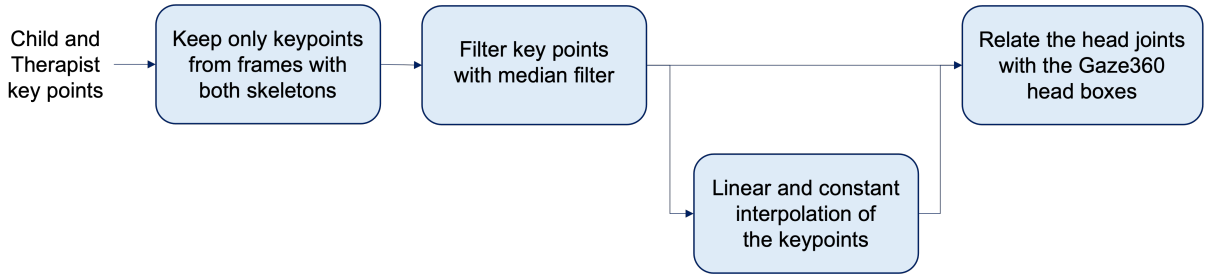


Figure 3.14: Data preprocessing scheme (School study)

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, shown in Table 3.5 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 were missing, 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 (joint 1 in Figure 2.3) were compared with the Densepose head boxes, outputted by the Gaze360 model. Both were represented in

Table 3.5: Percentage of lost data for Session 4 of the school study without using data interpolation. The red cells represent the session in which more than 2/3 of the data is lost [%] (School study)

	Child 6	Child 9	Child 10	Child 15	Child 19
Session 4	—	—	28	81	82

pixels and in the same coordinate system, presented in Figure 3.15.

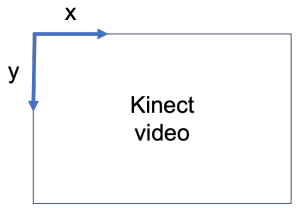


Figure 3.15: Coordinate system of the Densepose bounding boxes and the Kinect 2D joints given the Kinect video

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, according to Equation 3.5. Then, they were increased to compensate the errors from the Kinect skeletons detection, as shown in Equation 3.6, with $p = \{0.25, 0.50, 0.75\}$.

$$\begin{aligned} x_{left} &= box_{left} \times \frac{1920}{960} \\ x_{right} &= box_{right} \times \frac{1920}{960} \\ y_{up} &= box_{up} \times \frac{1080}{720} \\ y_{down} &= box_{down} \times \frac{1080}{720} \end{aligned} \tag{3.5}$$

$$\begin{aligned} x_{left} &= \frac{x_{left} + x_{right}}{2} - \frac{(x_{right} - x_{left})(1+p)}{2} \\ x_{right} &= \frac{x_{left} + x_{right}}{2} + \frac{(x_{right} - x_{left})(1+p)}{2} \\ y_{left} &= \frac{y_{left} + y_{right}}{2} - \frac{(y_{right} - y_{left})(1+p)}{2} \\ y_{right} &= \frac{y_{left} + y_{right}}{2} + \frac{(y_{right} - y_{left})(1+p)}{2} \end{aligned} \tag{3.6}$$

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.

3.3.2 Scene Geometry Analysis

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

angles.

For the pilot study, the targets were the NAO robot and the other person, which was the therapist for the ASD subject and the ASD subject for the therapist. On the other hand, for the school study, the targets were NAO, the other person and the computer. In the pilot study, the operator was not considered a target, since the ASD subjects did not interact with him/her. However, in the school study, the computer was considered a target. 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 subject/child). Since the angles varied according to the people's positions, they had to be calculated for each frame. The therapist and subject 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 3.7, where x and z were the 2D positions of the therapist or the patient, in the Kinect coordinate system (Figure 2.3). x_{diff} and z_{diff} were the difference between the 2D coordinates of the person and the target, given by Equations 3.8. An example is presented in Figure 3.16.

$$\alpha_{NAO} = \arctan\left(\frac{x}{z}\right) - \arctan\left(\frac{x_{diff}}{z_{diff}}\right) \tag{3.7}$$

$$\begin{aligned} x_{diff} &= x - x_{target} \\ z_{diff} &= z - z_{target} \end{aligned} \tag{3.8}$$

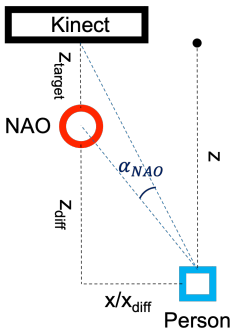


Figure 3.16: NAO standard angle (α_{NAO}) representation. The red circumference represents NAO, while the blue square represents the person from which the standard angle is calculated. The referential is located in the center of the Kinect as shown in Figure 2.3 from Section 2.1.

To obtain the standard angles for looking to the other person and to 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 on 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 3.9. Both conditions can be

combined and reduced to $x_{diff} \times x > 0$. In Figure 3.17 it is possible to observe these situations.

$$\alpha_{target} = -\arctan\left(\frac{z}{x}\right) + \arctan\left(\frac{z_{diff}}{x_{diff}}\right) \tag{3.9}$$

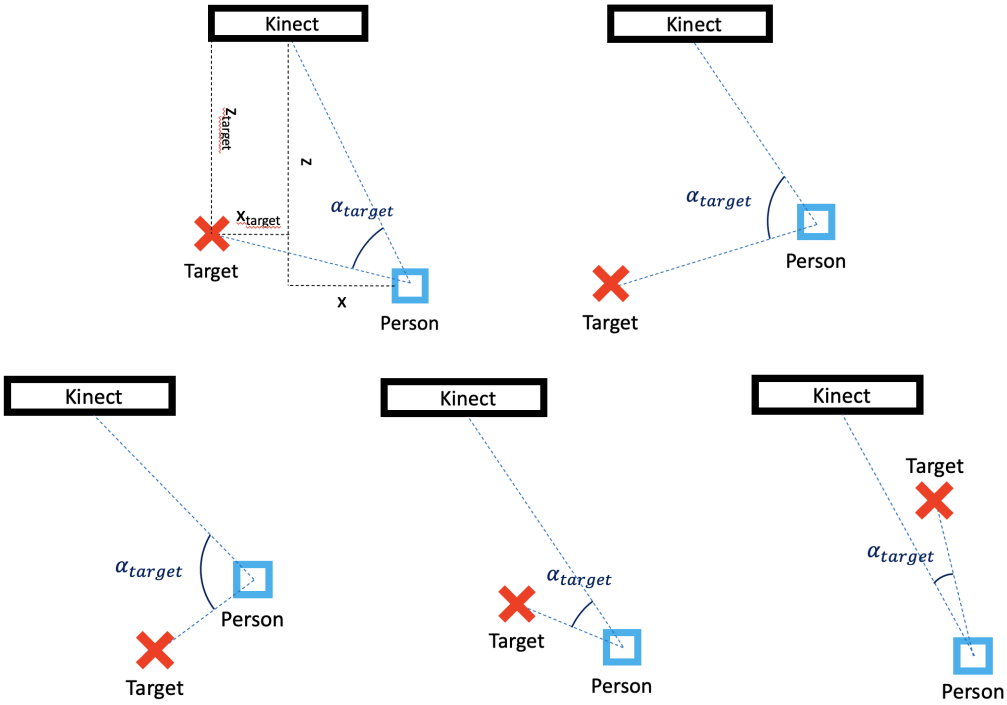


Figure 3.17: Standard angles (α_{target}) representation when $x_{diff} \times x > 0$. The red cross represents the target, while the blue square represents the person from which the standard angle is calculated. The referential is located in the center of the Kinect.

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

If the person and the target were on 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 3.10 was used. This situation is shown in Figure 3.18.

$$\alpha_{target} = \arctan\left(\frac{x}{z}\right) - \arctan\left(\frac{x_{diff}}{z_{diff}}\right) \tag{3.10}$$

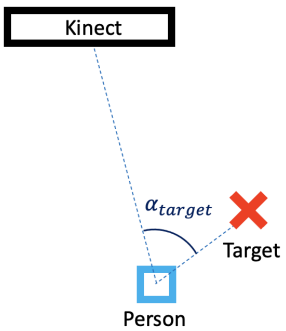
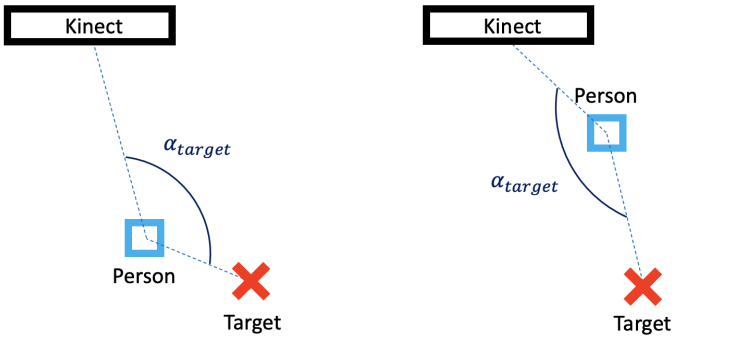


Figure 3.18: Standard angle (α_{target}) representation when $x_{diff} \times x \leq 0$ and $z_{diff} > 0$. The red cross represents the target, while the blue square represents the person from which the standard angle is calculated. The referential is located in the center of the Kinect.

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

If the person and the target were on 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 3.11 or 3.12, depending on relative positions between the person and the target. Equation 3.11 was used if $\arctan\left(\frac{z}{x}\right) \geq \arctan\left(\frac{z_{target}}{x_{target}}\right)$, while Equation 3.12 was used if $\arctan\left(\frac{z}{x}\right) < \arctan\left(\frac{z_{target}}{x_{target}}\right)$. Both situations can be observed in Figure 3.19.

$$\alpha_{target} = \begin{cases} \arctan\left(\frac{x}{z}\right) + \arctan\left(\frac{z_{diff}}{x_{diff}}\right) + \frac{x}{|x|}\frac{\pi}{2}, & \arctan\left(\frac{z}{x}\right) \geq \arctan\left(\frac{z_{target}}{x_{target}}\right) \quad (3.11) \\ -\arctan\left(\frac{z}{x}\right) - \arctan\left(\frac{x_{diff}}{z_{diff}}\right) - \frac{x}{|x|}\frac{\pi}{2}, & \arctan\left(\frac{z}{x}\right) < \arctan\left(\frac{z_{target}}{x_{target}}\right) \quad (3.12) \end{cases}$$



(a) $\arctan\left(\frac{z}{x}\right) \geq \arctan\left(\frac{z_{target}}{x_{target}}\right)$ (b) $\arctan\left(\frac{z}{x}\right) < \arctan\left(\frac{z_{target}}{x_{target}}\right)$

Figure 3.19: Standard angles (α_{target}) representation when $x_{diff} \times x \leq 0$ and $z_{diff} < 0$. The red cross represents the target, while the blue square represents the person from which the standard angle is calculated. The referential is located in the center of the Kinect.

3.3.3 Analysis of Gaze(360) Angles Distribution

After filtering the Gaze360 output angles, the angles distribution was analysed. As previously mentioned, the angles estimated by the Gaze360 model are in relation to the Kinect (Section 2.2.1), which means that, if a person is looking at the camera, the angle will be $0rad$.

Analysing the heat-scatters from the pilot study, shown in Figure 3.20, it is possible to observe that the biggest subjects' clouds have a deviation from the expected angle of looking at NAO, which was not confirmed by the OpenFace estimations. To verify if these offsets depended on the head pose of the subjects, the gaze estimations were compared with the WHENet and RT-Gene [7] head pose estimations. It was concluded that the Gaze360 offsets were not due to the head pose (Section 4.3.3).

Therefore, for each study (pilot and school), the effect of the offset correction was evaluated. The scheme in Figure 3.21 summarizes the procedure used to calculate the offsets for looking at each target.

First, the Gaze360 estimations were centralized to each target, according to Equation 3.13, using the α_{target}, from the previous section.

$$\alpha_{centralized} = \alpha_{Gaze360} - \alpha_{target} \quad (3.13)$$

Then, the histograms of the angles distribution were obtained for each centralization, as shown in

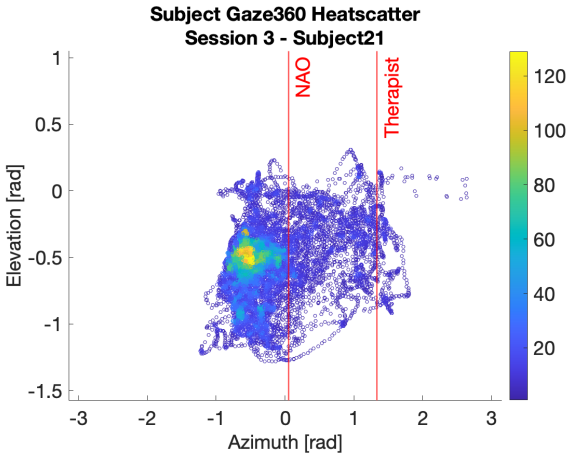


Figure 3.20: Angles distribution of the Gaze360 estimations, expressed in terms of elevation and azimuth angles, for Session 3 with Subject 21. The red lines represent the expected positions of the targets (NAO and Therapist) in the azimuth according to the scene geometry

Figure 3.21: Offset correction scheme

Figure 3.22, and for each person in each session.

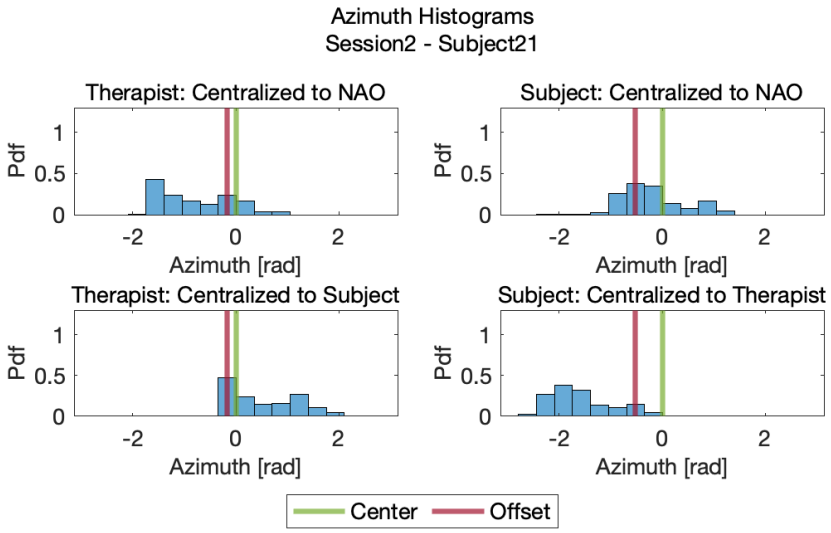


Figure 3.22: Centralized Histograms of the Gaze360 estimations for the Therapist (left) and the Subject (right) in Session 2 of Subject 21. The green lines represent the center of the histograms ($0rad$) and the red lines represent the computed offsets. Pdf: Probability distribution function

According to the histogram bar widths, the number of maximums and their positions vary. In this way, several bar widths were tested for each group of people, Therapist group and ASD group (subjects/children), using one of the sessions. For the pilot study, it is expected the existence of 2 maximums in each histogram. Thus, the chosen width was the one where most histograms had 2 maximums, considering all the subjects and therapist centralized histograms. For the school study, 3 maximums are expected in each histogram. A similar approach of the pilot study was used, but, 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).

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 ($0rad$) was used as the offset of that target. The intervals to find a maximum close to the center were established based on the scene geometry of each study. For both people in the pilot study and for the therapist in the school study, the defined interval was $[-\frac{\pi}{4}; \frac{\pi}{4}]rad$. For the children in the school study, the interval was $[-\frac{\pi}{6}; \frac{\pi}{6}]rad$, since the targets were closer in terms of radians. If no maximum was found in the defined intervals or if the bar of the obtained maximum had a lower value than the bar in the $0rad$, it was assumed that there was no offset for looking at that target.

The obtained offsets were later added to the calculated Areas of Interest.

3.3.4 Areas of Interest Definition

According to the literature review (Chapter 2), it was decided to use an Area of Interest for the definition of each target. Thus, the AOIs correspond to an area constructed around each target (NAO robot, other person and computer). If a person is fixating this area, it is considered that the person is looking to the target.

The AOIs were defined only 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, as shown in Figure 3.23. In Figure 3.24 it is possible to observe the top view of an AOI, together with some of the variables mentioned in the scheme of Figure 3.23.

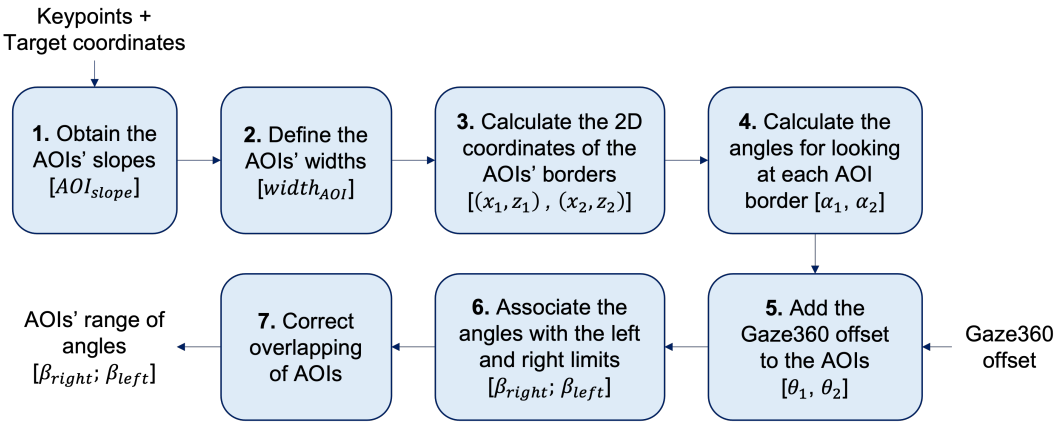


Figure 3.23: AOIs definition scheme

1. First, the slope of the AOI was obtained. 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, as shown in Figure 3.24. The slope was obtained using Equation 3.14, where x_{diff} and z_{diff} represented the difference between the 2D coordinates of the person and the target, as explained in Equation 3.8.

$$AOIslope = -\frac{x_{diff}}{z_{diff}} \tag{3.14}$$

2. Second, the width of the AOI was defined [m]. For this, two approaches were studied. First a geometrical approach was studied, where the real widths of the targets were used. The second approach is based on learning, being the best widths for the AOIs calculated using a session as training set of the system. Both approaches are explained in Sections 3.3.5 and 3.3.6.

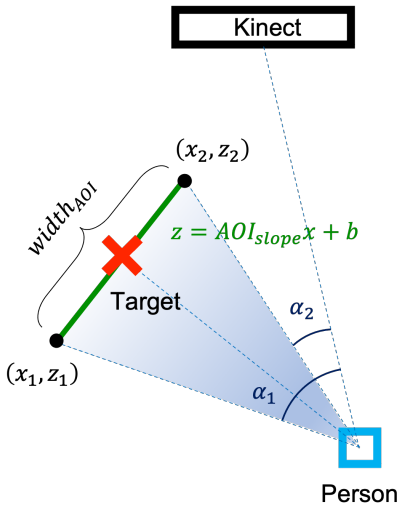


Figure 3.24: Representative top view of an AOI. The red cross represents the target, while the blue square represents the person in analysis. The green line corresponds to the AOI and the area in blue to the range of angles that correspond to the person looking at the target's AOI

3. Third, the 2D coordinates of the borders of the AOI were calculated [m]. Having the slope (AOI_{slope}), the width ($width_{AOI}$) and the center of the AOI ((x_{target}, z_{target})), two limits, one on each side of the target were obtained. In this way, (x_1, z_1) and (x_2, z_2) were derived using Equations 3.15.

$$\begin{cases} x_1 = x_{target} + \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \\ z_1 = z_{target} + AOI_{slope} \times \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \\ x_2 = x_{target} - \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \\ z_2 = z_{target} - AOI_{slope} \times \frac{width_{AOI}}{2} \times \sqrt{\frac{1}{1+AOI_{slope}^2}} \end{cases} \tag{3.15}$$

4. Fourth, the range of angles corresponding to looking at each AOI was obtained [rad]. To compute the relative positions between the person and the two extremities of the AOI ($(x_1, z_1), (x_2, z_2)$), the Equations 3.8 were used, with $(x_{target}, z_{target}) = (x_1, z_1)$ and $(x_{target}, z_{target}) = (x_2, z_2)$. Afterwards, Equation 3.7 was chosen when NAO was the target, while Equations 3.9, 3.10, 3.11 and 3.12 were used with 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. Fifth, the Gaze360 offsets [rad] obtained for each target through the centralized histograms (Section 3.3.3), were added to the calculated angles, α_k, with $k = 1, 2$, according to Equation 3.16.

$$\theta_k = \alpha_k + offset \tag{3.16}$$

6. Sixth, the computed angles were associated with the left and right limits of the AOI, obtaining $[\beta_{right}; \beta_{left}]$. Before this association, the angles θ_k were corrected to belong to the interval $[-\pi; \pi]$, according to Equation 3.17.

$$\theta_k = \begin{cases} \theta_k - 2\pi, \theta_k > \pi \\ \theta_k + 2\pi, \theta_k < -\pi \end{cases} \tag{3.17}$$

Since it is impossible for the AOIs to have a range higher than π, it was checked if the absolute difference between the 2 values of θ_k was lower than π. In these cases, the lowest value of θ_k was attributed to the variable β_{right} and the highest to β_{left}. Otherwise, the lowest value of θ_k was attributed

to the variable β_{left} and the highest to β_{right}. Equation 3.18 summarizes these situations.

$$\begin{cases} [\beta_{right}; \beta_{left}] = [\theta_{min}; \theta_{max}] \text{ , if } |\theta_1 - \theta_2| < \pi \\ [\beta_{right}; \beta_{left}] = [\theta_{max}; \theta_{min}] \text{ , if } |\theta_1 - \theta_2| > \pi \end{cases} \tag{3.18}$$

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

First, it was checked if an AOI is 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 therapist and child were inside the scene, shown in Figure 3.5b. Thus, the therapist or the child were always the ones blocking the view to the other targets. In the school study, if the AOI of NAO was covering or being totally covered by the computer AOI, the computer AOI was deleted, since NAO was the 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, using the Equation 3.19.

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{3.19}$$

For each target, the mean, μ, was defined as the mean value of the AOI limits at that instant (Equation 3.20). The standard deviation, σ, was calculated using an empirical rule. $k\sigma$, with $k = \{1, 2, 3\}$, was defined as half of the AOI (Equation 3.21). In this way, according to the empirical rule, 68%, 95% and 99.7% of the values were within k standard deviations of the mean, respectively.

$$\mu = \frac{\beta_{left} + \beta_{right}}{2} \tag{3.20}$$

$$k\sigma = \frac{width_{AOI}}{2} \tag{3.21}$$

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

3.3.5 Geometrical Approach

In the geometrical approach, referred in the step 2 of the AOIs definition (Section 3.3.4), the widths of the targets were decided based on their geometrical dimensions. The Gaze360 noise was calculated and added to the target limits to obtain the AOIs final widths, as shown in Figure 3.25.

The widths of NAO robot and the people were decided based on the literature. For the therapist and subject the average shoulders' width (bideltoid) of a male adult was considered. According to [54], this value is $47.6cm$. Figure 3.26a shows the bideltoid breadth measurement.

For NAO the shoulders' width was chosen, as well as the arms length, since NAO was using the arms to make the gestures. In this way, the NAO width corresponded to its shoulders' width plus both arms length. According to Figure 3.26b, the NAO width was $27.5 + 31.1 \times 2 = 89.7cm$.

The Gaze360 noise was obtained from the controlled experiments for the long distance benchmarking (Section 3.2.2), following the scheme presented in Figure 3.27. The 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 was obtained for each fixation point. At the

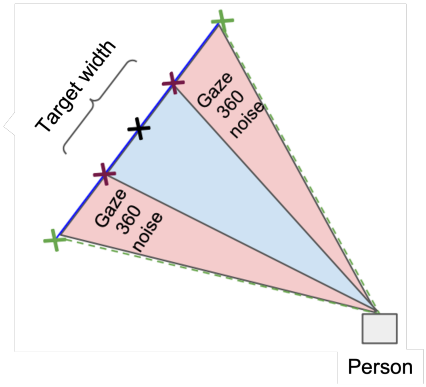
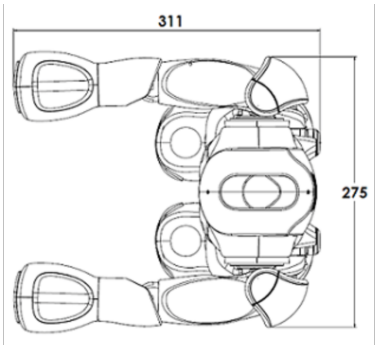


Figure 3.25: Representative AOI for the geometrical approach, including the target width and the Gaze360 noise

(a) Bideltoid breadth measurement (extracted from [55])



(b) NAO dimensions (extracted from NAO Software 1.14.5 documentation)

Figure 3.26: Targets (Other Person and NAO) dimensions

end, the highest standard deviation value was assumed to be the Gaze360 noise.

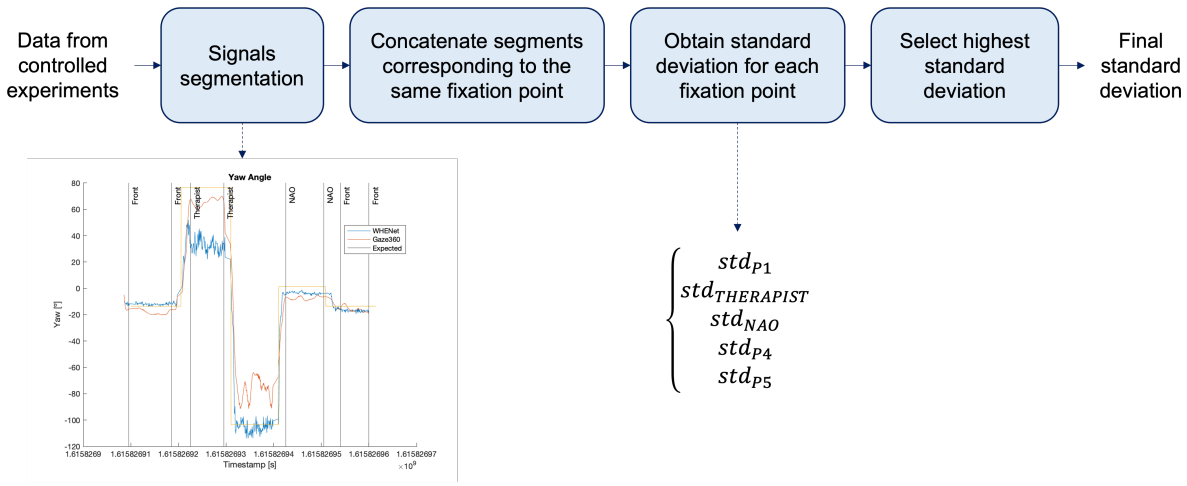


Figure 3.27: Gaze360 noise calculation scheme

After step 6 of the AOIs definition, the Gaze360 noise $[rad]$ was added to the obtained range of angles, according to Equations 3.22. At the end, the angles were corrected again to belong to the

interval $[-\pi;\pi]$, using Equation 3.17, with $\theta_k = \beta_{right}, \beta_{left}$.

$$\begin{aligned} \beta_{right} &= \beta_{right} - noise \\ \beta_{left} &= \beta_{left} + noise \end{aligned} \tag{3.22}$$

3.3.6 Learning Approach

In the learning approach, referred in the step 2 of the AOIs definition (Section 3.3.4), the best widths for each target were obtained through ground truth comparison. The scheme presented in Figure 3.28 summarizes the process.

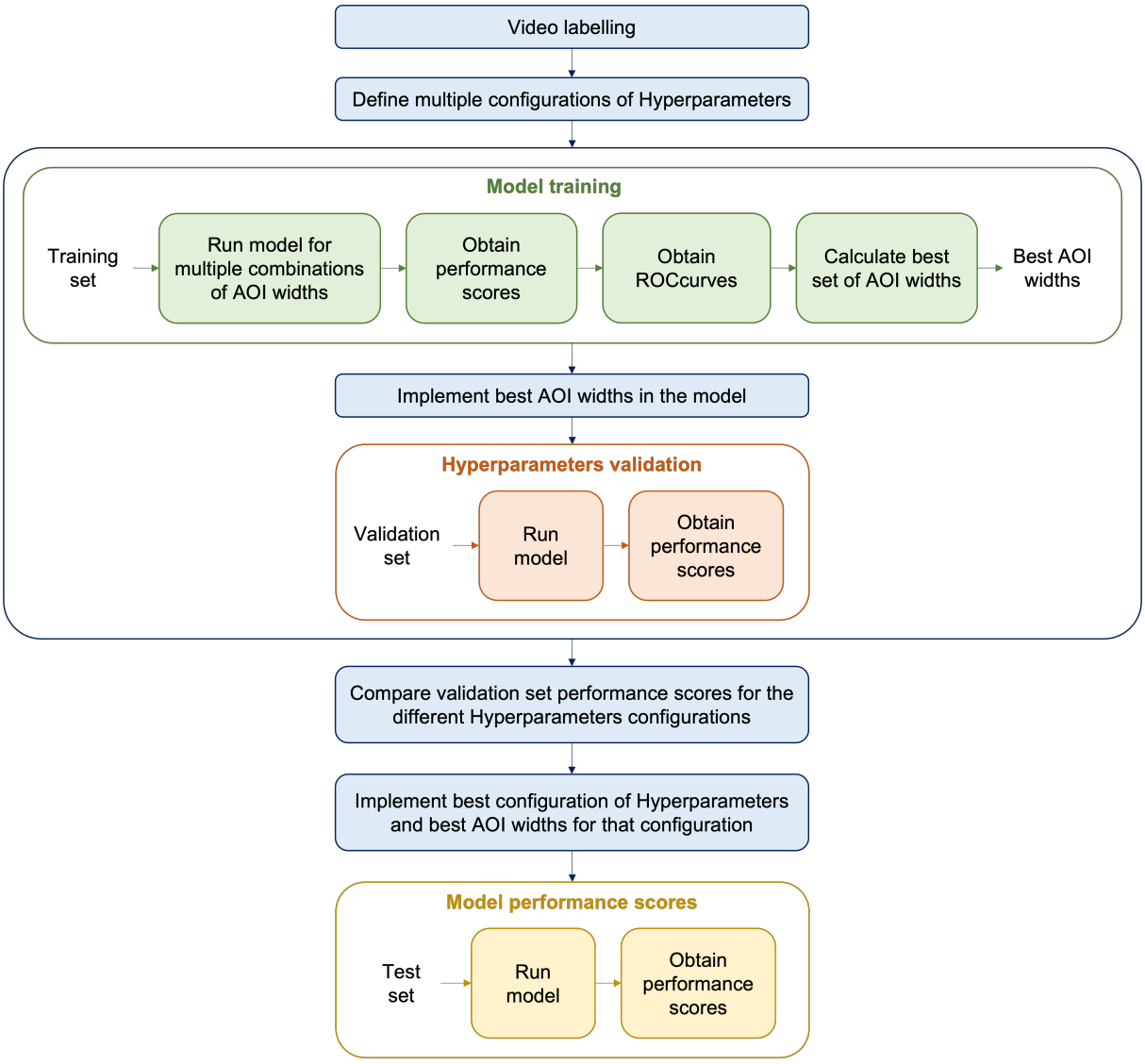


Figure 3.28: Learning approach scheme, composed by the system training (green blocks), validation (orange blocks) and testing (yellow blocks)

The data was divided in training, validation and test sets. From the 7 school study sessions, one was used as training set, other as validation set and the remaining 5 as test set. For the pilot study, a similar division of the data was done. The training set was used to find the best widths for each target. The validation set was used to make decisions relative to the system hyperparameters and validate the

choice of the best width. The test set was used to test the system and obtain the system performance scores.

The tested hyperparameters were the Densepose bounding boxes increase, ($p = \{0.25, 0.50, 0.75\}$ from Equation 3.6), the Gaze360 offsets correction and the variable $k = \{1, 2, 3\}$ from Equation 3.21, used to define the standard deviation of the Gaussians when 2 AOIs overlap.

The videos from the 3 datasets (training, validation and test) were labeled by two annotators, as will be presented in Section 4.2. Having the labelling from the training session, the proposed system was trained in order to calculate the best set of target widths. The training was done for multiple configurations of the hyperparameters, thus, for each configuration, a best set of target widths was obtained. Each set of widths consisted of a width for the NAO AOI, a width for the other person AOI and, in the case of the school study, a width for the Computer AOI.

For the pilot study, considering the scene geometry, the NAO width and the other person width were defined with values between 1.0m and 3.0m. The system was run for all the possible combinations of NAO and other person widths, with increments of 0.2m. In fact, the system was run for the 11*11=121 combinations of widths.

In the school study, there were 3 different targets. The values tested for the NAO width varied between 0.4m and 3.0m, with increments of 0.2m. For the other person, the tested widths 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. In this case, the system was run for the 14*9*4=504 combinations of widths.

To compute the best combination of widths, the method shown in Figure 3.29 was used. In this method, 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 3.23.

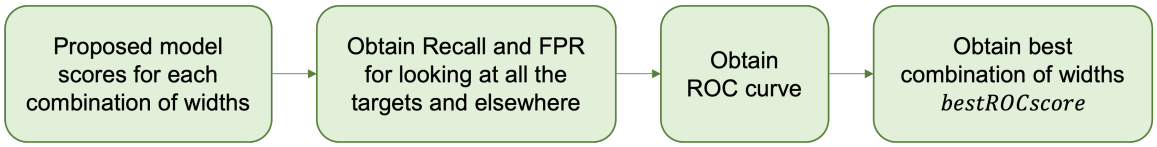


Figure 3.29: Best widths calculation scheme

$$bestROCscore = min(\sqrt{(1 - Recall)^2 + FPR^2}) \tag{3.23}$$

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

3.3.7 Fixation Signal

The attention indices were obtained based on the fixations towards each target (Section 2.2.2). The procedure used to compute the fixations is presented in Figure 3.30.

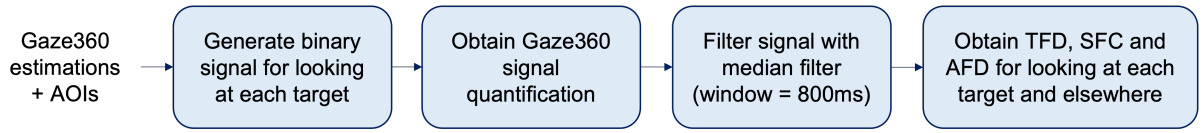


Figure 3.30: Attention indices calculation scheme

First, a quantification of the Gaze360 estimation was done. A binary signal for looking at each target was generated ($binary_{target}$, with $target = \{NAO, otherperson, computer\}$). For each frame, the binary signal was set to 1, if the Gaze360 estimation was inside the AOI range of angles, and 0, otherwise. The final Gaze360 signal ($fixation_{signal}$) was obtained according to Equations 3.24 and 3.25 for the pilot study and the school study, respectively. This step was done through the factorization of binary signals, using a different factor for each target, which facilitated the computation of the attention indices.

$$fixation_{signal} = binary_{NAO} + 2binary_{otherperson} \tag{3.24}$$

$$fixation_{signal} = binary_{NAO} + 2binary_{otherperson} + 3binary_{computer} \tag{3.25}$$

Then, it was considered that a person was fixating a target, if a fixation occurs. According to literature, a fixation was defined as at least $400ms$ of frames within the AOI [56]. In this way, the number of frames corresponding to $400ms$ was calculated. To remove most of the non-fixations, the final gaze signal was filtered with a median filter with a window of $800ms$.

Chapter 4

Experimental Results and Discussion

In this chapter the main results for the gaze estimators benchmarking and the gaze analysis of the two clinical studies are presented, as well as the discussion of the obtained results.

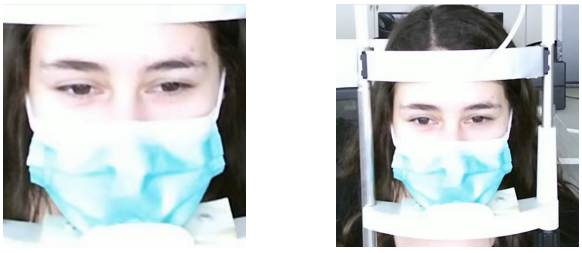
4.1 Benchmarking Gaze and Head Pose Estimators

This section presents the results of the experiments performed for the benchmarking of the gaze and head pose estimators at a ”short” and ”long” distance, described in Section 3.2. At the end, the most adequate estimator was chosen to be implemented in the proposed system.

4.1.1 Short Distance Tests

In the short distance benchmarking, the gaze was calculated using Tobii, the Gaze360 and the OpenFace estimators, in order to validate the last two and choose the most suitable one for the proposed system.

As explained before, the Gaze360 bounding boxes for the head were manually defined, considering two different resolutions ($275 \times 275px$ and $448 \times 448px$), as shown in Figure 4.1.

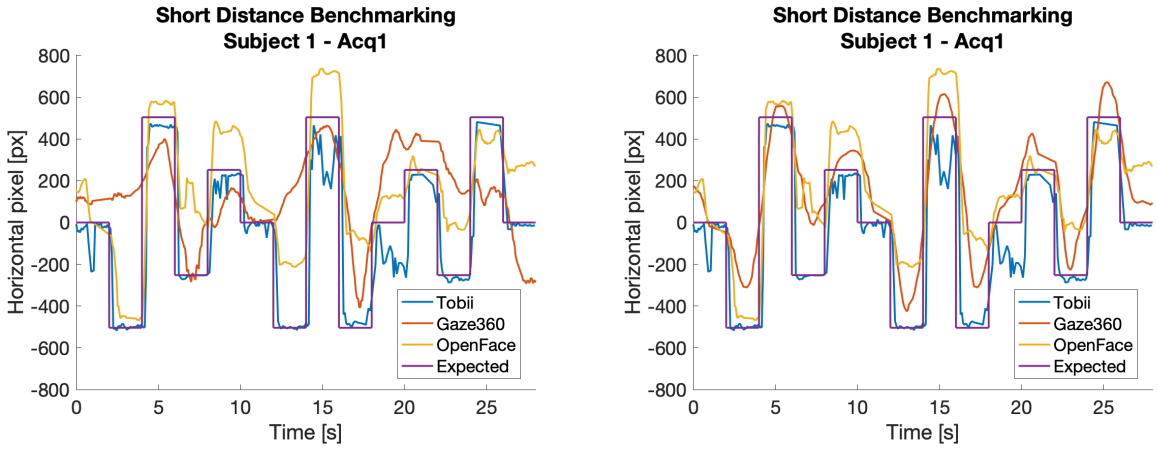


(a) $275 \times 275px$ (b) $448 \times 448px$

Figure 4.1: Cut frames using the different bounding boxes resolutions

Figure 4.2 shows the expected gaze based on geometry (fixation coordinates), the one given by the Tobii eye tracker and the ones estimated by the two algorithms (OpenFace and Gaze360). In Figure 4.2a, the bounding box resolution was $275 \times 275px$, while in Figure 4.2b was $448 \times 448px$. These plots correspond to the horizontal gaze points location on the Tobii screen, during one of the acquisitions.

Observing the results obtained for the Gaze360 model, the estimation is substantially better when the higher-resolution is used. With the lower resolution, the Gaze360 estimation is not able to follow the ”Expected Signal” for most of the fixation points. Conversely, with the higher resolution, the Gaze360



(a) $275 \times 275px$ (b) $448 \times 448px$

Figure 4.2: Gaze points (Points of Regard) estimations using the different head bounding boxes resolutions for the Tobii Eye Tracker (blue) and the Gaze360 (red) and OpenFace (yellow) models. The Expected signal based on geometry is plotted in purple

prediction is very similar to the ”Expected Gaze”, not reaching the expected value just for a few timestamps.

Comparing the 3 estimators, the Tobii predictions have a close behaviour to the ”Expected Signal” for most points, presenting some noise for points 8 and 10 (Figure 3.9). This noise is justified by the infrared lights projector and the camera from Tobii being located at bottom of the screen, while these 2 points are located in the top of the image. Therefore, the iris and pupil are not well detected.

In general, the OpenFace estimation also follows the expected signal, although showing a certain offset during some parts of the experiment, similarly to Gaze360.

To confirm the previous observations, the estimations were compared quantitatively by computing the RMSE between the estimations and the ”Expected Signal”, using Equation 3.3. The results were obtained for all the acquisitions and are presented in Table 4.1.

Table 4.1: RMSE of the Gaze points estimations (relative to the expected signal) for Tobii, the Gaze360 and the OpenFace models in each acquisition of each subject with a sampling frequency of $13Hz$ [px]

	Subject 1				**Subject 2**			
	Acq1	Acq2	Acq3	**Mean**	Acq1	Acq2	Acq3	**Mean**
Tobii Eye Tracker	115	138	142	**132**	105	128	120	**118**
OpenFace	253	198	224	**225**	302	273	265	**280**
Gaze360 (275px)	365	346	468	**393**	318	293	273	**295**
Gaze360 (448px)	241	244	223	**234**	280	223	251	**251**

As expected, Tobii is the estimator with lowest RMSEs, followed by Gaze360 when using a resolution of $448 \times 448px$ and, then, by OpenFace. Both the Gaze360 and the OpenFace RMSEs are partially justified by the delays during the transitions between the different fixations points. Therefore, both models are valid at a short distance, since it is considered that they achieve the expected performance when an adequate head bounding box is chosen.

Considering the lower RMSE of Gaze360 ($448px$) for most of the acquisitions, when compared with OpenFace, and its full-range estimation required for the therapy sessions and not present in the OpenFace, Gaze360 was picked to be implemented in the proposed system. Moreover, the OpenFace usually is not recognized at long distances.

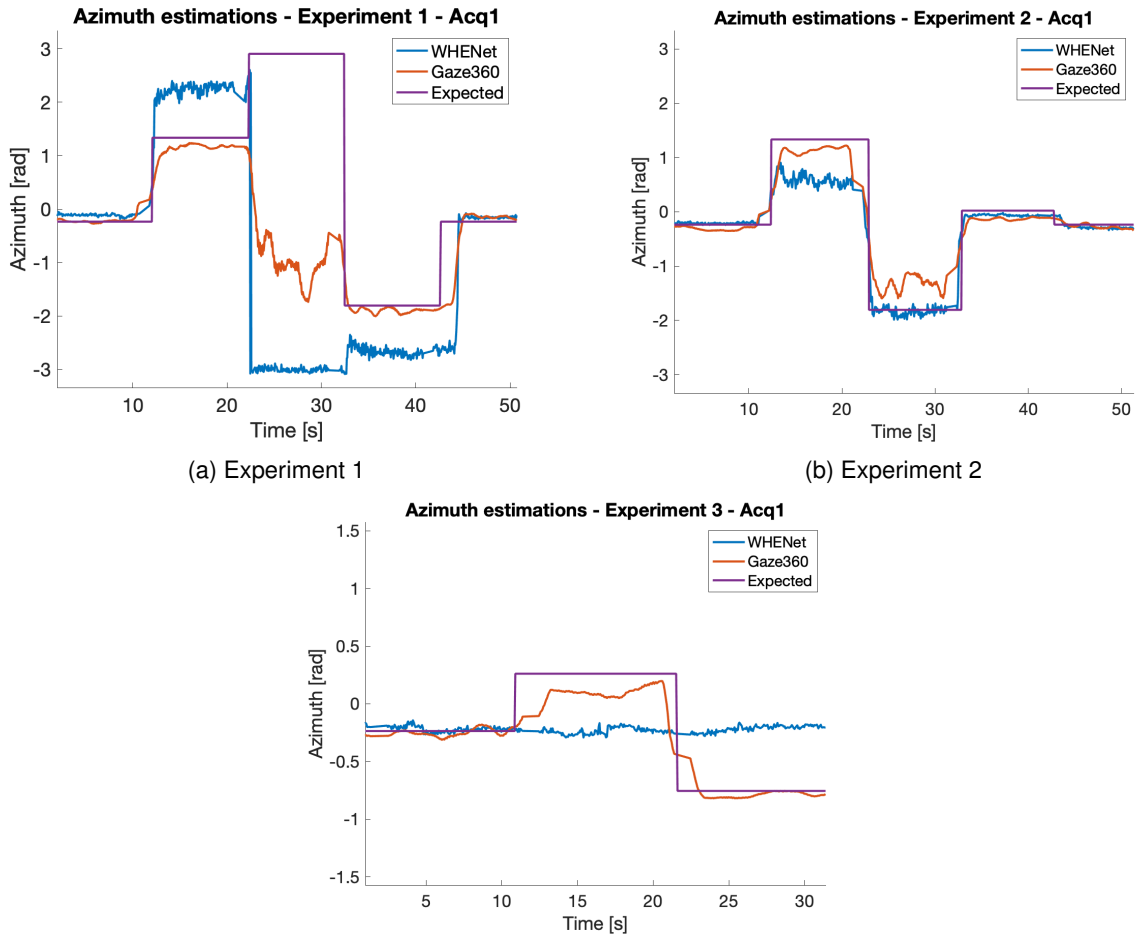
Moreover, the sensitivity of Gaze360 model to the chosen head bounding boxes also became clear from the results. The RMSE values decrease significantly for all the acquisitions when a bounding box

resolution of $448 \times 448px$ was used. Thus, it was decided to use the default face detector implemented with Gaze360, Densepose [40], to maintain the accuracy of the estimator.

4.1.2 Long Distance Tests

The long distance benchmarking consisted of 3 controlled experiments comparing the Gaze360 and the WHENet models, as explained in Section 3.2. The first experiment consisted of 4 fixations points placed around the subject (front (P1), left (THERAPIST), behind (P2), right (P3)). The subject was instructed to look to the fixation points in the order: P1-THERAPIST-P2-P3-P1. The second experiment consisted of 4 points all in front or to the sides of the subject (front (P1), left (THERAPIST), right (P3), in front below the Kinect (NAO)). The subject was instructed to look to the fixation points in the order: P1-THERAPIST-P3-NAO-P1. The third experiment consisted of 3 fixation points placed in the window in front of the subject. In this experiment the subject was instructed to look at each point without moving the head.

For the comparison of the gaze (Gaze360 with Densepose) and head estimators (WHENet), both models predictions were obtained for all the acquisitions done for each Experiment. The results are presented in Figure 4.3, together with the expected signals, obtained based on geometry .



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

Figure 4.3: Azimuth angle estimations for the WHENet (blue) and the Gaze360 (red) models in the long distance benchmarking (a) Experiment 1, (b) Experiment 2 and (c) Experiment 3

Figure 4.3a shows the results obtained for the first experiment, where the estimations of Gaze360 are better than the WHENet ones. For the WHENet, 3 of the 4 fixation points have an estimation error of at least 0.5 rad in relation to the expected signals. Only the angles towards the fixation point in front

($P1$) are closer to their expected value. On the contrary, for the Gaze360 model, just the estimation for the fixation point in the back ($P2$) is more noisy and does not reach the expected value. This noisier estimation is justified by the fact that all facial features are occluded, since the subject is backwards to the camera, deteriorating its detection by the face detector and the consequent estimation by Gaze360. Moreover, comparing both models, the Gaze360 estimation has a delay in relation to the expected signal. This delay can be justified by the time required for the eyes transitions between fixation points, which can take longer to reach since the Gaze360 model uses an input window of 7 frames to include temporal information [6].

For the second experiment the conclusions are fairly similar, although the differences between the two estimations are not so clear visually (Figure 4.3b). Both models follow the expected signal for all the fixation points, with each one having a notable offset for one of the fixations.

In the third experiment (4.3c), the WHENet model estimation is almost a constant signal towards the fixation point in front ($P1$), since there are no head movements during the acquisition. Observing the Gaze360 estimation, it is concluded that it follows its expected signal. The differences between the reaction of the estimators to Experiment 3 confirm the importance of using an eye-gaze estimator.

To confirm the previous observations and obtain a quantitative measurement of the models' performance, the RMSE between the estimations and the expected signals is computed, using Equation 3.3. The obtained results are presented in Table 4.2.

Table 4.2: RMSE of the Azimuth estimations (relative to the expected angles) of the WHENet and Gaze360 models in each acquisition of the 3 experiments with a sampling frequency of $13Hz$ [rad]

	Experiment 1				
	Acq1	Acq2	Acq3	Acq4	**Mean**
WHENet	0.77	0.68	0.71	0.40	**0.64**
Gaze360	1.14	0.94	1.15	0.84	**1.02**

	Experiment 2			
	Acq1	Acq2	Acq3	**Mean**
WHENet	0.46	0.43	0.36	**0.42**
Gaze360	0.37	0.37	0.48	**0.46**

	Experiment 3			
	Acq1	Acq2	Acq3	**Mean**
WHENet	0.43	0.43	0.42	**0.43**
Gaze360	0.29	0.33	0.25	**0.22**

Observing the RMSE values calculated for Experiment 1, both models have high errors. The error from WHENet is explained by the offsets present in most of the fixation points, while the Gaze360 error can be explained by two reasons: (1) the lack of capacity of Gaze360 model to follow the eye movements between fixation points without delays; (2) the high error obtained for the fixation point in the back ($P2$), which increases the RMSE value significantly.

For Experiment 2, the RMSE values are similar for both estimators, which confirms the previous observations. The Gaze360 RMSE values are lower than the RMSE values of the Experiment 1, especially due to the removal of the fixation point in the back ($P2$).

Observing the results obtained for Experiment 3, the WHENet model has a higher error, which was expected, since there were no head movements during the acquisition and the WHENet model does not estimate the eye gaze. The Gaze360 RMSE has a low value, confirming that it is valid at a long distance, when there are no head movements towards the target.

Overall, the Gaze360 model has a high performance when at least the profile facial features are visible. During the therapy sessions, the targets are most of the time in front of the people and the

previous condition is verified. On the other hand, the WHENet model has a better estimation when the facial features are not visible, however, since it is a head estimator, its estimations do not follow the gaze. Thus, the Gaze360 model was validated for this implementation and was inserted in the proposed system.

4.2 Data and Metrics

In this section, it is done a description of the clinical data, which was later analysed using the proposed system, and of the creation of the ground truth data. After, the used evaluation metrics and attention indices, along with how they were calculated, are presented

4.2.1 Clinical Data

Given that some sessions were not recorded, the data to analyse from both studies (pilot and school) had to be selected. For the pilot study, there were 5 sessions, from which the first and last were not recorded. Therefore, only Session 2, 3 and 4 were analysed. For the school study, there were 7 sessions, from which the first one was not recorded for any child and the second one consisted mainly of familiarization levels with the robot. Since the familiarization levels are only performed once and are not the main goal of the Protocol, the attention was only analysed for Sessions 3, 4, 5, 6 and 7. Moreover, since Child 20 was only present in Session 1 and 2, he/she was excluded from the analysis.

To evaluate proposed system performance, the data from the pilot study was split randomly in two sets. Session 3 was used as training/validation set, while Sessions 2 and 4 as test set.

Regarding the school study, the data was split randomly in three sets. Session 3 was used as train set, Session 6 as validation set and the remaining sessions as test set.

4.2.2 Ground Truth Data Labelling and Curation

To evaluate the proposed system, the ground truth was needed. Thus, the videos from the therapy sessions were labelled by two annotators. The labelling was done by selecting where the therapist and the ASD subject were looking at in the selected frame (NAO, Other Person (Therapist or ASD Subject), Computer or Elsewhere). Since more than 25 videos were acquired, with some having more than 15,000 frames, manually extracting the therapist/patient gaze target (NAO, Other Person, Computer or Elsewhere) would be an overwhelming task. Therefore, the videos were labelled every 3 seconds, a period which reflects the main changes in terms of fixations at the different targets.

Afterwards, the inter-annotator agreement was evaluated by computing the Cohen's kappa coefficient [1]. An interpretation of the coefficient is shown in the Table 4.3. At the end, only those frames with labels with total inter-annotator agreement were used for further processing.

Pilot study

Table 4.4 presents the results of the inter-annotator agreement for the pilot study. This table shows the calculated Cohen's kappa coefficient, obtained for the Subject and Therapist labels individually. The percentage of kept labels, after comparing the labels from both annotators, is also presented.

Comparing the results with the interpretation of the Cohen's kappa coefficient present in Table 4.3, it can be concluded that the agreement is moderate or strong for all the subject labels. On the contrary, for the therapist labels, it is always strong or almost perfect. This proves that it is harder to label the

Table 4.3: Interpretation of Cohen's kappa (extracted from [1])

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

subjects gaze. The percentage of useful labels is high for all the sessions, above 70%, which confirms the good inter-annotator agreement.

Table 4.4: Inter-annotator agreement and percentage of kept labels (Pilot Study)

		Cohen's kappa coefficient		**Final labels [%]**
		Subject labels	Therapist labels	Subject + Therapist labels
Session 2	Subject 8	0.87	0.92	89
	Subject 21	0.72	0.93	86
Session 3	Subject 8	0.82	0.89	84
	Subject 21	0.83	0.88	85
Session 4	Subject 8	0.60	0.87	73
	Subject 21	0.73	0.93	82

School study

Regarding the school study, the Cohen's kappa coefficient is higher than 70% for the children and therapist in all sessions (Table 4.5). Comparing these values with the interpretation of the Cohen's kappa coefficient in Table 4.3, it is considered that the agreement between annotators is high. The agreement is moderate for all the children's labels and moderate or strong for all the therapist labels. The children labels have a higher disagreement mainly because they have two targets (Therapist and Computer) in similar angles, making it difficult to distinguish between them.

The percentage of kept labels is shown in Table 4.6, and is higher than 75% for all the sessions, thus, confirming the good inter-annotator agreement.

Table 4.5: Inter-annotator agreement (School study)

	Cohen's kappa coefficient for the Child labels				
	Child 6	Child 9	Child 10	Child 15	Child 19
Session 3	0.77	0.81	0.84	0.81	0.87
Session 4	—	—	0.78	0.74	0.83
Session 5	0.72	0.87	0.83	—	—
Session 6	—	0.81	0.80	0.80	—
Session 7	—	—	0.88	—	—

	Cohen's kappa coefficient for the Therapist labels				
	Child 6	Child 9	Child 10	Child 15	Child 19
Session 3	0.84	0.88	0.91	0.90	0.87
Session 4	—	—	0.91	0.96	0.90
Session 5	0.83	0.92	0.91	—	—
Session 6	—	0.86	0.93	0.90	—
Session 7	—	—	0.92	—	—

Table 4.6: Percentage of kept labels for the Therapist and the Child (School study)

	Final labels [%]				
	Child 6	Child 9	Child 10	Child 15	Child 19
Session 3	78	80	85	81	85
Session 4	—	—	80	79	82
Session 5	77	87	82	—	—
Session 6	—	78	82	80	—
Session 7	—	—	87	—	—

4.2.3 Evaluation Metrics

To evaluate the proposed system, the confusion matrix for each session had to be analysed, in order to compute the performance metrics. Therefore, a 2×2 confusion matrix was calculated for each person (Therapist and Subject) looking at each target (NAO, Other Person and, in the school study, Computer) and Elsewhere. This was done by comparing the ground truth classification with the classification estimated by the proposed system (NAO, Other Person, Elsewhere and, in the school study, Computer). For each session, in the pilot study, $6 \times n_{Subjects}$ confusion matrices were computed (3 for each person) and, in the school study, $8 \times n_{Subjects}$ (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, as shown in Figure 4.4. The final confusion matrices consisted of 4 parameters used to compute the evaluation metrics: the count of true positives (TP), the count of true negatives (TN), the count of false positives (FP) and the count of false negatives (FN).

Having a confusion matrix for each session, the accuracy, precision, recall/true positive rate (TPR) and false positive rate (FPR) were obtained using Equations 4.1, 4.2, 4.3 and 4.4, respectively. The accuracy measures the percentage of correct estimations of the system. The precision measures the proportion of positive estimations that are actually positive in the ground truth data. The TPR measures the proportion of positives in the ground truth data that are correctly identified by the system. The FPR measures the proportion of negatives in the ground truth data that are incorrectly identified by the system

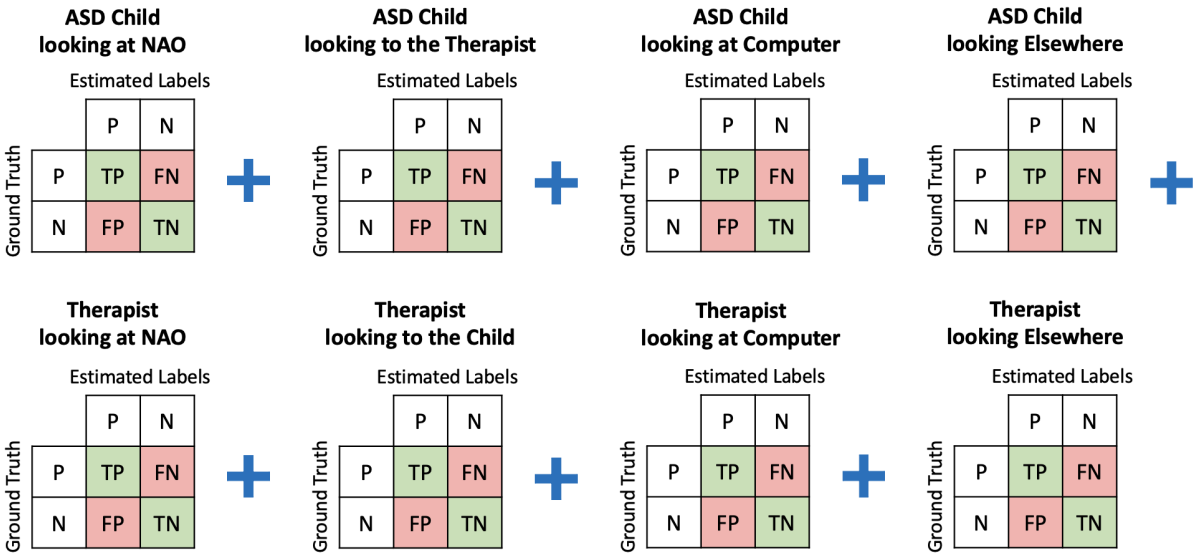


Figure 4.4: Components of the final confusion matrix for the school study

as positive.

$$\begin{cases} Accuracy = \dfrac{TP+TN}{TP+TN+FP+FN} & (4.1) \\ Precision = \dfrac{TP}{TP+FP} & (4.2) \\ Recall/TPR = \dfrac{TP}{TP+FN} & (4.3) \\ FPR = \dfrac{FP}{FP+TN} & (4.4) \end{cases}$$

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 (Therapist and Subjects) or by groups (Therapist and ASD group), in the learning approach. This was done by obtaining a confusion matrix for the therapist and one for the patients, by summing only the confusion matrices from each group, separately.

4.2.4 Attention Indices

To quantitatively evaluate the attention of the ASD subjects during the therapeutic sessions, the TFD, the SFC and the AFD were computed for looking at each target and elsewhere, based on the fixations towards each target. These attention indices (TFD, SFC and AFD) are presented in Section 2.2.2.

4.3 Proposed System

This section, presents the main results obtained for the proposed system, described in Section 3.3, along with their discussion. The results obtained by running the data from the clinical studies in the proposed system are presented following each step in Section 3.3: (i) the data preprocessing, (ii) the scene geometry analysis, (iii) the analysis of the Gaze(360) angles distribution and (iv) the areas of interest definition. At the end, the accuracy obtained for the geometrical and learning approaches in

both studies using multiple hyperparameters is presented.

4.3.1 Data Preprocessing and Curation

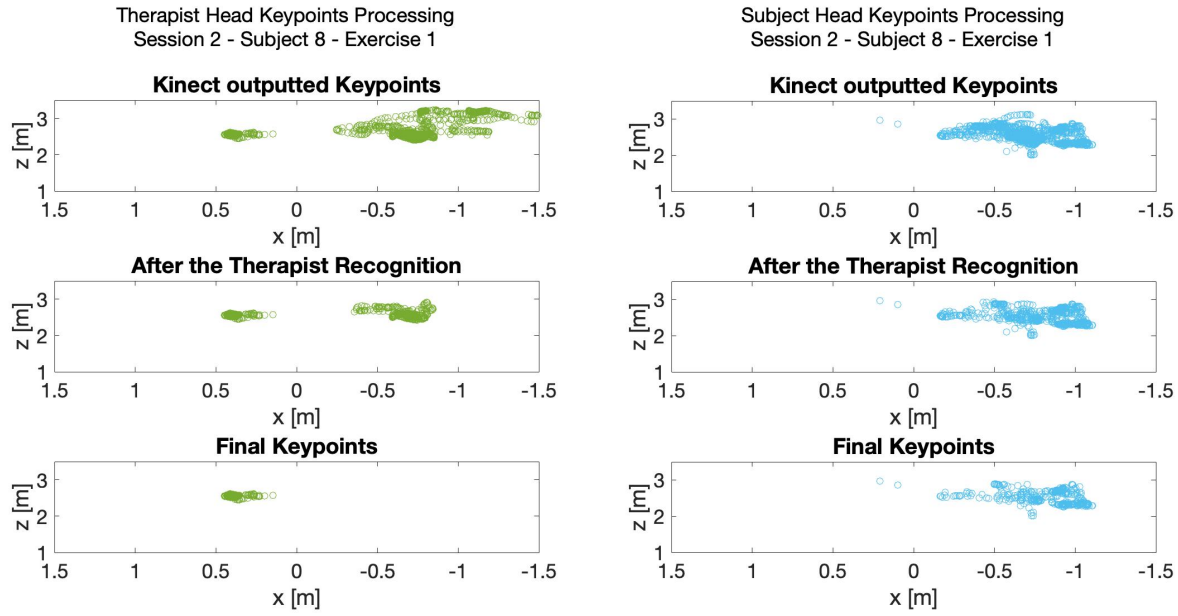
In this section, the results obtained for the data preprocessing described in Section 3.3.1 are presented for both studies. At the end, the percentage of lost data is calculated. As explained before, the preprocessing was different for each study, due to the different conditions.

Pilot Study

The Kinect keypoints data preprocessing for the therapist and the Subject 8 in Session 2 are presented in Figure 4.5. Each marker represents the 2D position of the head keypoint for each frame.

The Figure shows that the therapist skeleton computed by the Kinect is confused with the subject skeleton multiple times. This is due to the wrong reference of the red shirt during some frames of this session (segmentation failure), which was the first session with the Kinect camera. After doing the therapist recognition, by deleting the first frames with a wrong detection of the therapist skeleton, the amount of wrong keypoints decreases significantly. Then, through manual filtering, all the wrong therapist keypoints extracted by the Kinect are discarded.

For the subject, all the Kinect outputted keypoints appear to be in the right location when compared to the expected position from the setup (Figure 3.2b). The data processing only decreases the amount of useful keypoints, since it just keeps the frames in which both the therapist and the subject keypoints are well detected.



(a) Therapist

(b) Subject

Figure 4.5: Head keypoints data processing for the (a) Therapist and the (b) Subject in Session 2 with Subject 8. Each marker represents the 2D position of the head keypoint for each frame (Pilot study)

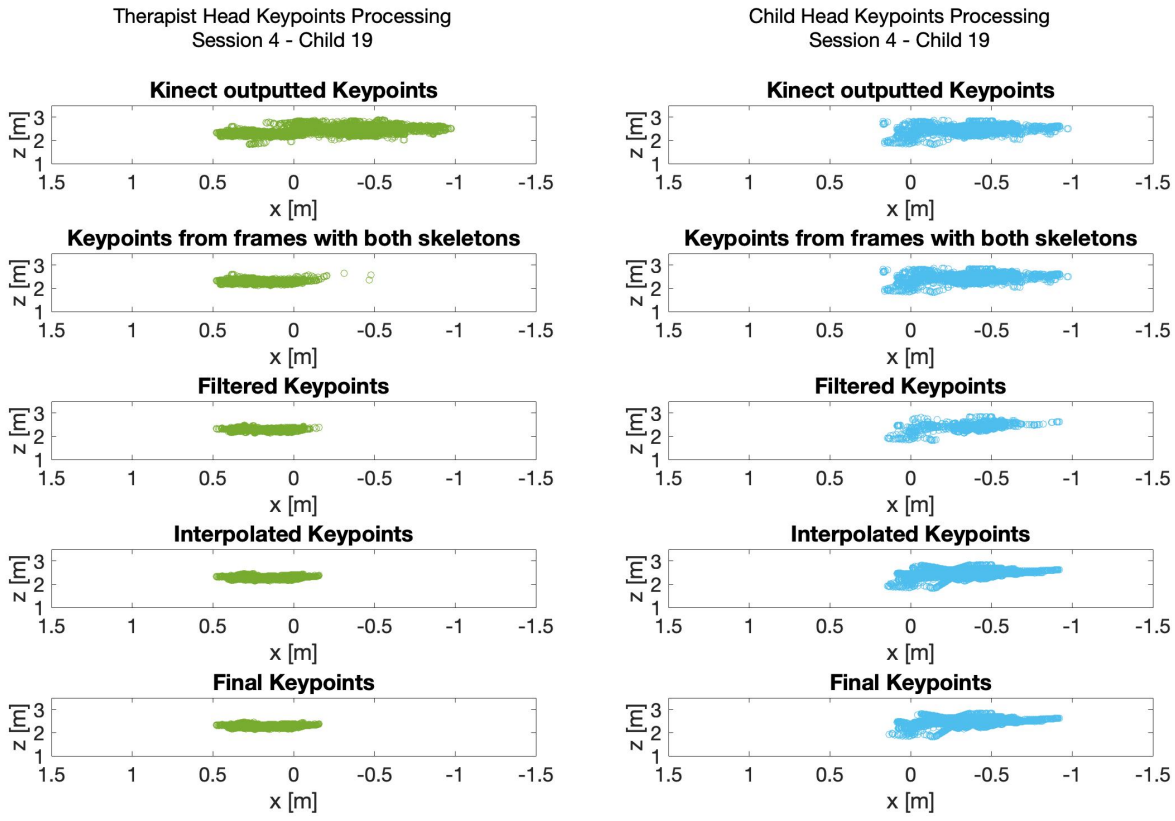
The percentage of lost data, after discarding the frames without the right detection of both skeletons and both bounding boxes, is presented in Table 4.7. For Session 2 the majority of the data is lost. This is explained by the wrong detection of the red shirt, as explained before. For the remaining sessions, approximately half of the data is lost. This is not only due to the data processing, but also due to the Kinect which is unable to output skeletons for all the acquired frames.

Table 4.7: Percentage of lost data after the pilot study data preprocessing [%] (Pilot study)

	Lost Data [%]		
	Subject 8	**Subject 10**	**Subject 21**
Session 2	83	73	75
Session 3	51	50	51
Session 4	53	50	60

School Study

The Kinect keypoints data processing for the therapist and the child in Session 4 with Child 19 are presented in Figure 4.6. Similarly to the pilot study, the therapist skeleton is confused with the child skeleton, due to the wrong detection of the red scarf. By keeping only the frames with both skeletons, most of the wrong skeleton detections are discarded and after filtering the keypoints, all the wrong detections are eliminated. As expected, by interpolating the keypoints, the amount of data increases significantly (4th plot of Figure 4.6b). At the last step, the number of keypoints decreases slightly, since the head interpolations outside the Densepose bounding boxes were discarded.

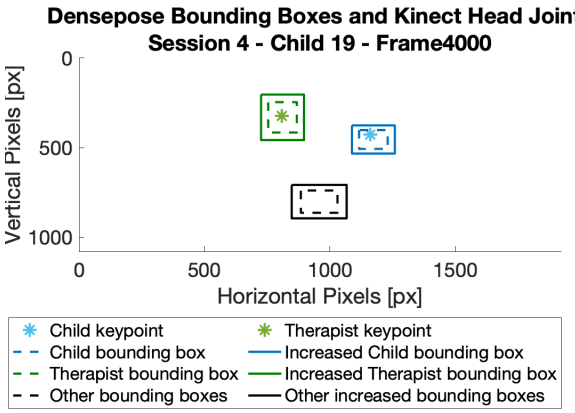


(a) Therapist

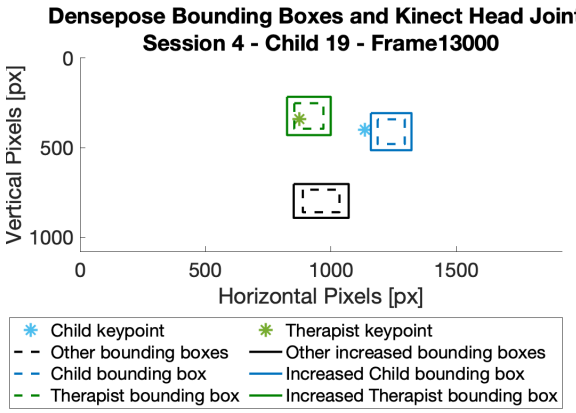
(b) Child

Figure 4.6: Head keypoints data processing for the (a) Therapist and the (b) Child in Session 4 with Child 19 (School study). Each marker represents the 2D position of the head keypoint for each frame

Figure 4.7 presents two examples of the last step of data processing. Both graphs represent two different frames with interpolated head keypoints and the respective Densepose head bounding boxes, in the 2D image coordinates. In Figure 4.7a, both keypoints are inside the increased Densepose boxes, meaning that this frame and the respective keypoints are kept. Instead, in Figure 4.7b, the interpolated child keypoint is outside the Densepose bounding box, meaning that the wrong interpolation is eliminated by discarding the respective frame.



(a) Interpolation inside the bounding boxes



(b) Interpolation outside the bounding boxes

Figure 4.7: Densepose bounding boxes and Kinect head joints in the 2D video image for 2 interpolated frames from Session 4 with Child 19 (School study)

Tables 4.8 and 4.9 show, respectively, the percentages of lost data without or with the use of interpolation. Both tables show the percentage of lost data for each session, with each child. Sessions with a percentage higher than 66% are marked in red.

For all sessions, the percentage of lost data is higher when the processing is done without interpolation. This is expected, since the interpolation reconstructs the keypoints from the discarded frames. Moreover, the amount of sessions with more than $2/3$ of the data lost is much lower with the interpolation, decreasing from 5 sessions to 1 session. Therefore, it was decided to incorporate the interpolation in the keypoints processing. Since the amount of data for Session 5 of Child 15 was still very low, this session was eliminated from the analysis.

Table 4.8: Percentage of lost data after the school study data preprocessing **without** Interpolation [%] (School study)

	Child 6	**Child 9**	**Child 10**	**Child 15**	**Child 19**
Session 3	67	13	1	15	11
Session 4	—	—	28	81	82
Session 5	51	81	41	99	—
Session 6	—	44	3	24	—
Session 7	—	—	8	—	—

Table 4.9: Percentage of lost data after the school study data preprocessing **with** Interpolation [%] (School study)

	Child 6	**Child 9**	**Child 10**	**Child 15**	**Child 19**
Session 3	52	10	1	11	9
Session 4	—	—	1	20	17
Session 5	24	55	1	95	—
Session 6	—	33	1	11	—
Session 7	—	—	0	—	—

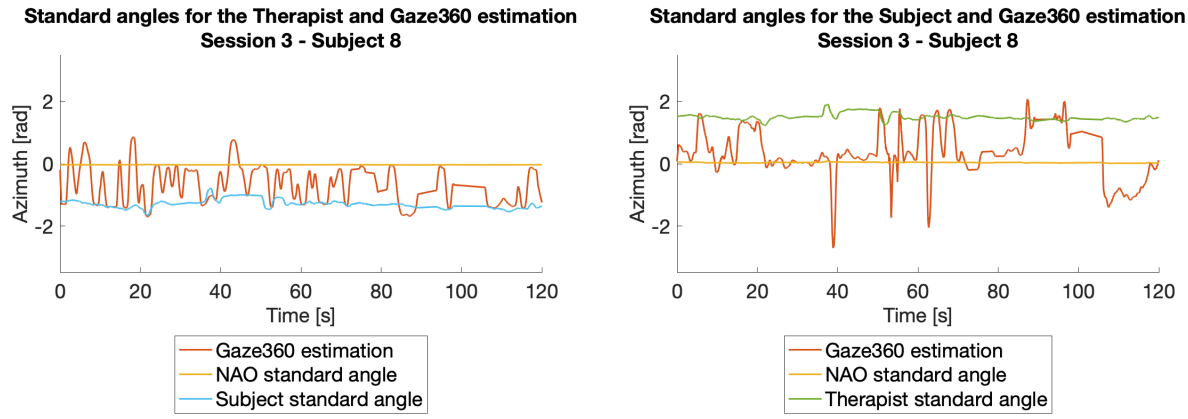
4.3.2 Scene Geometry Analysis

In this section, the results obtained after computing the standard angles, according to Equations 3.7, 3.9, 3.10, 3.11 and 3.12, for the first $120s$ of a session of each study, are presented in Figures 4.8 and

4.9. These results are shown for the therapist and ASD patient, along with the Gaze360 estimates, to show the gaze differences between them. The variation of the standard angles along time reflects the people's movements.

Pilot Study

Comparing the Gaze360 estimation with the standard angles, it can be observed that the Therapist gaze is distributed between the two targets (Subject and NAO), while the Subject looks mostly at one (NAO).

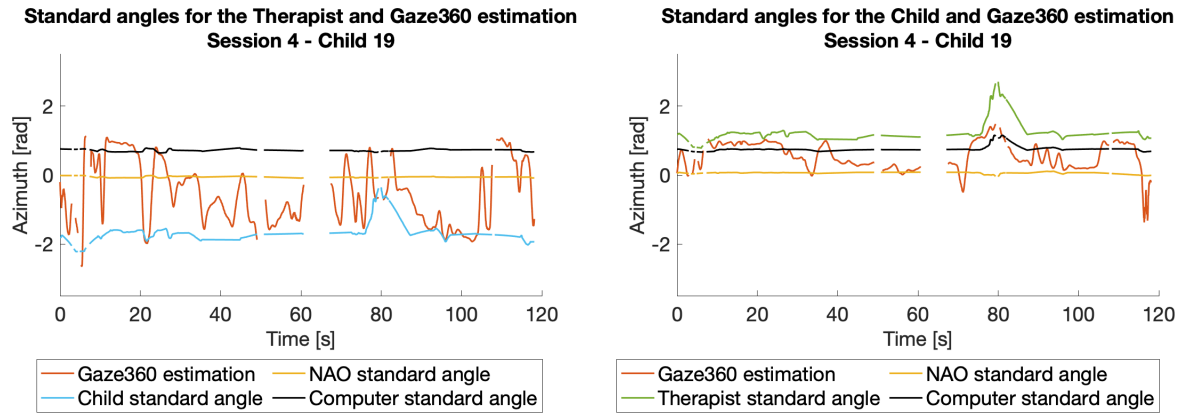


(a) Therapist (b) Subject

Figure 4.8: Standard angles and Gaze360 estimations for the (a) Therapist and the (b) Subject during the first $120s$ of Session 3 with Subject 8 (Pilot Study)

School Study

For this second study, the presence of a third target (Computer) is clear in the gaze pattern of the Therapist. For the Child, it is difficult to distinguish whether the Child is looking at this third target, the Computer or to the Therapist, since the standard angles for both targets are very close to each other and the Gaze360 estimation is in the middle of them.



(a) Therapist (b) Child

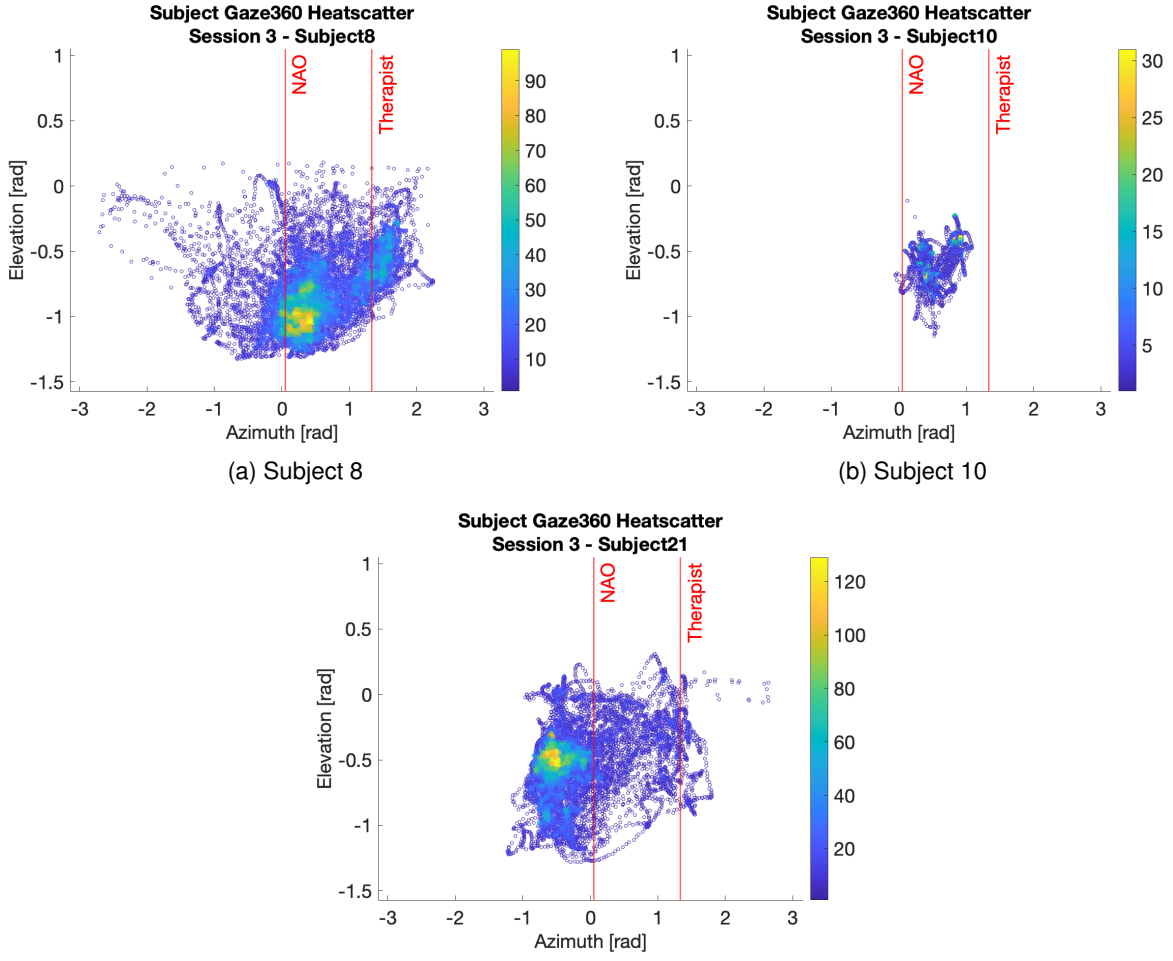
Figure 4.9: Standard angles and Gaze360 estimates for the (a) Therapist and the (b) Child during the first $120s$ of Session 4 with Child 19. The gaps across time correspond to the frames discarded during the data preprocessing from Section 4.3.1 (School Study)

4.3.3 Analysis of Gaze(360) Angles Distribution

In this section, the angles distribution of the Gaze360 estimations are presented. As explained in Section 3.3.3, an offset between the location of the biggest Subject's cloud and the expected angles of NAO is visible. Therefore, the Gaze360 model was compared with other gaze and head estimators to find the origin of this problem. Later, an automatized extraction of these offsets was performed, by obtaining histograms centralized to each target and finding the closest maximum to the center of each histogram. Since the maximums change with the bar widths of the histograms, the best bar widths were found for each study, as described in Section 3.3.3.

To analyse the distribution of the Gaze360 estimates, the scatter plots in Figure 4.10 were obtained for three different subjects during Session 3 of the Pilot Study. These plots represent the Gaze360 estimates in terms of azimuth and elevation. Since in the pilot study, the participants were in fixed positions, these plots can be interpreted as the location were people looked the most. The colors express the number of times the Subject's gaze was directed towards that zone, with a warmer color representing an higher number.

The azimuth angles for looking at NAO and to the Other Person, using the distances of Figure 3.2b, are also plotted as red lines.



(a) Subject 8

(b) Subject 10

(c) Subject 21

Figure 4.10: Gaze360 estimations, expressed in terms of elevation and azimuth angles, for Session 3 with Subjects (a) 8, (b) 10 and (c) 21

In the gaze plot of Subject 10, represented in Figure 4.10b, it is not possible to distinguish clouds, corresponding to looking at each target. The lack of this distinction plus the fact that his/her sessions

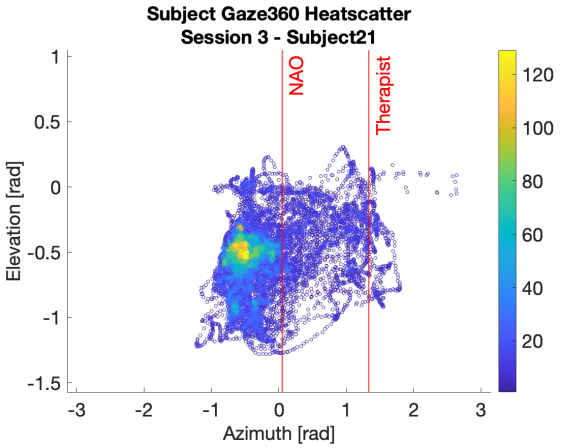
were shorter and consequently had a smaller amount of useful data led to the exclusion of this subject from the data analysis.

For the gaze plots of Subjects 8 and 21, represented in Figures 4.10a and 4.10c, the biggest clouds correspond to looking at NAO. However, their centroids are deviated from the expected azimuth angle. Thus, the Gaze360 estimations were compared with the OpenFace ones and with two head estimators (RT-Gene and WHENet) to understand the possible causes of these deviations. This comparison is reported in Figure 4.11.

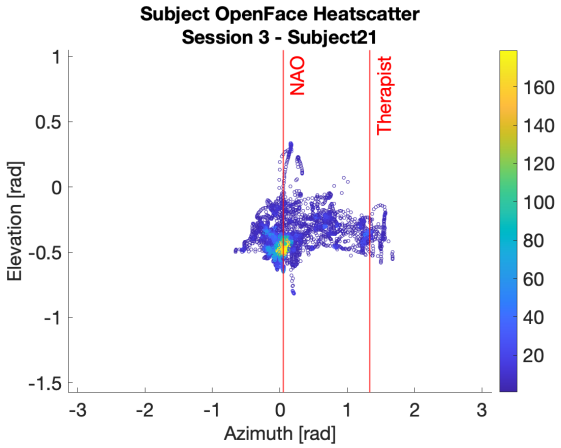
From the OpenFace output, shown in Figure 4.11b, the gaze estimations do not have an offset in the azimuth angle. The cloud corresponding to looking at NAO is centered at the expected angle. Thus, the Gaze360 offset is intrinsic to the model since it does not affect the other gaze estimator.

Observing the head pose plots from WHENet and RT-Gene (Figure 4.11c and 4.11d, respectively), the biggest cloud is centered at the expected angle in the former estimator, while, it has an offset to the opposite side of the Gaze360 cloud for the latter.

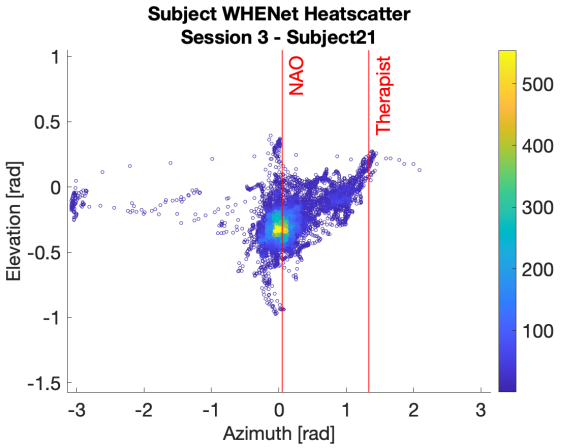
Therefore, Gaze360 has an intrinsic offset not related with the head pose.



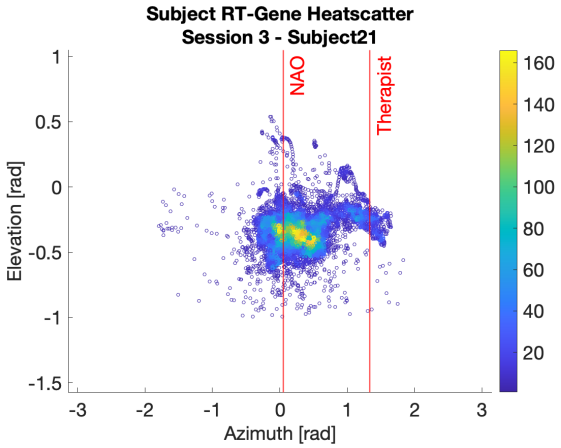
(a) Gaze360 gaze estimation



(b) OpenFace gaze estimation



(c) WHENet head estimation



(d) RT-Gene head estimation

Figure 4.11: Estimations, expressed in terms of elevation and azimuth angles, from the different gaze ((a) Gaze360 and (b) OpenFace) and head ((c) WHENet and (d) RT-Gene) models in Session 3 of Subject 21

Pilot Study

Before correcting the Gaze360 offsets, the best bar width for the histograms of the pilot study was found. Four bin widths were experimented on the eight centralized histograms (2 targets x 4 people - Subject 8, 21, Therapist (8 and 21)) of Session 3. The mean and 95% confidence interval of the number of peaks of these histograms were obtained (Table 4.10).

Table 4.10: 95% Confidence Interval (CI) of the number of maximums in the centralized histograms using 4 different bin widths for Session 3

Bin width [rad]			
0.25	**0.35**	**0.5**	**0.75**
4.1 ± 0.9	2.3 ± 0.3	1.8 ± 0.3	1.3 ± 0.3

For the pilot study, 2 maximums are expected for each histogram, given the existence of 2 targets. Observing the average number of maximums obtained for each bin width, the 2 widths with a value closer to 2 are $0.35rad$ and $0.5rad$. Considering there could be an operator in the room, the histograms can sometimes have 3 maximums. Thus, at least 2 maximums are expected, being the width of $0.35rad$ chosen.

Using the bin width of $0.35rad$, the histograms obtained for Session 3 of Subject 21 are presented in Figure 4.12. The location in the x-axis of the red lines corresponds to the maximums of the histogram, which represent the Gaze360 offset from the center of the each target. Therefore in this Figure, as observed in Figure 4.10c, it is confirmed that the Gaze360 estimations have an offset. Later, it is studied the effect of correcting these offsets in the proposed system performance.

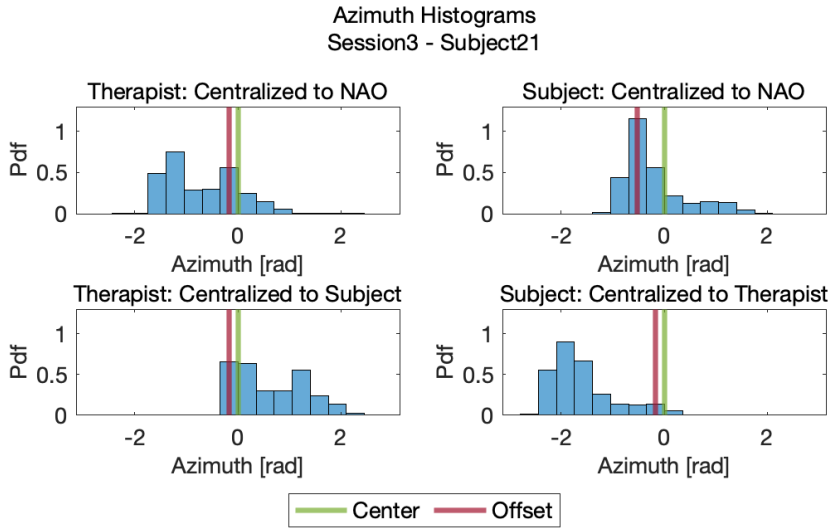


Figure 4.12: Centralized Histograms of the Gaze360 estimations for the Therapist and the Subject in Session 3 of Subject 21. The green lines represent the center of the histograms ($0rad$) and the red lines represent the computed offsets. Pdf: Probability distribution function

School Study

Before correcting the Gaze360 offsets for the school study, the best bar widths for the histograms were found using a process similar to the one used for the pilot study. However, since the increase of the Densepose bounding boxes, used as final step of the data processing, affects the amount of useful data

and, consequently, the histograms, multiple ratios were tested. The 95% CI of the number of maximums of the centralized histograms for each bin width was obtained for the Therapist and for the Subjects, since the targets relative position is different between them. The results are presented in Table 4.11.

Table 4.11: 95% Confidence Interval (CI) of the number of maximums in the centralized histograms of the subjects and therapist using 3 different bin widths for Session 3 with all the children

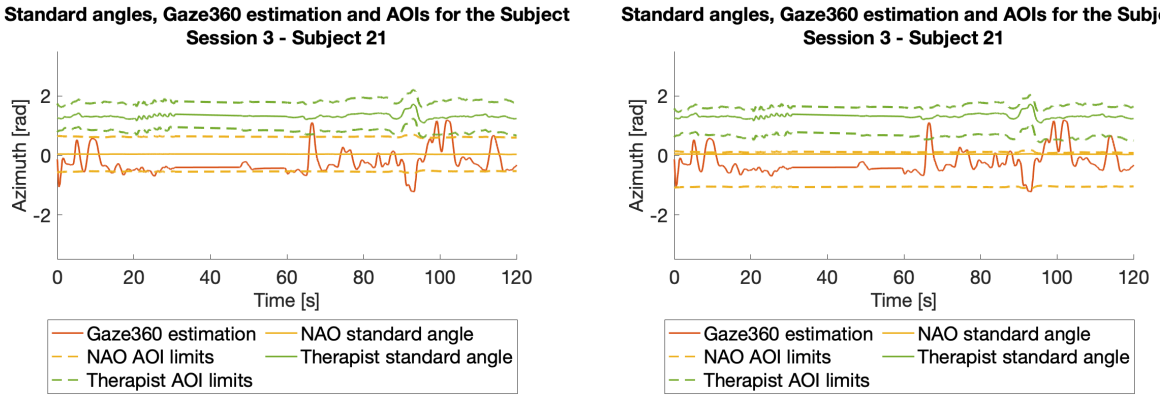
Therapist		**Bin width [rad]**		
		0.25	**0.35**	**0.50**
Gaze360 Bounding Boxes Increase	**25%**	3.6 ± 0.3	2.7 ± 0.4	2.1 ± 0.5
	50%	3.6 ± 0.4	2.7 ± 0.4	1.9 ± 0.4
	75%	3.8 ± 0.5	2.7 ± 0.4	1.9 ± 0.4

Subjects		**Bin width [rad]**		
		0.20	**0.25**	**0.35**
Gaze360 Bounding Boxes Increase	**25%**	4.2 ± 0.8	3.3 ± 0.8	1.7 ± 0.4
	50%	4.2 ± 0.9	3.2 ± 0.8	1.8 ± 0.4
	75%	4.3 ± 1.0	3.1 ± 0.8	1.8 ± 0.4

Regarding the confidence intervals, there are no big changes between the different Densepose bounding boxes increases. Moreover, for the school study, 3 maximums are expected for each histogram. Considering the obtained average number of maximums for each width, the best bin width for the Therapist histograms is $0.35rad$. While, for the Subject histograms, is $0.25rad$.

4.3.4 Areas of Interest Definition

Each AOI was defined according to the scheme in Figure 3.23 of Section 3.3.4. The results obtained for the Subject 21 in Session 3 of the pilot study are shown in Figure 4.13. This Subject had a visible Gaze360 offset on the cloud corresponding to looking at NAO, as shown in Figure 4.10c. Thus, the results are presented without the Gaze360 offsets correction in Figure 4.13a, and with the Gaze360 offsets correction in Figure 4.13b, where the shift in the azimuth of the NAO AOI is highlighted.

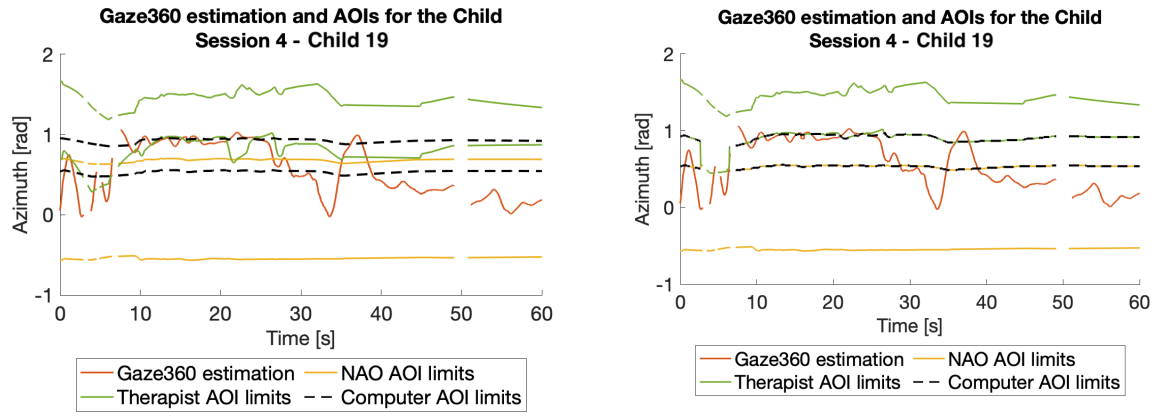


(a) Before adding the Gaze360 offsets

(b) After adding the Gaze360 offsets

Figure 4.13: Standard angles, Gaze360 estimation and Areas of Interest limits, (a) before and (b) after adding the Gaze360 offsets, for the Subject during the first $120s$ of Session 3 with Subject 21 (Pilot study)

The results obtained for the AOIs definition of the Child 19 in Session 4 of the school study are shown in Figure 4.14a. To overcome the AOIs overlapping problem, Gaussian distributions were estimated from the AOIs, see Figure 4.15, and used to determine the frontiers of the AOIs (intersections), as explained in Section 3.3.4. They were calculated for each instant with overlapping AOIs and for a fixed increase of the Densepose bounding boxes ($p = 25\%$ in Equation 3.6). To analyse the effect of the hyperparameter k, from Equation 3.21, the Gaussian curves were constructed for $k = 1$ and $k = 2$. As expected, increasing the k parameter, decreases the standard deviation of the Gaussians, changing the location of their intersection.



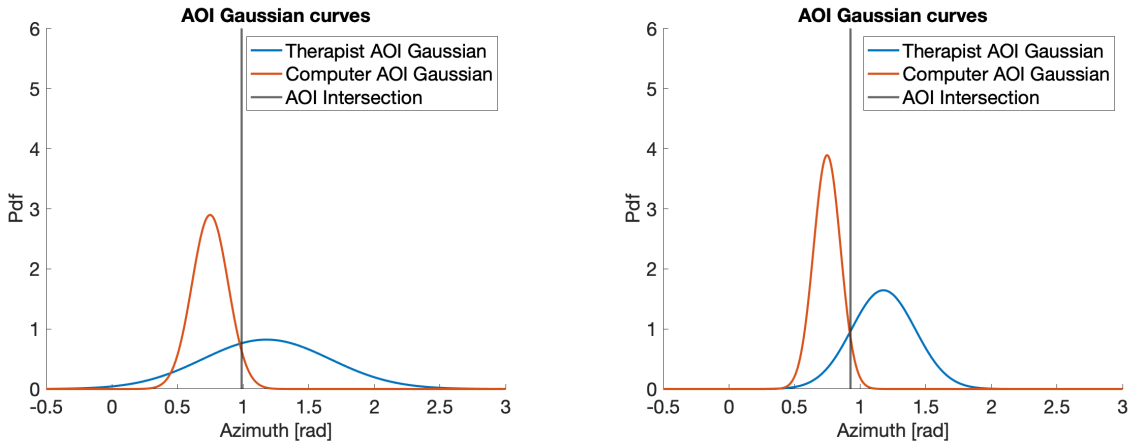
(a) Before correcting the AOIs overlapping

(b) After correcting the AOIs overlapping

Figure 4.14: Gaze360 estimation and AOIs limits, (a) before and (b) after correcting the AOIs overlapping, for the Child during the first $60s$ of Session 4 with Child 19. The gaps across time correspond to the frames discarded during the data preprocessing from Section 4.3.1 (School Study)

The final AOIs, after correcting their overlapping using $p = 25\%$ and $k = 2$, are shown in 4.14b.

In Figures 4.13 and 4.14, the Therapist AOI limits are more noisy than the other targets AOI limits, since the distance between the people is smaller than the distance to NAO and the Computer. Moreover, both people move during the sessions while the robot and the computer are in fixed locations for the majority of the sessions.



(a) $k = 1$

(b) $k = 2$

Figure 4.15: AOI Gaussian curves intersection for (a) $k = 1$ and (b) $k = 2$ (School study)

4.3.5 Geometrical Approach

In the geometrical approach, referred in step 2 of the AOIs definition described in Section 3.3.4, the widths of the AOIs are decided based on the geometrical dimensions of the targets and the Gaze360 noise obtained from the controlled experiments described in Section 3.2.2. In this section, the results obtained for the Gaze360 noise calculation are presented.

To calculate the standard deviation for looking at each fixation point, a manual segmentation was performed to the signals obtained from the first experiment of the long distance benchmarking (Figure 4.16). The standard deviation was only calculated for the fixation points in front of the subject (P1, NAO, P4 and P5) and for the fixation point corresponding to looking to the therapist (THERAPIST).

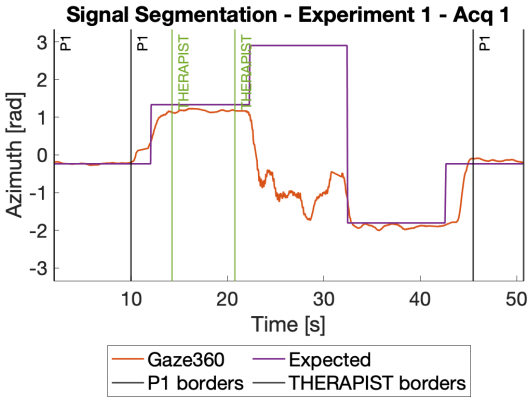


Figure 4.16: Gaze360 and expected signals segmentation for acquisition 1 of Experiment 1 of the long distance benchmarking

The computed standard deviations are presented in Table 4.12 The highest standard deviation corresponds to looking at the fixation point P1. This is expected, since there are more segments looking at this fixation point and, thus, more available data. This standard deviation was chosen as the Gaze360 noise, which was added to the predefined range of angles, as explained in Section 3.3.5.

Table 4.12: Standard deviations for looking at the several fixation points

	P1	**THERAPIST**	**NAO**	**P4**	**P5**
std [rad]	**0.069**	0.051	0.047	0.038	0.028

Pilot Study

After establishing the whole system, having the AOIs defined and the gaze estimations, the system hyperparameters were validated by computing the accuracy of the proposed system, comparing the gaze classification estimations with the ground truth classification. For the pilot study, the hyperparameters chosen were the Gaze360 offsets and the k, from Equation 3.21, used to define the standard deviation of the Gaussians when 2 AOIs overlap. Since most of the times all the performance scores improve together, only the accuracy was analysed for the hyperparameters validation.

To analyse the effect of each hyperparameter, the system was run with and without the Gaze360 offsets correction, and for different AOIs intersection points, dependent on the values of $k = \{1, 2, 3\}$ as explained in Section 3.3.4. The results for the several hyperparameters using the validation set are shown in Table 4.13.

Table 4.13: Geometrical approach accuracy of the proposed system classifying the gaze as looking at the different targets (NAO, Other Person and Elsewhere) for the different hyperparameters configurations using Session 3 as validation set [%] (Pilot Study)

$k\sigma$	**Without Offsets Correction**	**With Offsets Correction**
3σ	69.2	79.8
2σ	69.2	79.8
1σ	69.2	79.8

The system reacted the same way for the different k values, probably because there were only two targets, which means the AOIs did not overlap for the majority of the session.

By correcting the Gaze360 offsets, the system performance improved significantly, which proves the system sensitivity to this hyperparameter, as well as its importance. Consequently, the best hyperparameters configuration is to:

- Correct the Gaze360 offsets;
- Use $k = 2$.

Using these hyperparameters, the system has a good performance for all the metrics in Sessions 3, the validation set, and 4, the test set (Table 4.14). Thus, the system generalizes well for the chosen hyperparameters. However, the performance is worse for Session 2, which occurs due to the small amount of useful data in this session.

Table 4.14: Geometrical approach system performance scores, classifying the gaze, using the chosen hyperparameters configuration [%] (Pilot Study)

	Accuracy	**Precision**	**Recall/TPR**	**FPR**
Session 2 (Test set)	67.7	51.6	51.6	24.2
Session 3 (Validation set)	79.8	69.8	69.7	15.1
Session 4 (Test set)	82.4	73.8	73.0	12.9

School Study

After establishing the whole system, its hyperparameters, for the school study, were validated by computing the accuracy of the proposed system through the comparison of the gaze classification estimations with the ground truth classification. For the school study, the main hyperparameters are the Densepose bounding boxes ratio (p), the Gaze360 offsets and variable k, from Equation 3.21, used to define the standard deviation of the Gaussians when 2 AOIs overlap. To analyse the effect of each hyperparameter, multiple combinations were tested. The Densepose bounding boxes were increased by 25%, 50% and 75%. The system was tested with and without doing the Gaze360 offset correction. For the AOIs overlapping, k was tested for the values $k = \{1, 2, 3\}$. The results for the several hyperparameters using the validation set are shown in Table 4.15.

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. At the end, the best hyperparameters configuration corresponds to:

- Don't correct the Gaze360 offsets;
- Use $k = 1$;

- Use $p = 50\%$.

Table 4.15: Geometrical approach accuracy of the proposed system classifying the gaze as looking at the different targets (NAO, Other Person, Computer and Elsewhere) for the different hyperparameters configurations using Session 6 as validation set [%] (School study)

	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 system performance scores were computed for the test set, as shown in Table 4.16. The system 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 4.16: Geometrical approach system performance scores, classifying the gaze, using the chosen hyperparameters configuration [%] (School study)

	Accuracy	**Precision**	**Recall/TPR**	**FPR**
Session 3 (Test set)	81.1	62.1	62.0	12.6
Session 4 (Test set)	80.8	61.7	61.5	12.7
Session 5 (Test set)	79.2	58.3	58.1	13.8
Session 6 (Validation set)	80.1	60.2	60.1	13.2
Session 7 (Test set)	83.9	67.8	67.7	10.7

4.3.6 Learning Approach

In the learning approach, referred in step 2 of the AOIs definition described in Section 3.3.4, the best widths for the AOIs were found using one session as training set of the system. To determine the best widths, ROC curves (Figure 4.17), as explained in Section 3.3.6.

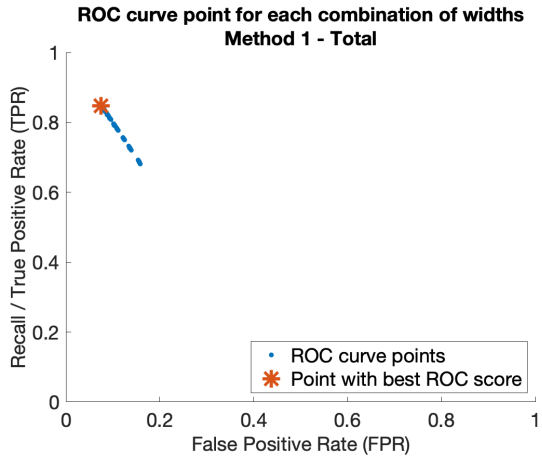


Figure 4.17: ROC curve points for each combination of widths using the obtained system evaluation metrics for Session 3 (training session) of the pilot study

After, the training set was used to calculate the best AOI widths combinations using different configurations of hyperparameters, the validation set was used to find the best configuration of hyperparameters and, finally, the test set was used to obtain the system performance scores when classifying the gaze.

Pilot Study

Using the training set, the best widths were obtained for the different hyperparameters configurations. For the pilot study, the main hyperparameters are the Gaze360 offsets and the k parameter, from Equation 3.21. Since, as observed in the geometrical approach, the AOIs do not overlap frequently in this study, the hyperparameter k was not analysed. For the learning approach, the effect of using group widths or total widths was also studied.

The best widths by group and in total, with and without the Gaze360 offsets correction, are presented in Table 4.17. The system performance scores computed for the test set are presented in Table 4.18.

Table 4.17: Best widths for Session 3 (training set) with and without the Gaze360 offsets correction [m] (Pilot study)

		Without offsets correction		**With offsets correction**	
		NAO	Other Person	NAO	Other Person
By group	ASD Group	3.0	1.4	3.0	1.0
	Therapist Group	2.6	2.0	3.0	1.6
In total		3.0	1.4	3.0	1.6

The performance scores are lower in Session 2 for all the hyperparameters configurations, which is explained by the few available data for this session.

All the results improve when correcting the Gaze360 offsets, with some metrics improving more than 10% for Session 2. For Session 4, the improvements are much smaller, demonstrating that the system sensitivity to the offsets correction is different between sessions. This fact can be justified by the presence of different offsets among the several sessions.

Comparing the results when using AOIs widths by group or in total, the changes in the scores are small and contradictory. For Session 2, without the offsets correction, the results are better using group widths, while with the offsets correction, they are better using the same widths for all people. For Session 4, the contrary happens. Without the offsets correction, the results are better using the total widths, while with the offsets correction, the performance is higher using the best widths by group.

Since Session 2 has fewer useful data than Session 4, it was considered that the results from the last Session are more reliable. Thus, the best configuration for this study corresponds to:

- Correct the Gaze360 offsets;
- Use $k = 2$;
- Use the best AOI widths by group.

Table 4.18: Learning approach accuracy of the proposed system classifying the gaze as looking at the different targets (NAO, Other Person and Elsewhere) for the different hyperparameters configurations using Session 2 and 4 as test set [%] (Pilot study)

Widths		**Without offsets Correction**	**With offsets correction**
By group	Session 2 (Test set)	73.6	79.5
	Session 4 (Test set)	86.9	**88.6**
In total	Session 2 (Test set)	72.9	80.3
	Session 4 (Test set)	87.2	88.4

The final system performance scores after defining the best configuration are in Table 4.19. Analysing the scores obtained for training set (Session 3) and the ones obtained for the test set (Session 2 and 4), it is concluded that the system generalizes well. The scores are high, meaning that the performance of the proposed system is good.

Table 4.19: Learning approach system performance scores, classifying the gaze, using the chosen hyperparameters configuration [%] (Pilot study)

	Accuracy	**Precision**	**Recall/TPR**	**FPR**
Session 2 (Test set)	79.5	69.2	69.2	15.4
Session 3 (Training set)	90.7	86.1	85.9	6.9
Session 4 (Test set)	88.6	83.0	82.8	8.5

Overall, it is considered that the proposed system for the pilot study using the learning approach has a good performance, given the study conditions, the wrong keypoints detection from the Kinect and the Gaze360 intrinsic offsets for the subjects.

School Study

As explained in the geometrical approach, for the school study, the main hyperparameters are the Densepose bounding boxes ratio (p), the Gaze360 offsets and variable k. For the learning approach, the effect of using group widths or total widths is also studied. To analyse the effect of each hyperparameter, multiple combinations were tested as shown in Table 4.20.

Table 4.20: Learning approach accuracy of the proposed system, classifying the gaze as looking at the different targets (NAO, Other Person, Computer and Elsewhere) for the different hyperparameters configurations using Session 6 as validation set [%] (School study)

		Without Offsets Correction			**With Offsets Correction**		
Widths	$k\sigma \setminus p$	**25%**	**50%**	**75%**	**25%**	**50%**	**75%**
By 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

Observing the results, the accuracy increases for all the hyperparameters configurations when the Gaze360 offsets are not corrected and when the best widths by group are used.

Analysing the hyperparameters effect, the Gaze 360 offsets correction is the main hyperparameter, being its variation the main source of change of the performance scores, proving that the proposed system is more sensitive to it. For all the other hyperparameters combinations, the accuracy is better without the offsets correction, as opposed to the pilot study. This occurs, because both the Therapist and the Computer targets (for the Child) are shifted to the same position when correcting the offsets. The main reason is the close location of these two targets, which imply the existence of only one maximum in the centralized histograms. Thus, one of the following two situations happens frequently: (1) the AOIs are too small or (2) one of them is totally overlapping the other. The last situation results in the exclusion of the Computer AOI, since the Therapist is closer to the Child and, therefore, the one blocking the view towards the Computer. Consequently, the performance decreases for these two targets, leading to lower total scores.

The parameter k, from Equation 3.21, seems to affect the system 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 system 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.

When the Gaze360 offsets are not corrected and the best widths by group are being used, the best performance is for $k = 2$ and $k = 50\%$. In this way, the best configuration of hyperparameters corresponds to:

- Don't correct the Gaze360 offsets;
- Use $k = 2$;
- Use $p = 50\%$;
- Use the best AOI widths by group.

Table 4.21: Learning approach system performance scores, classifying the gaze, using the chosen hyperparameters configuration [%] (School study)

	Accuracy	**Precision**	**Recall/TPR**	**FPR**
Session 3 (Training set)	83.8	67.5	67.4	10.8
Session 4	82.0	63.9	63.9	12.0
Session 5	84.6	69.4	69.0	10.2
Session 6 (Validation set)	82.2	64.5	64.3	11.8
Session 7	89.1	78.2	78.2	7.3

Analysing the performance scores, shown in Table 4.21, it was concluded that the proposed system is generalizing well, having high and consistent performance metrics for all the sessions, including accuracy values always above 80%. Overall, the scores are considered to be very good, given the study conditions and the wrong keypoints detection from the Kinect.

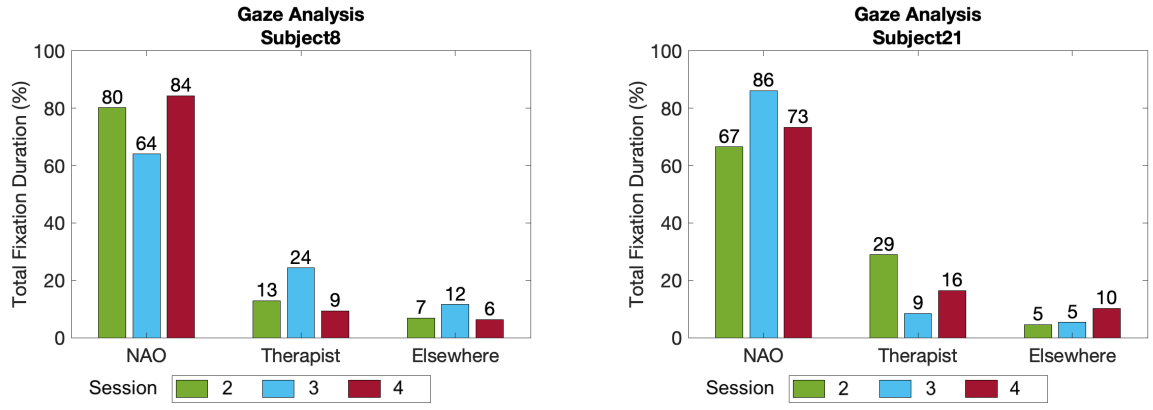
Comparing the system performance scores (Tables 4.14, 4.16, 4.19 and 4.21), the learning approach was considered to be the best one, outperforming the geometrical approach. It has higher scores for all the metrics in both studies, which was expected, since the best widths were obtained using the training session.

4.4 Attention Analysis

In this section, an attention analysis of the ASD subjects is done using the best approach and hyperparameters. This analysis allows to understand the on-task attention of each subject, by computing the attention towards each target and elsewhere. It can also give a feedback about future improvements on the setup and protocol of the therapy sessions.

4.4.1 Pilot Study

The attention of both subjects along the sessions is presented in Figure 4.18. These graphs show the TFD towards each target and elsewhere along the sessions.



(a) Subject 8

(b) Subject 21

Figure 4.18: TFD towards the targets and elsewhere along the sessions for Subjects (a) 8 and (b) 21 [%] (Pilot study)

Both subjects look more at NAO for all the sessions and less to elsewhere, which proves their engagement in the therapy sessions and their interest in the robot. There is a relation between looking at NAO and to the Therapist. For the sessions in which they look less at NAO, the attention towards the Therapist increased.

To analyse the fixations, the SFC and AFD were computed, as shown in Table 4.22. The percentage of fixations towards each target and elsewhere is stable for both subjects between sessions. Moreover, the AFD towards the robot is the highest for all sessions for both subjects, reinforcing their strong interest in NAO.

Table 4.22: SFC and AFD towards each target and elsewhere along the sessions for Subjects 8 and 21 (Pilot study)

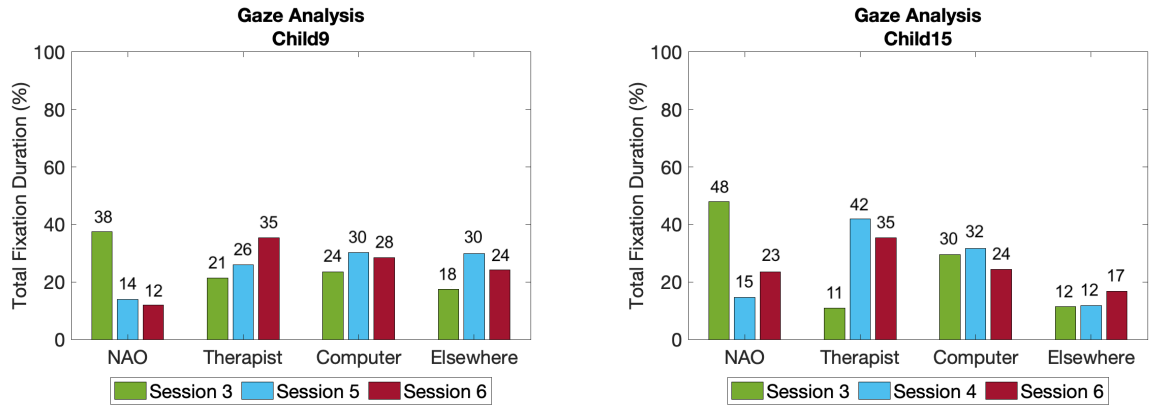
		Subject 8			Subject 21		
		NAO	**Therapist**	**Elsewhere**	**NAO**	**Therapist**	**Elsewhere**
Session 2	**SFC [%]**	68	20	11	40	37	48
	AFD [s]	7	4	4	9	4	1
Session 3	**SFC [%]**	47	29	25	37	20	43
	AFD [s]	10	6	3	22	4	1
Session 4	**SFC [%]**	54	29	18	48	23	30
	AFD [s]	11	2	2	14	7	3

According to the therapist, Subject 21 is more focused on the task, while, Subject 8 is more oriented towards personal interactions. This explains the higher values of AFD towards NAO for Subject 21. However, Subject 8 does not seem to interact more with the therapist, presenting lower levels of TFD, AFD and SFC for some of the sessions, when comparing with Subject 21. This lower result could be justified by the presence of another person in the room, the operator, which was not considered in this analysis, and by the fact that the protocol was designed for children and not adults, impacting their performances and levels of attention.

4.4.2 School Study

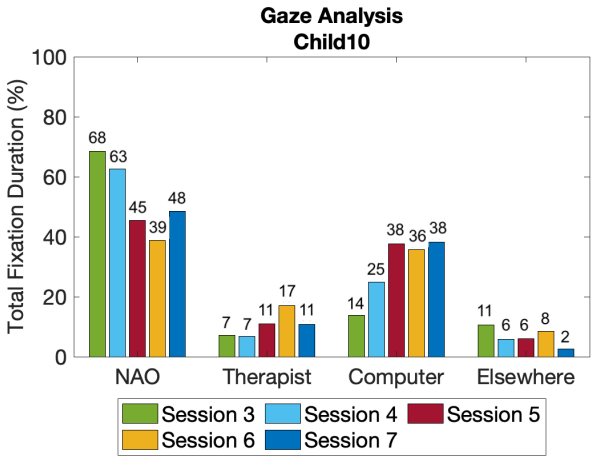
The TFD along sessions (Figure 4.19) was obtained for each child, as well as the individual accuracies in each session (Table 4.23), in order to relate with the therapist qualitative feedback of the sessions (Table 4.24). The attention was only studied for children with more than 2 analysed sessions (Child 9,

10 and 15).



(a) Child 9

(b) Child 15



(c) Child 10

Figure 4.19: TFD towards the targets and elsewhere along the sessions for Children (a) 9, (b) 15 and (c) 10 [%] (School study)

For all children, in the green session, Level 3 of the protocol was performed, in the light blue session, Level 4, and in the remaining sessions, Level 3 and 4. Thus, it is expected that the interest in the Computer is lower in the first session and higher in the remaining, due to Level 4. This is well verified for Child 10 (Figure 4.19c). For Children 9 and 15 (Figures 4.19a and 4.19b), 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 actualized in order to keep the children engaged and stimulated.

4.4.3 Further Insights of the School Study

In this section, the proposed system performance and estimated attention, and the qualitative therapist feedback are compared to take conclusions about their agreement. Observing the system accuracy obtained through the ground truth comparison of the classifications, for each child, it is visible that the system performs worse for Child 9, followed by Child 6 and 15 (Table 4.23).

Table 4.23: System accuracy, classifying the gaze, for each Child in each Session (School study)

	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 4.24, and the attention analysis, present in Figure 4.19, 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 system performance;
- Child 9 interacts well with the robot, however likes to touch it and uses the scenarios to play. Thus, he/she also moves a lot, which causes Kinect detection problems and deteriorates the performance of the proposed system. Due to the toys, the amount of time looking elsewhere is also higher than for the other children, as shown in Figure 4.19a.
- 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, justifying and his/her performance seems to have improved along sessions, which is reflected in the system attention estimations showing a high attention towards NAO (Figure 4.19c).
- 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.
- 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 of the proposed system.

Therefore, there is an agreement between the therapist feedback, the system performance scores and the system estimations of attention for each child. Moreover, after showing the Figures 4.19 to the therapist, she reported that the results obtained were in accordance with what she was expecting, 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 system as an explainable AI tool.

Table 4.24: Therapist qualitative analyse of the sessions (School study)

	Session 3	**Session 4**	**Session 5**	**Session 6**
Child 6	- Very curious about NAO; - Likes to touch NAO; - A lot of difficulty.	—	- Imitated the adult sometimes; - Offered a lot of resistance to the work; - Not an easy day.	—
Child 9	- Imitated the robot; - Did not wait for his/her turn; - Showed a lot of curiosity; - Likes to touch NAO.	—	- Recognized the scenarios; - Uses the scenarios to play.	- Interacted well; - Uses the scenarios to play.
Child 10	- Interacted well; - Did not imitate the robot; - Likes to give it orders; - Repeated therapist verbally; - Did not imitate all the gestures.	- Recognized some scenarios; - Had difficulty in memorization.	- Remembered some scenarios; - Interacted with NAO; - Performed the gestures.	- Interacted well; - Imitated the robot; - Spoke with NAO; - Gave verbal orders to NAO.
Child 15	- Interacted well; - Did not wait his/her turn; - Difficulty imitating NAO; - Better performance with therapist instructions.	- Difficulty imitating NAO; - Facility imitating the therapist; - Did not wait his/her turn; - Did not know the scenarios.	—	- Imitated the adult and NAO; - Did not wait his/her turn; - Can not remember the gestures, in his/her turn.
Child 19	- Did everything, waiting his/her turn.	- Did everything, waiting his/her turn.	—	—

Chapter 5

Conclusion and Future Work

ASD children show severe 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 thesis 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 leads to the selection of a camera which was placed at a certain distance from the children, hardening the gaze estimation process. Considering also the difficult conditions of the sessions, the data estimated from the Kinect camera was quite noisy, making this problem even more challenging, requiring a lot of data curation and filtering, where a considerable quantity of information was discarded, even when using interpolation to reconstruct the data.

The proposed system consists in extracting the gaze and defining the AOIs, followed by a gaze classification into different targets. Before extracting the gaze, a benchmarking of gaze and head pose estimators was performed, from which the Gaze360 model was chosen as the most adequate one. To define the AOIs, two approaches were analysed: a geometrical approach and a learning approach. In the geometrical approach the widths of the AOIs were defined based on the target geometry, while in the learning approach the system was trained to find the best widths for each AOI. Given the performance metrics of the system, the learning approach was chosen as the best one, outperforming the geometrical approach, with the proposed system reaching a total accuracy higher than **82.0%** for all the sessions of the school study. Comparing with the state of the art, our system performed better than the one proposed in [50], which was based on the head pose and achieved an accuracy of 73.5%. Moreover, for all the children in all the sessions, the system 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.

The directions for future work are multi-fold:

- New clinical studies: new clinical studies should be performed, including more participants and sessions, for a better understanding of the children's attention patterns.
- Improvements on the proposed attention system: there are some problems, mainly regarding the Densepose head bounding boxes and the Kinect skeletons detection. For the Densepose bounding boxes, there are two main issues. The model has a slow processing taking too long to extract the gaze from the videos, being 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. Although, they have to be precise, since the head bounding boxes affect the Gaze360 model significantly, as concluded in the gaze and head pose estimators benchmarking. The wrong skeletons detected by the Kinect are related with the incorrect distinction of the red scarf used by the therapist (segmentation failure) and with the lower Kinect rate for the calculation of 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, when 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.

- Improvements on the setup: to improve the Kinect performance in the school study, the therapist and child should be standing during the sessions, similarly to the pilot study. The system hyper-parameters showed to be very important to increase the system 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 in the school study sessions 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. Concerning the Total Fixation Duration 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 attract the children's attention and thus, when possible, the study should be conducted in a private space. They also affect the system performance, since they pass in the same azimuth angles as some of the targets. To solve this problem and distinguish between looking at elsewhere and to the targets, the elevation angle should be included in the proposed framework and analysed.

- Online adaptation of the attention system: the adjustment of the quantitative attention system 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, maximizing engagement and, 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. Thus, new quantitative systems could be created for the evaluation of the children facial expressions (affective) and of their performance in the imitation tasks (behavioral). In this way, it would be possible to observe the relation between the different types of engagement and have a complete overview of the impact of these type of therapies in children with ASD.

Bibliography

[1] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, pp. 276 – 282, 2012.

[2] R. Gelin, "NAO," in *Humanoid Robotics: A Reference* (A. Goswami and P. Vadakkepat, eds.), pp. 1–22, Springer Netherlands, 2018.

[3] F. Ahmed, P. P. Paul, and M. Gavrilova, "Kinect-based gait recognition using sequence of the most relevant joint relative angles," *Journal of WSCG*, vol. 23, pp. 147–156, 2015.

[4] B. Teke, M. Lanz, J. Kämäräinen, and A. Hietanen, "Real-time and robust collaborative robot motion control with Microsoft Kinect ® v2," *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*, pp. 1–6, 2018.

[5] 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.

[6] 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.

[7] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-time eye gaze estimation in natural environments," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 344–352, 2018.

[8] 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.

[9] T. Higuchi, Y. Ishizaki, A. Noritake, Y. Yanagimoto, H. Kobayashi, K. Nakamura, and K. Kaneko, "Spatiotemporal characteristics of gaze of children with autism spectrum disorders while looking at classroom scenes," *PLOS ONE*, vol. 12, no. 5, pp. 1–19, 2017.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] "Tobii eye tracking: An introduction to eye tracking and tobii eye trackers," in *Tobii Technology*, Whitepaper, 2010.

[17] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: Definition, epidemiology, causes, and clinical evaluation," *Translational pediatrics*, vol. 9 Suppl 1, pp. S55–S65, 2020.

[18] M. J. Maenner, K. A. Shaw, J. Baio, A. Washington, M. E. Patrick, M. Dirienzo, D. L. Christensen, L. D. Wiggins, S. Pettygrove, J. G. Andrews, M. Lopez, A. Hudson, T. Baroud, Y. Schwenk, T. White, C. R. Rosenberg, L. C. Lee, R. A. Harrington, M. Huston, A. S. Hewitt, A. N. Esler, J. A. Hall-Lande, J. N. Poynter, L. Hallas-Muchow, J. N. Constantino, R. T. Fitzgerald, W. M. Zahorodny, J. Shenouda, J. L. Daniels, Z. Warren, A. C. Vehorn, A. Salinas, M. S. Durkin, and P. Dietz, "Prevalence of autism spectrum disorder among children aged 8 years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016," *MMWR Surveillance Summaries*, vol. 69, no. 4, pp. 1 – 12, 2020.

[19] F. Marino, P. Chilà, S. T. Sfrazzetto, C. Carrozza, I. Crimi, C. Failla, M. Busà, G. Bernava, G. Tartarisco, D. Vagni, L. Ruta, and G. Pioggia, "Outcomes of a robot-assisted social-emotional understanding intervention for young children with autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 50, no. 6, pp. 1973–1987, 2020.

[20] 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)*, pp. 534–539, 2018.

[21] 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.

[22] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "The NAO humanoid: A combination of performance and affordability," *CoRR*, 2008.

[23] N. I. Ishak, H. Md. Yusof, S. N. Sidek, and N. Rusli, "Robot selection in robotic intervention for ASD children," in *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pp. 156–160, 2018.

[24] M. A. Saleh, F. A. Hanapiah, and H. Hashim, "Robot applications for autism: A comprehensive review," *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 6, pp. 580–602, 2021.

[25] 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.

[26] 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.

[27] H.-L. Cao, P. G. Esteban, M. Bartlett, P. Baxter, T. Belpaeme, E. Billing, H. Cai, M. Coeckelbergh, C. Costescu, D. David, A. De Beir, D. Hernandez, J. Kennedy, H. Liu, S. Matu, A. Mazel, A. Pandey, K. Richardson, E. Senft, S. Thill, G. Van de Perre, B. Vanderborght, D. Vernon, K. Wakanuma, H. Yu, X. Zhou, and T. Ziemke, "Robot-enhanced therapy: Development and validation of supervised autonomous robotic system for autism spectrum disorders therapy," *IEEE Robotics Automation Magazine*, vol. 26, no. 2, pp. 49–58, 2019.

[28] 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.

[29] J. Kang, R. Kim, H. Kim, Y. Kang, S. Hahn, Z. Fu, M. Khalid, E. Schenck, and T. Thesen, "Automated tracking and quantification of autistic behavioral symptoms using Microsoft Kinect," *Studies in health technology and informatics*, vol. 220, pp. 167–70, 2016.

[30] I. Budman, G. Meiri, M. Ilan, M. Faroy, A. Langer, D. Reboh, A. Michaelovski, H. Flusser, I. Menashe, O. Donchin, and I. Dinstein, "Quantifying the social symptoms of autism using motion capture," *Scientific Reports*, vol. 9, no. 1, p. 7712, 2019.

[31] S. Giancola, A. Corti, F. Molteni, and R. Sala, "Motion capture: An evaluation of Kinect v2 body tracking for upper limb motion analysis," in *MobiHealth*, pp. 302–309, 2016.

[32] E. Dell'Aquila, G. Maggi, D. Conti, and S. Rossi, "A preparatory study for measuring engagement in pediatric virtual and robotics rehabilitation settings," *HRI '20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, p. 183–185, 2020.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] Khan and Lee, “Gaze and eye tracking: Techniques and applications in ADAS,” *Sensors*, vol. 19, p. 5540, 2019.

[38] M. Borys and M. Plechawska-Wójcik, “Eye-tracking metrics in perception and visual attention research,” *European Journal of Medical Technologies*, vol. 3(16), pp. 11–23, 2017.

[39] M. Böhme, A. Meyer, T. Martinetz, and E. Barth, “Remote eye tracking: State of the art and directions for future development,” *2nd Conference on Communication by Gaze Interaction - COGAIN 2006: Gazing into the Future*, pp. 10–15, 2006.

[40] 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.

[41] M. Patacchiola and A. Cangelosi, “Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods,” *Pattern Recognition*, vol. 71, pp. 132–143, 2017.

[42] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” *arXiv preprint arXiv:1804.02767*, 2018.

[43] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 215500–215509, 2018.

[44] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97, pp. 6105–6114, 2019.

[45] N. W. Rim, K. W. Choe, C. Scrivner, and M. G. Berman, “Introducing Point-of-Interest as an alternative to Area-of-Interest for fixation duration analysis,” *PLOS ONE*, vol. 16, no. 5, pp. 1–18, 2021.

[46] M. Hochhauser, A. Aran, and O. Grynszpan, “Investigating attention in young adults with autism spectrum disorder (ASD) using change blindness and eye tracking,” *Research in Autism Spectrum Disorders*, vol. 84, p. 101771, 2021.

[47] T. Fujioka, K. J. Tsuchiya, M. Saito, Y. Hirano, M. Matsuo, M. Kikuchi, Y. Maegaki, D. Choi, S. Kato, T. Yoshida, Y. Yoshimura, S. Ooba, Y. Mizuno, S. Takiguchi, H. Matsuzaki, A. Tomoda, K. Shudo, M. Ninomiya, T. Katayama, and H. Kosaka, “Developmental changes in attention to social information from childhood to adolescence in autism spectrum disorders: A comparative study,” *Molecular Autism*, vol. 11, p. 24, 2020.

[48] W. Cao, W. Song, X. Li, S. Zheng, G. Zhang, Y. Wu, S. He, H. Zhu, and J. Chen, “Interaction with social robots: Improving gaze toward face but not necessarily joint attention in children with autism spectrum disorder,” *Frontiers in Psychology*, vol. 10, p. 1503, 2019.

[49] 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.

[50] 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)*, pp. 1–4, 2018.

[51] X. Xiong and F. De la Torre, “Supervised descent method and its applications to face alignment,” in *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 532–539, 2013.

[52] P. I. Wilson and J. Fernandez, “Facial feature detection using Haar classifiers,” *Journal of Computing Sciences in Colleges*, vol. 21, no. 4, p. 127–133, 2006.

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

[54] 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.

[55] T. N. Garlie and H. Choi, “Characterizing the size of the encumbered soldier,” *US: Army Natick Soldier Research, Development and Engineering Center (NSRDEC)*, 2014.

[56] 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.