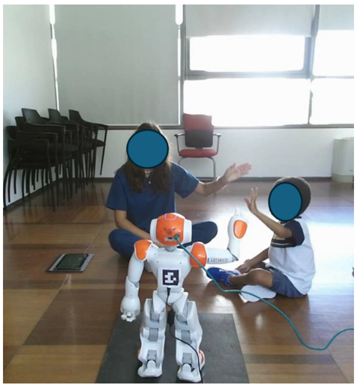
UNIVERSIDADE DE LISBOA

INSTITUTO SUPERIOR TÉCNICO

POLITECNICO DI MILANO



Analyse&Act: Automatic metrics and innovative protocols in robotics for children with ASD

Laura Joana Espinosa Fortes Ferreira dos Santos

Supervisors : Doctor José Alberto Rosado Santos-Victor
Doctor Alessandra Laura Giulia Pedrocchi

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering at IST/ULisboa and the PhD Degree in Bioengineering at Politecnico di Milano

Jury final classification: Pass with Distinction and Honour

2025

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

POLITECNICO DI MILANO

Analyse&Act: Automatic metrics and innovative protocols in robotics for children with ASD

Laura Joana Espinosa Fortes Ferreira dos Santos

Supervisors : Doctor José Alberto Rosado Santos-Victor
Doctor Alessandra Laura Giulia Pedrocchi

Thesis approved in public session to obtain the PhD Degree in Electrical and Computer Engineering at IST/ULisboa and the PhD in Bioengineering at Politecnico di Milano

Jury final classification: Pass with Distinction and Honour

Jury

Chairperson : Doctor Pedro Manuel Urbano de Almeida Lima, Instituto Superior Técnico, Universidade de Lisboa;

Members of the Committee :

Doctor Yukie Nagai, International Research Center for Neurointelligence, University of Tokyo, Japão;
Doctor Angelo Cangelosi, School of Engineering, University of Manchester, Reino Unido;
Doctor Alessandra Laura Giulia Pedrocchi, Politecnico di Milano, Itália;
Doctor Franca Garzotto, Politecnico di Milano, Itália;
Doctor José Alberto Rosado dos Santos Victor, Instituto Superior Técnico, Universidade de Lisboa;
Doctor João Miguel Raposo Sanches, Instituto Superior Técnico, Universidade de Lisboa.

Funding Institutions - Fundação para a Ciência e a Tecnologia and Fondazione Milano per Expo

2025

“Eu fui à terra do bravo,
Bravo, meu bem,
Para ver se embravecia.
Cada vez fiquei mais manso,
Bravo, meu bem,
Para a tua companhia.”

(Música tradicional dos Açores,
eternizada por Zeca Afonso,
harmonizada pelas Sopa de Pedra)

“I went to the land of the brave
Brave, my love,
To see if I would grow fierce.
Each time I became gentler,
Brave, my love,
For your company.”

(Traditional music from the Azores,
immortalized by Zeca Afonso,
harmonized by Sopa de Pedra.)

Acknowledgments

The PhD and the beginning of my research path took me to the ”land of the brave”. Along the way, I ”embraved(grew fierce)” and realized that the strength required for research comes primarily from the ”gentleness” developed in reviewing and understanding errors, results, and papers. This strength also depends on the ”company” we keep—those who accompany us during what appears to be a solitary period. The fact that the first two years of this thesis took place during a global pandemic made this ”company” even more valuable. The list of people to thank is long, as this project was carried out in two countries. However, expressing my gratitude to each one of them is essential because, without them, this entire document would be different.

First, I deeply thank my supervisors, Prof. José Santos-Victor and Prof. Alessandra Pedrocchi, for their invaluable lessons and firm commitment to this work. I am grateful for their constant encouragement to strive for excellence, their insistence on rigorous and well-conducted research, and the immense freedom they gave me to explore my ideas while guiding me whenever I needed.

The countless meetings, revisions, emails, and visits to clinical institutions were essential to this thesis. For that, I am eternally grateful. Phrases like ”This is not a race; it is a marathon. The struggle continues!” and ”The best is the enemy of the good” will continue to resonate in my mind, and I hope they will guide me throughout the rest of my research journey.

To my guardian angel, Alice Geminiani, I believe it would take several lifetimes to truly thank her—for all the revisions of my writing in its rough drafts, for her countless suggestions for future work, and for the many moments when I felt the ’land of the brave’ would swallow me, yet she always offered a simple solution, reassuring me that everything would be fine. A special thanks also to Catarina Barata for fully embracing this project and for the many discussions and ideas.

This thesis was funded by FCT Portuguese Foundation for Science and Technology (projects SFRH /BD/145040/2019 and C645008882-00000055 Center for Responsible AI) and Fondazione Milano per Expo/Fondazione Bracco, to whom I am deeply grateful. I also thank other key partners for their support: Associazione Paolo Zorzi (project *IOGIOCO*) and the European Union (for funding the MUSA project under the National Recovery and Resilience Plan Mission 4 Component 2 Investment Line 1.5).

In this project, three clinical institutions chose to take part, and they were fundamental to the impact of this thesis, as they played a important role in the results obtained. To the children, parents, and clinical staff who participated in this study, my gratitude is immense. More specifically, at Fondazione Don Gnocchi, I thank Dr. Ivana Olivieri for believing in me even when there was no formal project; Arianna Caglio for testing the protocols multiple times, contributing countless ideas through repeated trials, and spending many hours with me refining them; Dr. Silvia Annunziata for her valuable feedback on our results, as well as for recruiting and evaluating many children for our projects; and last but not least, Dr. Anna Cavallini, who brought new energy to this project, worked tirelessly with ethical committees, introduced innovative ideas, and, most importantly, led this project toward a randomized controlled trial.

At the Associação Portuguesa de Autismo (APPDA), I am especially grateful to Inês Neto for facilitating the integration of this project within the institution and to therapists Inês Tecedeiro and Sara Ferreira for helping us adapt the protocol to a school setting.

At the Centro de Apoio ao Desenvolvimento (CADIn), I deeply thank Ana Rita Gonzalez and Susana Lúcio for developing a new protocol and for managing the complex logistics of coordinating multiple therapists and children simultaneously for the study. Finally, I extend my gratitude to Dr. Bernardo Barahona-Corrêa for integrating this work into CADIn and for the discussions that enriched this project.

I deeply thank the people who have worked on this project and from whom I have learned so much about the importance of communication. First, to Gabriele, who made me see everything in a different light. Then, to all the other students with whom I have worked: Noa, Bárbara, Alessia, Alice, Daniela, Filippo, Martina, Catarina, Bernardo, and Mariana.

This PhD brought me new friendships that have deeply marked my journey over these five years. To Paul, I deeply thank you for all the advice for a cleaner code, for the emotional support during the journey through bugs and beyond, and most of all, for the philosophical discussions on robots, research, and life. To Hugo, I thank for the many aesthetic suggestions for papers and presentations, the continuous support during our intercontinental calls, and for developing an ethical perspective on the impact of coaching robots. The Nearlab and Vislab are two places where I grew considerably, and they have become true homes, especially because of the people I met there who made everything lighter during the darkest hours. To them, I express my deepest thanks: Akhil (and Anju), Dimitris, Atabak, Alessandra Trapani (and Gabriele and Tobe), Mattia, Francesco Sheiban, and Benedetta. To Alberto, Emilia, Plinio, Avelino, Heitor, Nuno, José Gaspar, and Alexandre Bernardino, I thank you for your invaluable feedback on my research and the insightful discussions about life after the PhD.

I am grateful to all my friends who have been a refuge for talking about everything in life and who, despite all my ”I can't” moments, continued to invite me: to the new friends from Conservatorio, who have become a safe haven of hugs at any time of day (Giorgia McKenzie, Tommaso, Maori, and Giorgia Ciriello); to my old Politecnico friends, with whom I continued discovering Italian culture, full of good food and laughs (Chiara, Adriano, Giando, Ale, Beatrice, Fede, Pana); to my international friends who were always there even though we no longer live in the same city (Alejandra, Ava, Alexandra, and Marine); to my friends from Técnico, whose friendships grew during this period (Inês, Miguel, Diogo, Manuel, André, Sofia); to my musician friends, who are an important part of who I am today and were always available for a chat (Maria, Mafalda, Mariana, Inês, Olívia, and Gameiro); and to the ones who have been listening to me about every single topic for years and choose to continue—Bárbara and Joana.

This paragraph is dedicated to those who don't know they had an enormous impact on the continuation and completion of this thesis. Marco Simonini, to whom I am grateful for his commitment to establishing the double degree program, maintaining correspondence, and advocating for various institutions, and who helped me navigate countless bureaucratic hurdles. To Ana Santos and Julia Oliveira, I thank you for your quick email responses that resolved several issues at Técnico. Lastly, I thank those whose simple everyday gestures were fundamental: Nati, the cleaning lady of building 32, who always started the day with a smile and gave me a boost of energy, and Letizia and the staff from the Union Bar who always made sure I had something to eat, even at 3 or 4 p.m.

My deepest gratitude goes to my family. To my sister, for her immense care, understanding, and constant belief in me. To my grandmother, the personification of the verb ”embrave”, for always transmitting her strength, for all the calls, and the countless frozen meals I brought to Milan. To my parents, who have lived in the ”land of the brave” as long as I can remember and have sparked my curiosity about it. To my mum, for her persistence in seeing this process through and her constant presence throughout it. To my father, for passing on his calmness, showing me that it is truly the key to thriving in the land of the brave. To both of them, I am forever grateful for all the opportunities they gave me, for their support, dedication, and, most importantly, their love.

Abstract

Robotics in autism spectrum disorder (ASD) already has a two-decade history. Some individuals with ASD are particularly drawn to robots due to their predictability and repeatability, showing improvements in social, communication and motor capabilities following robot-assisted therapies. However, the simplified external characteristics of robots often extend to their control systems, limiting their applicability in a disorder characterized by its heterogeneity. People with autism exhibit highly varied behaviours across different contexts, days, or even within the same day. Therefore, adaptable control systems are required, but they rely on quantitative measures, which are currently lacking in the autism field. This lack of quantitative measures has also hindered the ability to consolidate evidence from multiple clinical studies, as each relies on its own subjective measures. Consequently, robots are still far from being integrated into standard clinical practice.

The main goal of this thesis was to address this gap by developing and testing new protocols and quantitative measures for clinical use. Three sub-goals were established:(i) implementation of protocols that consider the heterogeneity of the disorder; (ii) construction of quantitative measures for evaluation intra and inter-sessions; (iii) design of experimental studies to provide evidence of the system's impact.

The contributions of this thesis concern both in new protocols and new quantitative measures. All developments were carried out in close collaboration with clinical institutions specialized in Autism in both Portugal and Italy, where this research took place. Four distinct robotic protocols were developed: one for pre- and post-treatment evaluation (scale protocol); another centred on the therapeutic improvement of the gestural and mirroring skills, including multiple levels to consider the heterogeneity of ASD (hierarchical protocol); and two additional protocols derived from tests conducted in Portugal and Italy using the hierarchical protocol. Interestingly, both of these adaptations evolved into single-level protocols incorporating holistic games to facilitate gesture presentation, making integration into clinical practice easier. In total, 33 children and 11 therapists have been involved in the test of the hierarchical protocol, a relatively large sample size in ASD research.

Cameras were used as sensors for quantitative measurement, as children with ASD often prefer non-intrusive devices due to their sensitivity to touch. We developed a new mirroring measure that considered the duration and latency at start of exercises captured differences between neurotypical and ASD children. Additionally, a gesture recognition system was created by combining a kinetic parameter - used to determine the beginning and end of a gesture- with a ResNet architecture, for the classification of the gesture. The integration of this system into one of the protocols led to increased engagement in the child with ASD that tested it.

Given the significance of attention in ASD, a neural network was developed for automatic attention classification following the assessment of the scale protocol. Furthermore, an attention biomarker was introduce for use during the therapy sessions, measuring fixation time on different targets of interest using geometrical assumptions. Not only did this system outperform existing methods, but it also revealed a correlation between the obtained fixation time and one of the clinical scales used in our study - an important step toward establishing evidence for the potential impact of robotic therapies in the ASD field.

Future work will focus on integrating these metrics in real-time into our robotic protocols to create a personalized system, enhancing engagement in long-term therapy sessions. In addition, we aim to increase parental involvement to facilitate the extension of these therapies into home settings.

Keywords: Autism Spectrum Disorder, quantitative measures, gesture recognition, attention classification, adaptive robotics

Resumo

A robótica na perturbação do espectro do autismo (PEA) já tem uma história de duas décadas. Alguns indivíduos com PEA são atraídos por robôs devido à sua previsibilidade e repetibilidade, mostrando melhorias nas capacidades sociais e motoras após terapias assistidas por robôs. No entanto, as características externas simplificadas dos robôs estendem-se aos seus sistemas de controlo, limitando a sua aplicabilidade numa perturbação caracterizada pela heterogeneidade. As pessoas com PEA exibem comportamentos muito variáveis, sendo necessários sistemas de controlo adaptativos. Infelizmente, estes dependem de medidas quantitativas, que escaceiam nesta área. Esta carência também dificultou a consolidação das evidências de múltiplos estudos clínicos, já que cada um depende das suas próprias medidas subjectivas. Consequentemente, os robôs não estão ainda integrados na prática clínica.

O principal objectivo desta tese foi lidar com esta lacuna, desenvolvendo e testando novos protocolos e medidas quantitativas para uso clínico. Foram estabelecidos três sub-objectivos:(i) implementação de protocolos sensíveis à heterogeneidade do distúrbio;(ii) construção de medidas quantitativas para avaliação intra e inter-sessões;(iii) concepção de estudos experimentais para fornecer evidências do impacto do sistema.

Todos os desenvolvimentos foram realizados em colaboração com instituições clínicas especializadas em autismo onde esta investigação se realizou, em Portugal e em Itália. Foram desenvolvidos quatro protocolos robóticos distintos: um para avaliação pré e pós-tratamento (protocolo da escala); outro centrado na melhoria terapêutica das capacidades gestuais e de imitação, incluindo múltiplos níveis para considerar a heterogeneidade da PEA (protocolo hierárquico); e outros dois protocolos derivados de testes realizados em Portugal e Itália usando o protocolo hierárquico. Ambas as derivações evoluíram para protocolos de nível único, incorporando jogos holísticos para facilitar a apresentação de gestos e facilitar a integração na prática clínica. No total, 33 crianças com PEA e 11 terapeutas participaram no teste do protocolo hierárquico, uma amostra considerável nesta área de investigação.

Para a medição quantitativa, câmaras foram utilizadas como sensores pois crianças com PEA frequentemente preferem dispositivos não intrusivos, devido à sua sensibilidade ao toque. Desenvolvemos uma nova medida de espelhamento que considerou a duração e a latência no início dos exercícios, capturando as diferenças entre crianças neurotípicas e com PEA. Além disso, foi criado um sistema de reconhecimento de gestos, combinando um parâmetro cinético - usado para determinar o início e o fim de um gesto - com uma arquitectura ResNet para a classificação do gesto. A integração deste sistema num dos protocolos levou a um aumento do envolvimento da criança com PEA que o testou.

Dada a importância da atenção na PEA, foi desenvolvida uma rede neuronal para a classificação automática da atenção após a avaliação do protocolo da escala. Além disso, foi introduzido um biomarcador de atenção para uso durante as sessões de terapia, medindo o tempo de fixação em diferentes alvos de interesse, utilizando pressupostos geométricos. Este sistema não só superou os métodos existentes, como também revelou uma correlação entre o tempo de fixação obtido e uma das escalas clínicas utilizadas no nosso estudo - um passo importante para estabelecer evidências sobre o impacto potencial das terapias robóticas no campo da PEA.

O trabalho futuro irá focar-se na integração destas métricas nos nossos protocolos robóticos em tempo real, a fim de criar um sistema personalizado, aumentando o envolvimento nas sessões terapêuticas a longo prazo. Além disso, pretendemos aumentar o envolvimento dos pais para facilitar a extensão destas terapias para ambientes domiciliários.

Palavras-chave: Perturbação do Espectro do Autismo, medidas quantitativas, reconhecimento de gestos, classificação da atenção, robótica adaptativa

Sommario

La robotica nel disturbo dello spettro autistico (ASD) ha già una storia di due decenni. Alcuni individui con ASD sono Interagiscono facilmente con i robot a causa della loro prevedibilità e ripetibilità, mostrando miglioramenti nelle capacità sociali e motorie dopo terapie assistite da robot. Tuttavia, le caratteristiche esterne semplificate dei robot spesso si estendono ai loro sistemi di controllo, limitando la loro applicabilità in un disturbo caratterizzato dalla sua eterogeneità. Le persone con autismo presentano comportamenti altamente variabili, pertanto, sono necessari sistemi di controllo adattativi. Purtroppo, questi si basano su misure quantitative, che attualmente mancano in questo campo. Questa mancanza ha anche ostacolato la possibilità di consolidare evidenze da diversi studi clinici, poiché ciascuno si basa sulle proprie misure soggettive. Di conseguenza, i robot non sono ancora integrati nella pratica clinica.

L'obiettivo principale di questa tesi era colmare questa lacuna sviluppando e testando nuovi protocolli e misure quantitative per l'uso di robot nella terapia dei disturbi nello spettro dell'autismo. Sono stati stabiliti tre sotto-obiettivi: (i) implementazione di protocolli che considerino l'eterogeneità del disturbo; (ii) costruzione di misure quantitative per la valutazione intra e inter-sessioni; (iii) progettazione di studi sperimentali per fornire prove dell'impatto del sistema.

Tutti gli sviluppi sono stati realizzati in stretta collaborazione con istituzioni cliniche specializzate in autismo sia in Portogallo che in Italia, dove è stata condotta questa ricerca. Sono stati sviluppati quattro distinti protocolli robotici: uno per la valutazione pre e post-trattamento (protocollo della scala); un altro centrato sul miglioramento terapeutico delle abilità gestuali e di imitazione, con più livelli per considerare l'eterogeneità dell'ASD (protocollo gerarchico); e due protocolli aggiuntivi derivati da test condotti in Portogallo e Italia utilizzando il protocollo gerarchico. Entrambi questi sviluppi si sono evoluti in protocolli a singolo livello che incorporano giochi olistici per facilitare la presentazione dei gesti, rendendo più facile l'integrazione nella pratica clinica. In totale, 33 bambini e 11 terapisti sono stati coinvolti nel test del protocollo gerarchico, un campione relativamente ampio nella ricerca sull'ASD.

Per la misurazione quantitativa, sono state utilizzate telecamere come sensori poiché i bambini con ASD spesso preferiscono dispositivi non invasivi a causa della loro sensibilità al tatto. Abbiamo sviluppato una nuova misura di *mirroring* che considerava la durata e la latenza all'inizio degli esercizi, rilevando differenze tra bambini neurotipici e bambini con ASD. Inoltre, è stato creato un sistema di riconoscimento dei gesti combinando un parametro cinetico - utilizzato per determinare l'inizio e la fine di un gesto - con un'architettura ResNet, per la classificazione del gesto. L'integrazione di questo sistema in uno dei protocolli ha portato a un maggiore coinvolgimento nel bambino con ASD che lo ha testato.

Data l'importanza dell'attenzione nell'ASD, è stata sviluppata una rete neurale per la classificazione automatica dell'attenzione dopo la valutazione del protocollo di scala. Inoltre, è stato introdotto un biomarcatore dell'attenzione per l'uso durante le sessioni di terapia, misurando il tempo di fissazione su diversi obiettivi di interesse utilizzando assunzioni geometriche. Non solo questo sistema ha superato i metodi esistenti, ma ha anche rivelato una correlazione tra il tempo di fissazione ottenuto e una delle scale cliniche utilizzate nel nostro studio - un passo importante verso la creazione di prove per il potenziale impatto delle terapie robotiche nel campo dell'ASD.

Il lavoro futuro si concentrerà sull'integrazione di queste metriche in tempo reale nei nostri protocolli robotici per creare un sistema personalizzato, migliorando il coinvolgimento nelle sessioni di terapia a lungo termine. Inoltre, intendiamo aumentare il coinvolgimento dei genitori per facilitare l'estensione di queste terapie in ambito domestico.

Parole chiave: Disturbo dello Spettro Autistico, misure quantitative, riconoscimento di gesti, classificazione dell'attenzione, robotica adattiva

Contents

List of Tables **xiii**

List of Figures **xix**

Acronyms **xxv**

1 Introduction **1**

1.1 Problem statement 2

1.2 Objectives 3

1.3 Summary of contributions 5

1.4 List of publications 5

2 Background and Literature Review **9**

2.1 Autism Spectrum Disorder 9

2.1.1 A multifactorial disorder 10

2.1.2 The therapies 12

2.2 Social Robots in Autism 12

2.2.1 Setup and Protocol Design 13

2.2.2 Performance Evaluation Measures 15

2.2.3 Randomised Controlled Trials 16

Meta-analysis 19

2.2.4 Technical limitations and clinical domain constraints 20

2.3 Technical Challenges 21

2.3.1 Pose Measurement Methods 22

2.3.2 Action Analysis from Video 25

2.3.3 Gaze Measurement Methods 26

2.4 Conclusions 28

3 Clinical Protocols **29**

3.1 The basic mirroring protocol 29

3.2 The hierarchical protocol 31

3.2.1 The five levels 32

3.3 The sensorial protocol 33

Luminous stimuli 34

Auditory stimuli 34

Movement stimuli 34

3.4 The scale protocol 35

3.5 The bingo protocol 36

3.6 Conclusions 37

4 Action Analysis from video 39

4.1 Mirroring measure 39

4.1.1 *Robot Coach* Protocol 39

4.1.2 *Adult Coach* Protocol 40

4.1.3 Results 42

Robot Coach Protocol 42

Adult Coach Protocol 43

4.1.4 Discussion 44

4.2 Gesture recognition system 44

4.2.1 Activity Detection 45

4.2.2 Activity recognition 45

4.2.3 Results 46

Activity Recognition performance 47

Comparison of the Activity Detection methods 48

Segmentation Quality by human expert validation 49

Real-time Online Recognition during Therapy 51

4.2.4 Discussion 51

4.3 Conclusion 51

5 Attention biomarker 53

5.1 Attention classification system 53

5.1.1 Benchmark of the 360º Estimators 54

5.1.2 Gaze Estimation and Data Processing 56

5.1.3 Areas-of-Interest construction 57

5.1.4 Attention classification 59

5.1.5 Results 59

Data division and hyperparameters tuning 60

Attention classification and comparison with the Therapist feedback 61

5.1.6 Discussion 63

5.2 Online attention classifier 63

5.2.1 Results 64

5.2.2 Discussion 66

5.3 Neural Network for attention classification 66

5.3.1 Results 68

5.3.2 Discussion 70

5.4 Conclusions 70

6 3D body pose estimation 71

6.1 Offline version 71

6.1.1 Focal length estimation 71

6.1.2 Controlled Experimental Setting 72

6.1.3 Real World Setting 73

6.2 Online version 73

6.2.1 Evaluation experiments 74

6.3 Results 74

6.3.1 Offline version 74

6.3.2 Online version . . . 78

6.4 Discussion and conclusions . . . 79

7 Clinical Studies 81

7.1 A Pilot Study . . . 81

7.2 A Pilot Randomized Controlled Trial . . . 86

7.2.1 Generalizability of the model . . . 86

7.2.2 Comparison between clinical and automatic measures of attention . . . 87

7.3 Discussion . . . 89

8 Conclusions 91

8.1 Future Work . . . 92

8.2 Final considerations . . . 94

Bibliography 95

List of Tables

2.1 Outcome Measures of Randomized Controlled Trials - Part I 17

2.2 Outcome Measures of Randomized Controlled Trials - Part II 18

3.1 Target poses and ranges of variability (α) used in the feedback system of the *Robot Coach* protocol. 31

4.1 Control angles of each exercise of the *Robot Coach* protocol. 40

4.2 Latency times of the Therapist and ASD child and mean latencies of the neurotypical adults and children for the exercises in the *Robot Coach* protocol. The zero values represent situations in which the subject was already in the starting position at the onset of the analysis. 42

4.3 First positive feedback times (FPFT) and Number of attempts until the first feedback of the Therapist, ASD child, neurotypical adults and children for the exercises in the *Robot Coach* protocol. 42

4.4 Duration of successful trials and number of failures for the Therapist, ASD child, neurotypical adults and children in the exercises of the *Robot Coach* protocol. The empty cells are exercises not completed in the preset time. 43

4.5 Movement variability: *Intraset distance* in the children set, adults set, therapist and ASD child sets for each sport of the *Adult Coach* protocol. 43

4.6 Repeatability among pairs: Mean of the *intraset distances* for neurotypical subjects and *intraset distance* for ASD child-therapist pair for each type of exercise in the *Adult Coach* protocol. 43

4.7 Mirroringability: For each sport in the *Adult Coach* protocol, *intraset distances* of the pair child-respective adult (left column) compared to the mean of *intraset distances* for the pairs child-other adults (right column). The coloured cells mark the lower distances, being green when verified by the couple child-respective adult (mirroring verified), and red, otherwise. 44

4.8 Index of the keypoints associated to each different body segment. For index meaning please refer to Figure 2.8. 46

4.9 Mean prediction probabilities in output from the Softmax layer of ten gestural trials of neurotypical subjects. 47

4.10 Comparison between results on ASD children obtained with two segmentation methods: with Sliding Window[1] and with the Kinetic Parameter. Bold indicates the main measure analysed, F1-score, since the dataset is unbalanced. 49

4.11 Mean delays between expected and predicted starting and stopping instants for the gestures classified as correct and incorrect by the blinded assessor. 51

5.1 Average RMSE of WHENet and Gaze360 azimuth estimates (rad). Experiment 3 focused only on the movement of the eyes and it was the experiment in which Gaze360 presented better results. Bold indicates the lowest RMSE per experiment. In Experiment 2, no value is marked due to their similarity. 56

5.2 Percentage of lost data for the children who attended that therapy session: with and without data interpolation. The red cells represent the session in which more than 2/3 of the data was lost (%). 56

5.3 Attention classification system's accuracy for the different hyperparameters configurations in the validation set (%). Bold indicates the highest accuracy for each approach. 61

5.4 Accuracy of our attention classification system using the best hyperparameters configuration (%). Session 6 was the validation set for both approaches. Session 3 was the training set on the learning approach, but part of the test set on the geometrical. All remaining sessions were part of the test set for both approaches. Bold indicates the highest accuracy for each session. 62

5.5 Our system attention classification accuracy for each child in each session and therapist's qualitative analysis. Green: Good system accuracy ($> 85\%$) or positive therapist feedback; Yellow: Average system accuracy ($80\% - 85\%$) or neutral therapist feedback; Red: Low system accuracy ($< 80\%$) or negative therapist feedback; P.: Performance; T.: Likes to touch robot (for our attention system, it was a negative characteristic, although it is a positive protocol remark, since the children show interest in the robot). 62

5.6 RMSE of short distance validation process (pixel) for the three face detectors tested and comparison with the gold standard Tobii T60. 64

5.7 RMSE (in degrees) and Face detection time (in seconds) of long-distance acquisitions for the three face detectors tested and comparison with the gold standard Tobii T60. 65

5.8 Inter-rater reliability - Cohen's kappa coefficient. NA represents the session which was not labelled by the second rater. 66

5.9 3D estimation accuracy overview for the two sessions (S1 and S2) of the two children. T represents the therapist, C, the child and R, the robot. The arrow indicates the object of interest, namely T ->C represents the Therapist looking to the Child. 66

5.10 Accuracy and precision of the original algorithm and the original algorithm with the elevation computation. 69

5.11 Accuracy and precision of explicit method solution. 69

5.12 Accuracy and precision obtained for the Multilayer Perceptron (MLP) architecture and for the combination of the Multilayer Perceptron with the Convolutional Neural Network (CNN) architecture. In the $Subject$ column, additional information about the gender is provided: male (M) and female (F). 69

6.1 Results from the RANSAC model using two different loss functions (sum of squares or weighted sum of squares). Each row corresponds to a linear model ($f = Slope \times height + Intercept$) generated by the RANSAC model with a coefficient of determination R^2. The selected model is underlined in green. 76

6.2 Height and Focal Lengths of participants in the controlled scenario. 76

6.3 3D RMSE results and processing rate for the CRMH-p and BEV model for each participant in the depth $1.9m < z < 2.5m$ corresponding with the region where therapies take place. The best results for each height are marked in bold. 76

6.4 Focal lengths estimations for the children and adults analyzed in the clinical sessions. The names of the children and therapists are represented by a code for anonymization purposes. 76

6.5 Percentage of ground truth skeletons in which both skeletons were detected by the Kinect and the proposed model. Each row represents one session done by a therapist (characters' code) and a child (numerical code). The best results for each pair (therapist-child) are marked in bold. 77

6.6 Mean Dice similarity between the skeletons of the ground truth and the ones of the analysed system (Kinect or CRMH-p) with the centroid correction. Each row represents one session done by a therapist (characters' code) and a child (numerical code). The best results for each child and each therapist are marked in bold. 77

6.7 Mean Dice similarity for the skeletons identified by the proposed model (CRMH-p) but not by Kinect. Each row represents one session done by a therapist (characters' code) and a child (numerical code). 78

6.8 Comparative Frames per second (FPS) analysis between sequential and parallel processing using RabbitMQ communication system. 78

6.9 Comparison of communication, processing, and total delay times across sequential and parallel systems using RabbitMQ communication system. 78

6.10 Comparative Frames per second (FPS) analysis between RabbitMQ and Mounted Directories techniques using sequential processing 79

6.11 Comparison of communication, processing, and total delay times between RabbitMQ and Mounted Directories techniques using sequential processing. 79

6.12 Frames per second (FPS) analysis of the full system using sequential processing and RabbitMQ communication technique. 79

6.13 Communication, processing, and total delay times for the full system using sequential processing and RabbitMQ communication. 80

7.1 Principal characteristics of the children who arrived to the end of the study. 86

7.2 Accuracy of our attention classification system in the CADIn study using the widths of the APPDA study and the widths obtained from the learning approach. Bold indicates the highest accuracy for each session. 87

7.3 Median[1st quartile; 3rd quartile] of the Total Fixation Duration towards the three areas of interest (NAO, therapist and elsewhere) for the control and experimental group and respective p-values. Bold indicates the significant p-values. 88

7.4 Median[1st quartile; 3rd quartile] of the LUI scale for the control and experimental group and respective p-values. 88

List of Figures

1.1 Design cycle of the general approach of the thesis. 4

2.1 Early Social Communication Scale (ESCS) setup 11

2.2 Distribution of the papers about robots in autism across the years. 13

2.3 Humanoid robots used in the literature. 13

2.4 Types of controllers reported in the analysed studies. 14

2.5 Number of sessions reported in the analysed studies. 15

2.6 Number of participants reported in the analysed studies. 15

2.7 Meta-analysis with the Randomized Controlled Trials selected. In a), it is presented the analysis of the ICF criteria d720 - Complex interpersonal interactions and in b) the analysis of the ICF criteria d160. 21

2.8 Skeleton format of the 25 joints Microsoft Kinect identifies. 23

2.9 Block diagram illustrating the CRMH model process. The model takes an RGB image as input, and then detects individuals, providing bounding boxes around them. Following this, each individual is reconstructed through their skeletons, utilizing an additional input of focal length and the SMPL model. Ultimately, the process culminates in a 3D scene, delivering an accurate mesh representation of the individuals [2]. 24

2.10 Overview of the BEV framework (extracted from [3]). Given an RGB image, BEV first estimates the 3D translation of all people in the scene via compositing the front view and the bird's-eye-view predictions. Then guided by the 3D translation, we sample the mesh feature of each person to regress their age-aware SMPL+A parameters [3]. 25

2.11 Full architecture of Gaze360. Each image passes through a backbone network (ImageNet - pretrained ResNet18), which produces a representation that goes to a bidirectional Long Short-Term Memory network. This network outputs a gaze direction and a quartile that represents the confidence in the gaze direction measurement [4]. 27

3.1 Diagram of the triadic interaction (adult-child-robot), illustrating the geometry between the participants and the perception system (Kinect) used during a protocol session (a) and block diagrams of the two games protocols. (b) *Robot Coach* protocol: first the NAO shows the exercise, then the adult repeats it while the robot is mirroring, and finally the child does the exercise, mirrored by the robot. (c) *Adult Coach* protocol: the adult leads the game, showing the exercise, which is then repeated by the robot and finally by the child while the robot is mirroring. Adapted from [5]. 30

3.2 Robot NAO performing the 19 gestures of the robotic-assisted therapy 32

3.3 Total body gestures included in the Hierarchical Protocol. 32

3.4 Example of dialogue in Level 4 (N: NAO; T: Therapist; C: Child). 33

3.5 The sensorial protocol with its three parts: luminous stimuli, auditory stimuli and movement stimuli. 34

3.6 The scale protocol setup. The orange star represents the mechanical toy; DX (right), SX (left) and Behind are the three posters on the wall. The blue x represents the therapist, while the green one, the child. 35
3.7 The scale protocol's overview 36
3.8 Block Diagram illustrating the flow and structure of the bingo protocol, emphasizing the key gestures and stages of interaction. 36
3.9 Illustrative Paper with Cartoon Theme. This figure showcases the selection of a theme suitable for the child in the Bingo Robotic Game. The therapist has laid out an illustrative paper on the ground, featuring engaging cartoons that correspond to the chosen theme, which could be food, animals, or cartoon characters. 37
3.10 Example of dialogue in Bingo robot (N: NAO; T: Therapist; C: Child). 37

4.1 Example of a set of observations using (a) single subject analysis and (b) paired analysis. In (a), circles and stars indicate the data points for one example adult and child, respectively; in (b) triangles indicate data points for one example pair. 41
4.2 Main steps of the data processing starting from raw keypoints to obtain an RGB pose feature image. In particular, N refers to the number of frames and K is the number of keypoints considered (23 in this work) [6]. 46
4.3 Block diagram of the online activity recognition system. After extracting the keypoints between the start and end points, the RGB pose feature images are generated and passed through the ResNet architecture. If the prediction probability exceeds a threshold, the probabilities from the last softmax layer are stored in a buffer. Once the buffer is full, if the gesture with the highest probability matches the one presented by the therapist, the robot provides feedback; otherwise, a new frame is added. 47
4.4 Confusion matrix obtained by applying the activity recognition model on the Neurotypical dataset. These results were good since the accuracy of 96% was much larger than the chance level (1/17 gestures × 100=6%). 48
4.5 Confusion matrix obtained with the complete online method with 3 neurotypical subjects and 8 repetitions.The small decrease of accuracy in relation to the offline method was expected since it depends on the activity detection part. Nevertheless the accuracy 92% is still above of the chance level (6%). 49
4.6 Confusion Matrices of the online application on the *Retrospective ASD children dataset* without (a) and with (b) the Kinetic Parameter algorithm. The improvement in terms of the accuracy is noticed by the reduction of the number of misclassified gestures. 50

5.1 Full overview of the attention classification system. The system is composed of two main blocks: the attention angle extraction and the Areas of Interest (AOIs) definition (Sections 5.1.2 and 5.1.3). After these two blocks, we compared the extracted angle with the defined AOIs to classify the attention (Section 5.1.4) into one of four targets: NAO robot, Other person, Computer, Elsewhere. For the choice of the attention estimator, we did an initial benchmark of gaze and head pose estimators (Section 5.1.1). For the definition of the AOIs, an initial data preprocessing (Section 5.1.2), followed by the analysis of the scene geometry (Section 5.1.3) was required. 54
5.2 Long Distance Benchmarking Setup: Experiment 1. The red crosses represent the 4 fixation points and the blue square represents the subject. 55
5.3 Long Distance Benchmarking Setup: Experiment 2. The red crosses represent the 4 fixation points and the blue square represents the subject. 55

5.4 Long Distance Benchmarking Setup: Experiment 3. The red crosses represent the 3 fixation points and the blue square represents the subject. 56

5.5 Standard angle (α_{target}) representation. The green cross represents the target, while the blue square represents the person from whom the standard angle is calculated. The origin of the reference frame is located in the centre of the Kinect and the coordinates of the person and the target are represented by (x, y) and (x_{target}, y_{target}), respectively. The variables x_{diff} and z_{diff} denote the differences between the 2D coordinates of the person and the target. 57

5.6 Representative top view of an Area-of-interest (AOI). The green cross represents the target, while the blue square represents the analysed person. The dark green line corresponds to the AOI width and the area in blue to the range of angles (receptive field) for looking at the target. 58

5.7 Example of decision boundary at a given frame for two overlapping Areas of Interest (AOIs) (computer and therapist). After establishing the probability density function (Pdf) for each AOI, the decision boundary is decided based on the intersection of the two gaussian curves. 59

5.8 Setup representation with the therapist (green triangle), the ASD child (blue square), NAO (red circumference), the computer (black square) and Kinect (black circumference). 60

5.9 Fixation maps of the therapist (a) and the children (b) gaze, considering the fixations of all sessions and all children. The red vertical lines represent the positions of the different targets. 62

5.10 Total Fixation Duration towards the targets and elsewhere along the sessions for Children (a) 9, (b) 10, (c) 15, (d) 19, (e) 6 (%). In Session 3, Level 3 was performed. In the remaining sessions, Level 4 was performed by all children except Child 6 which repeated Level 3. 63

5.11 Benchmark setups for validating different face detection algorithms. (a) and (b) illustrate the short-distance validation process using the Tobii T60, where the points in (b) appeared sequentially as indicated by the numbers on the right, starting and ending at the same point. (c) and (d) depict the long-distance validation process, where the blue crosses represent the points the subject focused on. 65

5.12 Example of offset between predicted Gaze360 and standard angle for Kinect camera in azimuth direction 67

5.13 Architecture of CNN model. 68

5.14 Architecture of MLP model. 68

6.1 Skeletons' scheme of the systems. The correspondences between joints used to evaluate are marked by circles of the same colour. Since the systems' skeletons are not a perfect match, for some joints (for example 14 in (a)) we had to use the mean of two associated markers. 73

6.2 Architectures tested for evaluation of the performance and decision on the integration in a real-time system: (a) Sequential processing using RabbitMQ system; (b) Sequential processing using Mounted Directories; (c) Parallel processing using RabbitMQ system. In all cases, the goal was to transmit frames captured by the Kinect to Docker for processing with the CRMH-p algorithm to extract the participants' skeletons. 75

6.3 Depth values for the child's hip joint using different systems: CRMH, BEV, CRMH-p and Optitrack. The proposed model improved the performance of the original CRMH. In the middle depth range ($1.9m < z < 2.5m$), the accuracy of the proposed model is notably high. 77

7.1 ABAS-II Scores at different time points [T0 (baseline)-T1 (end of the intervention protocol)-T2 (6 months after the end)] for the (a) Conceptual Adaptive Domain (CAD), (b) Social Adaptive Domain (SAD), (c) Practical Adaptive Domain (PAD), (d) General Adaptive Composite Score (GAC). 82

7.2 Griffith-III scores at different time points (T0-T1-T2) for (a) the Communication Scale (B scale), (b) the General Quotient and (c) the number of Gestures performed in the MacArthur's scale. 83

7.3 ABAS-II Scores for the familiarization and the training groups at different time points (T0-T1-T2) for the (a,b) Conceptual Adaptive Domain (CAD), (c,d) Social Adaptive Domain (SAD), (e,f) Practical Adaptive Domain (PAD), (g,h) General Adaptive Composite Score (GAC). The red asterisks represent the outliers. 84

7.4 Griffith-III scores for the familiarization and the training groups at different time points (T0-T1-T2) for (a,b) the Communication Scale (B scale), (c,d) the General Quotient and (e,f) the number of Gestures performed in the MacArthur's scale. 85

7.5 Boxplot of LUI scale at T0 and T1 for (a) the control and (b) the robotic group. 88

7.6 Boxplot of Total Fixation Duration (TFD) towards Elsewhere at T0 and T1 for (a) the control and (b) the robotic group. 89

7.7 Evolution of the LUI considering the median TFD when looking at elsewhere for both the control and robotic groups. 89

Acronyms

AOI Area-of-interest.

APPDA Associação Portuguesa para o Autismo e as Perturbações do Desenvolvimento.

ASD Autism Spectrum Disorder.

BEV Bird-Eye View.

CADIn Centro de Apoio ao Neurodesenvolvimento.

CNN Convolutional Neural Network.

CRMH Coherent Reconstruction of Multiple Humans.

CRMH-p CRMH-personalized.

DSM-V Diagnostic and Statistical Manual of Mental Disorders-5th edition.

ESCS Early Social Communication Scale.

FDG Fondazione Don Carlo Gnocchi.

FPS Frames per second.

LUI Language Use Inventory.

MLP Multi-Layer Perceptron.

RCTs Randomized Controlled Trials.

ResNet Residual Neural Networks.

RMSE Root mean squared error.

SARs Social Assistive Robots.

SMPL Skinned Multi-Person Linear Model.

TD Typical Development.

TFD Total Fixation Duration.

Chapter 1

Introduction

Robots have demonstrated their potential in diverse clinical settings. The two most common settings are the surgical theatre and the rehabilitation unit. In the first scenario, robots assist in performing surgical procedures, with roles that range from instrument control to automated surgical table. In the second scenario, robots physically assist or aid patients to achieve their goals [7]. From this second scenario, another setting can be derived, where robots provide assistance to users through social rather than physical interaction [8]. These robots, known as Social Assistive Robots (SARs), operate in less controlled and more unpredictable environments compared to their counterparts in surgical and rehabilitation settings. One of the earliest applications of SARs is in autism therapy, a field that exemplifies these unique characteristics [8].

Autism Spectrum Disorder (ASD) is a neurodevelopment disorder whose prevalence has increased in the last years. In Europe, the estimated prevalence from population studies among children with 5-18 years old was 0.9% in 1990s and raised to 1.4% in the studies published between 2015 and 2020 [9]. Part of this prevalence increment is justified by the broadening of the diagnostic criteria. In the first description of the disorder, in the third edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-III), ASD was characterized by “perversive lack of responsiveness to other people” and “gross deficits in language development”. Nowadays, according to the fifth edition of Diagnostic and Statistical Manual of Mental Disorders (DSM-V), ASD has two main diagnostic criteria: the presence of deficits in social communication and social interaction or restrictive and repetitive behavioural patterns [10]. Overall, tracking the prevalence of ASD has been a very challenging task not just because of the changing diagnostic criteria, but also due to the lack of biological diagnostic markers and the heterogeneity of the symptoms in the different children[11].

The heterogeneity intrinsic to the disorder calls for multidisciplinary and personalised treatments with no ’one-fits-all’ solution. Earlier treatment leads to a better prognosis [8]. Initial signs of autism during the first two years of life include no response to name when called, no or limited use of gestures in communication, and lack of imaginative play [12]. These signs prompt paediatricians to refer patients to neurologists and psychiatrists for comprehensive evaluations.

Some of the initial signs of ASD prolong through life namely the lack of gestural production, especially intransitive gestures. This type of gestures convey socio-communicative intent, and their recognition is particularly difficult for children with ASD compared to gestures involving objects [13]. In addition, these children present deficits in motor skills like imitation. For a certain mirroring task, they can achieve the goal of the task but with different kinematics from the subject who was the demonstrator [14]. [15] showed that these skills are related to some communication skills. In a more overall view, [16] demonstrated that the severity of motor impairments is directly correlated with impairments in social and communication skills. One of the skills impaired is joint attention [15].

Joint attention refers to the ability to share attention between a person and a social partner on an aspect of the environment (object or people) through eye-gazing, pointing or other verbal or non-verbal indications [17]. A clinical scale that measures this skill is the Early Social Communication Scale (ESCS). This scale is characterised by 17 semi-structured situations and encourages the interactions between the tester and the child. Through a videotape, an operator observes if the child shows any joint attention behaviour. The scoring of the scale is attributed according to the frequency of those behaviours.

Children with autism usually show deficits in joint attention, producing less declarative pointing gestures than typically developing children [18]. Children with ASD also present some difficulties in other types of attention such as social attention, defined by the focus given to a social stimulus and on-task attention, which is the focus given to a target object being essential for the learning of new skills[19].

Children with ASD have exhibited distinctive behaviours, such as novel joint attention and spontaneous mirroring, during interactions with SARs. Researchers attribute this to the fact that robots occupy a middle ground between non-social inanimate objects and highly complex social beings, which can sometimes overwhelm or confuse children with ASD [8]. SARs have been employed across a range of therapeutic applications, including improving social skills, emotional regulation, and motor training. These interventions are collectively known as robotic-assisted therapies.

1.1 Problem statement

The introduction of new technologies in the autism therapy field is particularly challenging, due to several reasons caused primarily by the children's heterogeneity in terms of capabilities and behaviours. This heterogeneity occurs not just between children but also between sessions of the same child. In order to reduce this heterogeneity, most studies in robotics for autism tend to focus on a subset of clinical conditions, which leads to several pilot studies. However, there are relatively few Randomized Controlled Trials (RCTs), which are considered the gold standard in clinical trial design [20].

In RCTs, children with ASD are randomly assigned to one of two groups: an experimental or robotic group, which participates in the new robotic therapy, and a control group, which usually continues standard care without the robotic intervention. The control group helps to determine how children would progress without the new therapy, accounting for confounding factors that might influence outcomes. This design minimizes differences between groups, ensuring the most reliable evidence of the intervention's impact [20].

The lack of these studies in the autism field is associated with a consequent low level of clinical evidence of the effectiveness of the robotic therapy approaches. This clinical evidence is required and essential for these technologies' introduction in clinical practice.

However, the heterogeneity reported previously is responsible for several contradictory results during RCTs: while some children improve, others do not. In this way, RCTs in robotic-assisted therapies often show no impacting results because the heterogeneity smooths out the differences between the outcomes of control and experimental groups, limiting the clinical power of RCTs. RCTs frequently provide real insights about the benefit of the treatment in subgroups of the included patients, reducing the statistical power of the results. For example, So et al. [21] and Zheng et al.[22] verified that when they did a sample division, certain children particularly improved and others did not.

Moreover, since this heterogeneity is also present in the different capabilities of the children, it is difficult to find a unique biological characteristic (biomarker) that can be used to classify the evolution of the children during therapy sessions. In order to have a more global overview, the clinicians prefer to use scales or questionnaires as the previously mentioned ESCS. These evaluation instruments can also be compiled by the children's parents and evaluate the children's behaviour during artificial tasks that specifically elicit certain behaviours or during daily life tasks. The scores are then computed by an observer during or after the session. Thus, there is a lack of objective quantitative measures, which does not allow a clear comparison between studies and the setting of precise directions to move forward.

The lack of quantitative measures is also related with the extreme sensitivity to touch of these children, preventing the use of several types of sensors. Therefore, in robotic-assisted therapies, the most used sensors are cameras. Moreover, in most of the therapies, children are allowed to move freely in the room. Thus, measuring some of the capabilities referred before such as gesture performance or attention becomes harder since the gold standard sensors such has optoelectronic systems or eye-tracking devices as Tobii can not be used. In addition, children are frequently held by therapists or closely supervised, leading to occlusions that present significant challenges for computer vision systems.

Most studies in this field are small-scale pilot studies involving few children, resulting in limited datasets. This scarcity of data makes it difficult to apply advanced machine learning techniques, particularly modern neural networks, to develop new measurement methods. Moreover, due to the sensitive nature of the data (video recordings of children), ethical constraints often prohibit data sharing between research groups, further hindering the creation of large, diverse datasets. This lack of generalizable algorithms represents a major bottleneck in the field, perpetuating the challenges associated with achieving reliable quantitative measures.

1.2 Objectives

Our main goal is to develop flexible protocols and new quantitative measures for robotic-assisted therapies in autism which can be used in clinical practice. To achieve this goal, we establish the following sub-goals:

- Development of a robotic protocol related to the clinical practice of this disorder for an easier technology transfer into the clinical setting

- Construction of quantitative measures based on computer vision and machine learning algorithms that can evaluate the children during the sessions and pre- and post-treatment (online and offline evaluations).

- Design and implementation of several experimental studies (from Pilot Studies to RCTs) to identify the best target population and provide evidence of the effect of the system on this population.

This thesis work was developed between Politecnico di Milano, in Italy and Instituto Superior Técnico, in Portugal and involved a close collaboration with three clinical institutions: Fondazione Don Carlo Gnocchi (FDG) in Milan, Italy; Associação Portuguesa para o Autismo e as Perturbações do Desenvolvimento (APPDA), in Lisbon, Portugal and Centro de Apoio ao Neurodesenvolvimento (CADIn), in Lisbon, Portugal.

We present our contributions in a compact form with one chapter for each type of contribution. These contributions resulted from a dynamic and cyclic development process covering four methodological steps: literature review; development of computer vision algorithms; protocol design and testing (Figure 1.1). These steps were iterated in different orders in four moments of the development of this work:

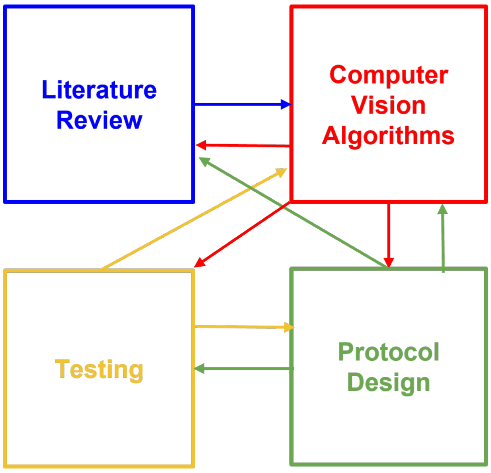


Figure 1.1: Design cycle of the general approach of the thesis.

(i) We started by reviewing the literature and talking to the clinicians of our first partner institution, FDG. From this initial search, we decided to focus on motor skills, specifically on imitation skills to indirectly work on social skills. Thus, we designed a new protocol where a robot imitated a child, that was tested on two ASD children. These tests highlighted the importance of creating a mirroring metric to assess participants' performance and tailoring protocols to match each individual's abilities (Section 4.1). This metric was established through the extraction of movement features of the child obtained through a depth camera.

(ii) From the testing of this quantitative measure, we designed a new protocol with different levels adjusted to the severity of Autism and focus on a specific skill: gestures production (Section 3.2). For this gestural protocol, we developed a feedback system based on the recognition of gestures through a Convolutional Neural Network (Section 4.2). Moreover, a quantitative metric focused on children's attention during therapy was constructed from a gaze tracker algorithm present in the literature whose inputs are just video frames (Section 5.1). We adapted this metric so that it could be used both offline for therapy evaluation and online to provide a feedback to the child (Section 5.2).

(iii) Based on the results of a pilot study conducted with children with autism in Milan, we decided to enhance our protocol to better address the needs of children with lower levels of autism, emphasizing luminous, auditory, and movement stimuli (Section 3.3). Additionally, a study carried out in Lisbon prompted the development of a simpler protocol to complement rehabilitation sessions. This protocol involved a bingo game with a robot (Section 3.5). At the same time, to establish a robust outcome measure, we created a protocol with the robot based on the ESCS (Section 3.4). In this case, as the children remained in fixed positions, we developed a deep learning method to enhance the accuracy of our attention metric.

(iv) In the end, we extended a pose estimator to be applied in children to predict better the positions of the several participants and, in this way, improve the accuracy of the created metrics (Section 6.1). We have also explored the possibility of using this estimator online (Section 6.2).

In summary, in this document, Chapter 2 describes our literature review, Chapter 3 focus on the clinical protocols, Chapter 4 on the metrics related with imitation and gesture, Chapter 5 on the metrics related with the attention and Chapter 6 describes the adaptation of a pose estimator. Finally, Chapter 3 presents the results of our contributions on the two largest clinical studies developed during this thesis.

1.3 Summary of contributions

The contributions of this work can be summarized into two parts:

(i) Protocols

1. Hierarchical protocol: Designed for exploring intransitive gestures with a robot, where children progress through different levels.
2. Sensorial protocol: Developed based on the results of the first clinical study, this protocol focuses on the robot's core functionalities, such as lights and sounds.
3. Scale protocol: To obtain more precise outcome measure, a protocol was created integrating the robot with the ESCS, allowing for the evaluation of a child's attention.
4. Bingo protocol: As an alternative to the sensorial protocol and following findings from the clinical study developed in Portugal, this protocol introduces a collaborative bingo game involving the robot, the child, and the therapist.

(ii) Algorithms/Measures

1. Mirroring measure: Using data from a depth camera, this algorithm evaluates children's movements during exercises and adjusts the difficulty of rehabilitation protocols accordingly.
2. Gesture recognition algorithm: Developed for the first protocol, this real-time algorithm identifies intransitive gestures performed by a person and provides immediate feedback and motivation to the child during therapy.
3. Attention biomarker: Recognizing the importance of attention in autism therapy, this algorithm determines where the child is looking at any given moment. We posed the problem as a classification problem, and joined a gaze estimator available in literature with some geometrical assumptions. In this way, we summarized the engagement of the child versus different targets during a clinical session.
4. Neural Network Classifier: For the scale protocol, a Multilayer Perceptron combined with a Convolutional Neural Network was used to address the attention classification problem in a more constrained environment.
5. 3D body pose estimator for children: An existing 3D pose estimator was modified and adapted for children by adjusting a specific parameter. This algorithm was designed for both offline and real-time use during therapy sessions.

1.4 List of publications

The work developed in this thesis was partially published in journals and conferences listed below. * represents an equal contribution.

Journal Papers

(A) 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, 2021

(B) A. S. Ivani*, A. Giubergia*, L. Santos, A. Geminiani, S. Annunziata, A. Caglio, I. Olivieri, and A. Pedrocchi, “A gesture recognition algorithm in a robot therapy for ASD children,” Biomedical Signal Processing and Control, 2022.

(C) L. Santos, S. Annunziata, A. Geminiani, A. Ivani, A. Giubergia, D. Garofalo, A. Caglio, E. Brazzoli, R. Lipari, M. C. Carrozza, E. Ambrosini, I. Olivieri, and A. Pedrocchi, “Applications of robotics for Autism Spectrum Disorder: a scoping review,” Review Journal of Autism and Developmental Disorders, 2023

(D) B. Silva*, L. Santos*, C. Barata, A. Geminiani, G. Fassina, A. Gonzalez, S. Ferreira, B. Barahona-Corrêa, I. Olivieri, A. Pedrocchi and J. Santos-Victor, “Attention Analysis in Robotic-Assistive Therapy for Children with Autism,” IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2024

(E) S. Annunziata, L. Santos, A. Caglio, A. Geminiani, E. Brazzoli, E. Piazza, I. Olivieri, A. Pedrocchi and A. Cavallini, “Interactive mirroring Games wIth sOCial rObot (IOGIOCO): a pilot study on the use of intransitive gestures in a sample of Italian preschool children with Autism Spectrum Disorder,” Frontiers in Psychiatry Autism, 2024

Conference Papers

(F) L. Santos, A. Geminiani, I. Olivieri, J. Santos-Victor, A. Pedrocchi, “CopyRobot: Interactive Mirroring Robotics Game for ASD Children”, XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019, IFMBE Proceedings, vol 76, Springer, Cham.

(G) A. Geminiani, L. Santos, C. Casellato, A. Farabbi, N. Farella, J. Santos-Victor, I. Olivieri, A. Pedrocchi, ”Design and validation of two embodied mirroring setups for interactive games with autistic children using the NAO humanoid robot,” 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 2019, pp. 1641-1644

(H) L. Santos, S. Annunziata, A. Geminiani, E. Brazzoli, A. Caglio, J. Santos-Victor, A. Pedrocchi, I. Olivieri, “Interactive Social Games with a Social Robot (IOGIOCO): Communicative Gestures Training for Preschooler Children with Autism Spectrum Disorder”, I Congresso Annuale Rete IRCCS Neuroscienze e Neuroriabilitazione, 2020

(I) G. Fassina, L. Santos, A. Geminiani, A. Caglio, S. Annunziata, I. Olivieri, A. Pedrocchi, “Development of an Interactive Total Body Robot Enhanced Imitation Therapy for ASD children.”, International Conference on Rehabilitation Robotics, 2022

(J) L. Santos, B. Silva, F. Maddaloni, A. Geminiani, A. Caglio, S. Annunziata, I. Olivieri, C. Barata, J. Santos-Victor, A. Pedrocchi, “Sharing Worlds: Design of a Real-Time Attention Classifier for Robotic Therapy of ASD Children”, International Conference on Rehabilitation Robotics, 2022

(K) L. Santos, M. Murgo, B. Silva, A. Geminiani, A. Caglio, S. Annunziata, C. Barata, J. Santos-Victor, and A. Pedrocchi, “An attention classifier for the evaluation of a robotic therapy in children with autism spectrum disorder,” in Gruppo Nazionale di Bioingegneria - 8th Congress of Bioengineering, 2023.

(L) L. Santos, B. Carvalho, C. Barata, and J. Santos-Victor, “Extending 3d body pose estimation for robotic-assistive therapies of autistic children,” in IEEE RAS EMBS 10th International Conference on Biomedical Robotics and Biomechatronics (BioRob 2024), 2024

The association between these publications and the thesis chapters is the following:

- Chapter 2 - [C] and [G]
- Chapter 3 - [E], [F] and [H]
- Chapter 4 - [A], [B] and [I]
- Chapter 5 - [D], [J] and [K]
- Chapter 6 - [L]
- Chapter 7 - [D] and [E]

Chapter 2

Background and Literature Review

We start by giving a general overview of the research on Autism Spectrum Disorder, focusing on the diagnosis, the several factors involved in the disorder and the current therapies. Then, we present a synthesis of robotics for autism, considering both the clinical and engineering aspects. The chapter takes inspiration from our journal article ”Applications of robotics for Autism Spectrum Disorder: a Scoping Review”, for the Review Journal of Autism and Developmental Disorders [23]. Special focus is given to the metrics used since their definition is a central topic throughout the thesis. Moreover, ten randomised controlled trials are analysed in detail to understand the impact of this type of therapy in children with Autism. We then focus on the engineering challenges presented in the Review and show the solutions already found in other fields.

2.1 Autism Spectrum Disorder

The diagnosis of autism in children occurs through observation by the clinicians of child's interactions with different people combined with detailed developmental history generally provided by the parents. Through these data, as previously presented in Chapter 1, the clinicians verify whether the child has the two core symptoms of autism present in the Diagnostic and Statistical Manual of Mental Disorders-5th edition (DSM-V): social communication and interaction deficits and restrictive and repetitive behaviours. Depending on the severity of the symptoms, the level of support required by ASD individuals can be classified into three levels: level 3 in which the subject is ”requiring very substantial support”, level 2 in which the subject is ”requiring substantial support”, and level 1 in which the subject is ”requiring support”. The severity of symptoms can vary between individuals and during the developmental trajectory of the same individual. The early signs begin around 2 years of age but the definitive diagnosis appears around 4-5 years old [24].

In the past, clinical assessment solely was considered the most reliable way to diagnose autism. Recent evidence shows that this is not true, especially for toddlers and preschool children, with whom instrument scores, like scales, are used [24]. The most widely standardized instruments are the Autism Diagnostic Observation Schedule (ADOS), a semi-structured observation of the child's behaviour and the Autism Diagnostic Interview-revised (ADI-R), a semi-structured interview with the parents [25]. These two scales evaluate the presence or absence of different abilities in the child. Till this moment there is not enough evidence for the presence of specific biomarkers in autism which could considerably help the diagnosis.

2.1.1 A multifactorial disorder

ASD is regarded as a multifactorial disorder influenced by both genetic and environmental factors, though none appear to be uniquely specific to ASD [12]. Twin and family studies estimate autism's heritability to range from 40% to 90%, with sibling studies playing a critical role in advancing the understanding of autism's neurobiological mechanisms [24]. Both magnetic resonance imaging (MRI) and electroencephalography (EEG) have emerged as possible technologies for the detection of biomarkers related to autism. By biomarkers we intend any medical signs which can be measured quantitatively and reproducibly [26]. In the case of MRI and EEG, significant differences were observed in the brain structures and brain connectivity [24, 27] in studies that compared the autism population with the neurotypical population. These biomarkers could be particularly important for the ease of the diagnosis, but since these studies always have small sample sizes, significance levels of evidence are never reached. Their replication is also prevented by the intrinsic heterogeneity of the disorder. In addition, the acquisition methods of these two technologies interfere with one of the symptoms of autism, the sensory anomalies (hypersensory and hyposensory responsiveness) that in the DSM-V were included in the restrictive behaviour symptoms [24]. For instance, MRI necessitates the use of noise-cancelling headphones, and EEG requires wearing a specialized cap, both of which can be uncomfortable or distressing for individuals with ASD.

Alternative biomarkers could be the attention and the gestural repertoire of the children with ASD. Related to the main symptom of social communication and interaction deficits, these types of non-verbal communication (attention and gestures) appear often altered in individuals with autism [28]. Attention can be defined as the process of concentrating on selected items of the environment, to the exclusion of other unattended stimuli [28]. When these items are social items, such as people, we talk about social attention. Social attention appears from infancy and it is highly adaptive, since it enhances the opportunities of social experiences important for the development of social communication skills such as language processing or emotion recognition [29]. However, ASD individuals appear to have a preference for non-social stimuli, spending less time focusing on social stimuli when compared with neurotypical children [29]. Moreover in terms of face-processing, ASD subjects spend significantly less time looking to the eyes than neurotypical subjects, developing an alternative strategy of looking more to the mouth [30].

The lack of social attention in children with ASD is hypothesized to impact the development of other social cognitive skills such as joint attention [28]. Joint attention refers to the social coordination of one's attention with that of another person to a common fixation point in order to share information [31]. This process begins between the 2nd to 12th months of age and it is constituted by two types of behaviours. The first that appears is the "Responding to joint attention" (RJA) that consists of the ability to follow the direction of attention of another person. Secondly, there are the "Initiating joint attention" (IJA) behaviours which refer to the ability to spontaneously direct the attention of others to share their experience of an object or an event. Joint attention skills include gestures such as pointing, coordinated looks between objects and people, and showing.

A lower tendency to join joint attention behaviours is considered a prominent factor in young children with ASD. It can be measured through a clinical scale called Early Social Communication Scale (ESCS), especially designed for the assessment of non-verbal communication skills [32]. This scale may be used in neurotypical children between 8 and 30 months of age or children with developmental disorders whose verbal age estimate is in the same range. It normally takes 15-25 minutes to be administered. Usually, the child and an adult tester are seated at a table where several objects are presented 2.1. 17 semi-structured tasks elicit different interactions between child and adult tester. A camera is oriented to capture three quarters to full-face view of the child while also capturing a profile viewer of the tester.

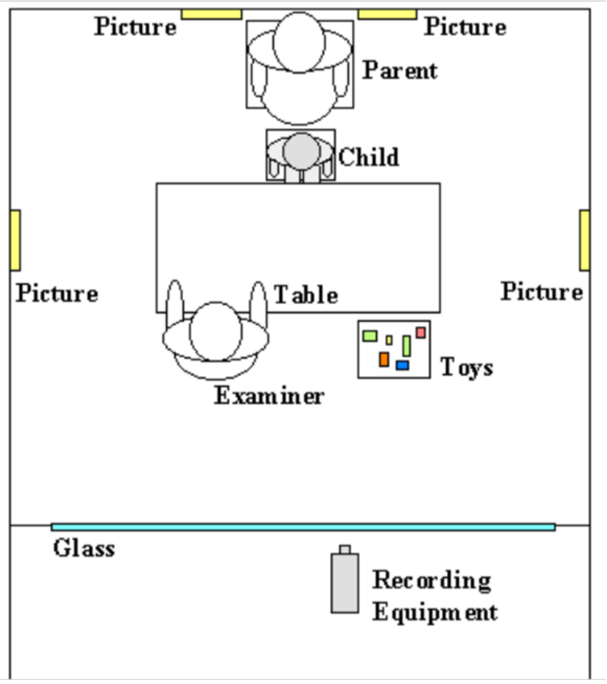


Figure 2.1: Early Social Communication Scale (ESCS) setup

Parents can be present or not during the test. The setup is shown in Figure 2.1. After an operator codes the different behaviours of the child. Examples of IJA behaviours are establishing eye contact with the tester, alternating the eye contact between the tester and the objects shown, pointing to some objects, or showing an object to the tester. Regarding the RJA behaviours, the child should follow the pointing of the tester both in terms of his/her proximal point (finger) and the object pointed [32].

Among the non-verbal communication skills in which children with autism present particular deficits, there is the gestural production. Gestural production and language in autism seem to not have any differences in children with autism when compared with neurotypical subjects in the first year of life, but present significant differences in the second year [33]. Children with autism produce a reduced number of gestures, namely communicative gestures [33, 34]. Joint attention gestures, like pointing, are the communicative gestures that present a significant reduction in autistic children when compared with neurotypical controllers [34].

Gestural production is normally elicited by imitation [33]. Children with ASD often show deficits in this skill that reflect deficits in both social and cognitive processes. Imitation is an important mechanism for transmitting information and learning, especially for individuals who have not acquired language. Imitation can be divided into two types: high-fidelity imitation in which subjects reproduce both the form and the end result of the modelled action; emulation in which the final result is achieved but the subject chooses its own action [35]. Individuals with ASD show significant differences compared to neurotypical subjects in high-fidelity imitation but not in emulation [35]. Similarly, focusing on the imitation of gestures, autistic children can often choose the appropriate gestures (emulation) but they are less accurate than the gestures produced by neurotypical children (high-fidelity imitation) [36]. In addition, groups of subjects with more severe forms of ASD show greater imitation deficits than those with less severe forms of ASD. However, it is not clear whether an imitation deficit leads to the development of ASD or it is a symptom of this disorder [35].

2.1.2 The therapies

Regarding interventions for ASD, they are mainly focused on behavioural and developmental therapies, since no medication is able to reduce the core symptoms of autism [12]. Some pharmacological intervention is used for co-occurring disorders such as ADHD or obsessive-compulsive disorder. Behavioural therapies have focused on young children but can be applied to people of any age [12]. Early intervention is prioritised because the core symptoms limit the capability of children to learn and are particularly challenging for their parents. Nowadays, the therapies with more clinical evidence are the naturalistic developmental behavioural interventions (NDBI) which join the principles of the behavioural approaches like ”Applied Behaviour Analysis” (ABA), in more naturalistic scenarios of the developmental approach (e.g: ”Developmental, Individual-differences, and Relationship-based” (DIR)).

In ABA, positive reinforcements reward desired behaviours and the trials are discretised to provide opportunities for engaging and repetitive practice [37]. It is an effective tool for reducing negative behaviours while learning a new skill. In the DIR, floor-based play with the child is conducted by the therapist or the parent, with the intervention tailored to the child's specific needs [38]. Many current interventions have evolved from traditional ABA, incorporating more naturalistic, child-initiated developmentally appropriate strategies and tasks, rather than relying on discretized activities. However, there remains a lack of clear evidence regarding the optimal intensity of treatment or the most suitable approach for each child, as no direct comparisons between different treatments have been conducted [24].

Clinical trials in autism are mainly limited by the cost, the time and limited outcome measures, being far behind other clinical research [24]. Even in cases in which there are some significant treatment differences between the children of different groups (treatments vs no treatment or treatment 1 vs treatment 2), individual results are very variable and some children do not improve. That is why 'biomarker' based psychological intervention is considered a promising area, although these therapies are still being developed. Meanwhile, parents are limited to what is available or marketed in their region and since no clear evidence exists in relation to the treatments, they often search for alternative therapies. One of the alternative therapies that could be a fertile place for biomarkers-based therapies is robotic-assisted therapy. In this document, robotic-assisted therapy is referred to as robotic therapy for purposes of brevity.

2.2 Social Robots in Autism

Recently, robots have been used for treating autism thanks to their predictability and repeatability, which makes them easier to understand by children. An extensive review of this topic in the literature was done as an initial step of this thesis. The keywords used were ”robot” and ”autism”, the databases used were three (*Web of Science*, *Pubmed* and *Scopus*) and the publication date was chosen between January 2016 and October 2020. These dates were chosen since in the beginning of this thesis in November 2019, the last available review in the topic was from 2016.

In total, the three databases selected 804 papers. Then, we eliminated all duplicates and all works that verified our exclusion criteria, namely, papers that did not have any robot, did not describe well the role of the robot or in which the robot was used as a model of the disorder and not for an interaction. We also excluded the papers that were not related mainly to autism, presented the results for the people with ASD aggregated with other disorders, or had the same study design and participants of other studies. Finally, since we were interested in the quantitative significance, we excluded papers in which less than two children interacted with the system. In the end, 146 papers were analysed. We focused our analysis first on the setups and protocols used in this field, followed by the evaluation measures chosen on the different studies. After a comprehensive review of the topic, we scanned the randomised controlled trials present in our sample and developed a meta-analysis, a type of analysis in which the data of different studies is combined to determine overall trends.

2.2.1 Setup and Protocol Design

One of the main conclusions is that the number of papers on this topic has risen exponentially in the last years (Figure 2.2). Most studies focus on improving the social skills of children [39, 40, 41, 42, 43, 44, 45], namely joint attention or emotion recognition. Robots are also used to train motor skills, namely imitation which, as previously mentioned, is important for for learning[46, 47, 48, 49, 50]. In addition, there are several studies in which robots participate in the diagnosis of autism [51, 52, 53, 54, 55].

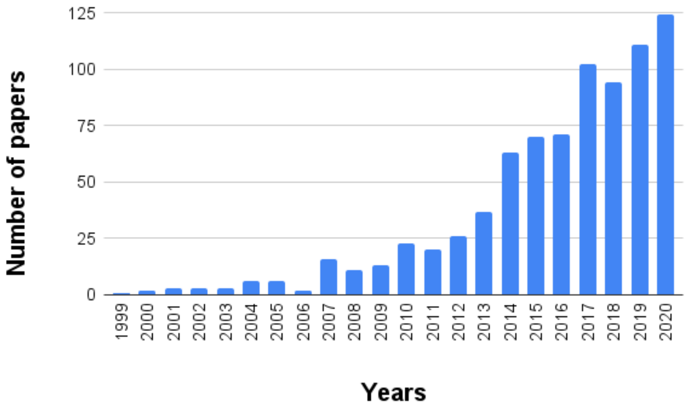


Figure 2.2: Distribution of the papers about robots in autism across the years.

Overall, 76% of the studies use humanoid robots since they have simpler expressions, which ease the work with ASD children when compared with non-humanoid robots. NAO is the robot most frequently mentioned in the majority of the papers (Figure 2.3) [56, 57, 58, 59, 60, 61, 62, 63, 64, 18]. Its widespread use can be attributed to its status as a commercial robot. It has 25 degrees of freedom, 16 LEDs in the eyes and two loudspeakers, ideal for clinical uses. The other humanoid robots by order of utilisation frequency are Zeno [65, 66, 67, 50, 68, 53, 69], ActroidF [70, 71, 72, 73, 45], CommuU [71, 72, 17] and Kasper [74, 75, 76]. All have more facial expressions than NAO and have been used for training different skills, but their availability in the market is much lower than NAO's one.

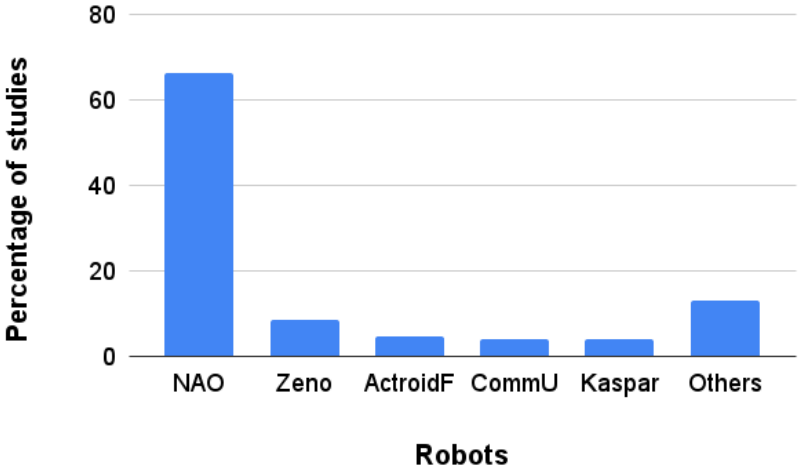


Figure 2.3: Humanoid robots used in the literature.

In most cases, the robots are controlled through a Wizard of Oz strategy, in which an operator is in another room, controlling the robot's actions [77, 70, 45, 78]. There is a general tendency to increase the adaptability and autonomy of the robot (Figure 2.4) due to the heterogeneity of symptoms and behaviours previously mentioned [79, 22, 80]. Therefore, four other types of robotic control can be considered, according to [81]:

(i) hybrid control: the robot has some autonomous behaviours, but most of them are activated through an interface controlled by an adult or a child [82, 83, 84, 85, 86]

(ii) semi-autonomous control: all behaviours of the robot are autonomous, but they need to be approved by a therapist [87, 88, 78, 52, 89]

(iii) fully autonomous control: the robot behaves independently without any supervision, being an open loop control system [90, 91, 92, 93].

(iv) autonomous and adaptive control: the robot has the objective of adapting the exercises and protocols to each subject, maintaining people's engagement throughout the therapy [79, 22, 80]. In this case, a biomarker, like the attention of the child is chosen to control the robot, in a controlled feedback (closed loop) manner.

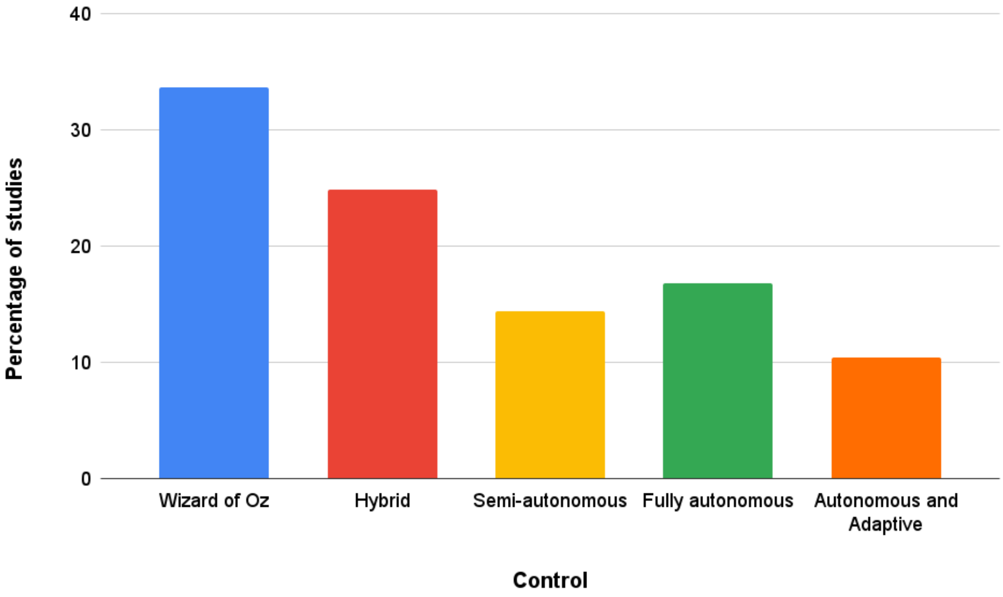


Figure 2.4: Types of controllers reported in the analysed studies.

Concerning the protocol, in some cases, the robot interacts solely with the child in a dyadic interaction [77, 70, 88], while in other cases the session includes another agent, which can be the therapist, the researcher or the parent, in a triadic interaction [94, 91, 95, 96]. Triadic scenarios are less common than dyadic ones (78% vs 21% of the analysed studies), but they are of outmost importance since social capabilities and engagement improvement can be achieved through triadic interactions [87, 67].

In terms of session duration, the majority of the studies reports a short duration (inferior to 15 min), due to the low span of attention of children with ASD [97, 98, 99, 100]. Many studies are pilots that have just one session [101, 102, 103], to analyse the protocol and platform feasibility, and the subject's engagement towards the new robotic platform (Figure 2.5). Moreover, in the studies with a larger number of sessions, few authors have analysed the effects in a long-term scenario, with just 12% of the studies having a follow-up evaluation [94, 104, 78, 96]. These facts compromise the construction of general conclusions about the robots in Autism.

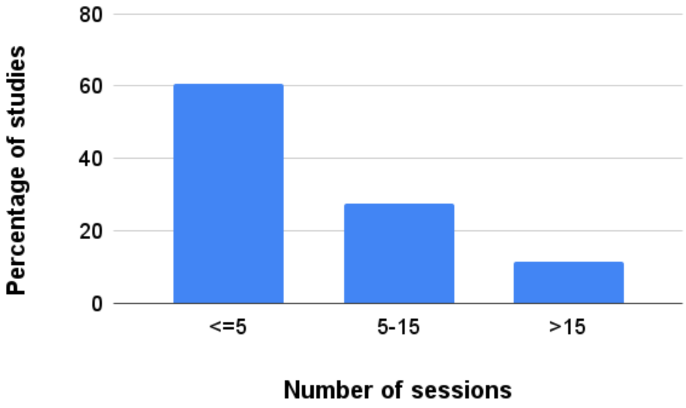


Figure 2.5: Number of sessions reported in the analysed studies.

Regarding the number of participants, most papers describe pilot studies with a low number of participants, typically less than 10 (Figure 2.6)[105, 106, 107]. Reduced sample sizes are associated with the difficulty of homogenising the groups of children who participated in the robotic therapies, a common problem in the autism field as described in Section 2.2. Therefore, these studies have low statistical power. Trying to overcome this problem, few studies have more than 30 participants [108, 78, 77]. Another way to deal with the intrinsic heterogeneity of ASD children and extract meaningful results for the studies is by using a control group. About 13 % of the studies have chosen a control group of ASD subjects performing standard therapies [107, 61, 60].

Concerning the age of the participants, most studies focus on primary school children [109, 110, 111], probably due to better compliance with the humanoid robot and task comprehension. Given the importance of early intervention on ASD, numerous studies have been conducted involving sessions with preschool children [112, 113, 53].

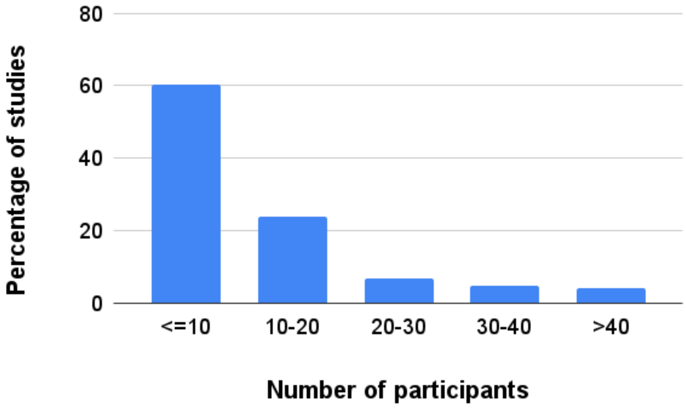


Figure 2.6: Number of participants reported in the analysed studies.

2.2.2 Performance Evaluation Measures

For the evaluation of the children's evolution during the therapies, the measures used can be divided into two types: (i) qualitative, in which a therapist, a researcher or a parent describes the child's performance based on their previous knowledge, for example through clinically validated questionnaires [108, 101, 114]; and (ii) quantitative, in which a numerical value is associated with the child evolution for a given therapy. Both types of measures can be calculated from direct observation during or after each session, using a video.

The quantitative measures can be divided further into three types: manual, automatic and physiological. First, in the manual collected measures, an operator is responsible for manually classifying the child, either counting the number of times a certain behaviour is verified [41, 42, 47, 48] (ex: number of times the child did the exercise correctly, number of times the child looked at the robot) or by filling certain scales (ex: Early Social Communication Scale [115, 112, 18]). In the automatic measures, specific sensors, such as the Tobii eye tracker [41], or computer vision algorithms are used to provide these measurements automatically. For example, [67] used the dynamic time warping algorithm to compare the movement of the robot and the one of the child and obtain a metric of the performance. Another example is [98], which used an emotion recognition algorithm to classify and rank children's emotions during therapy. To further improve the accuracy, recently, physiological measures have been taken from the children to assess the stress, for example, through salivary tests [85] or heart rate variability [116] or the attentiveness through the EEG power density [42].

In summary, researchers are trying to convert manual and qualitative measures into automatic and quantitative data. The wider implementation of automatic measures would allow a better comparison between the studies and a deeper understanding of the impact of the robots on ASD.

2.2.3 Randomised Controlled Trials

From the set of analysed studies, we selected the Randomized Controlled Trials (RCTs) to understand better the impact of robots on autism in studies with a higher statistical power. Ten papers were extracted. These RCTs reflected the diversity of purposes of the robotic therapies, going from the training of joint attention [22, 117] to gesture imitation and recognition [118, 119, 21, 18].

All the studies had a control group, formed by ASD children not performing any therapy while waiting to be admitted to the robotic one (waiting list group) [118, 119, 18, 22]. Some authors have included other control groups to provide further comparisons with the ASD children, for example, a group of children doing another therapy, e.g. rhythm therapy, as in [117] or a Typical Development (TD) group as in [119].

The outcome measures are detailed in Table 2.1 and Table 2.2 for each study. Seven out of the ten selected studies had outcome measures with a significant positive effect ($p<0.05$) towards the robotic group in parametric and non-parametric tests[118, 119, 18, 63, 39, 120, 121]. This effect was confirmed during the follow-up in three of the seven studies [63, 119, 118], demonstrating the long-term maintenance of the skills learnt during robotic therapy.

The three studies that did not verify a significant positive effect justified the result with different reasons: the lack of adaptability of the robot, which became boring for the children [117] and the heterogeneity of the children [22, 21]. In particular, Zheng et al. [22] explained their results by splitting the robotic group into several groups, considering the improvement or not of an outcome measure. Since they had three outcome measures, one clinical and two from the automatic system, six statistical tests were done. They verified that with this sample division, the changes within each subgroup were significant for the system measures, not for the clinical one. When looking to the clinical characteristics of the participants, just the age was close to the significance, showing that younger participants tended to improve their system metrics. So et al. [21] showed a similar problem in their results, where their variability was justified by the different severity of autism, cognitive functioning and communication skills of the participants.

Table 2.1: Outcome Measures of Randomized Controlled Trials - Part I

References	Outcome Measures	Significant effect in favour of the robotic group
Pop et al., 2017	Frequency of correct strategies to be used in a social situation	no
	Frequency of used rational beliefs	yes
	Frequency of used irrational beliefs	no
	Emotional intensity rating scale for angry and sadness	yes
	Frequency of using adaptive or dezadaptive behaviours	no
Korte et al., 2020	Number of self-initiations	yes
	Social Responsiveness Scale teachers	no
	Social Responsiveness Scale parents	yes
Marino et al., 2020	Test of Emotion Comprehension	yes
	Emotional Lexicon Test	yes
Srinivasan et al., 2016	Joint Attention Test	no
Yun et al., 2017	Autism diagnostic observation schedule	no
	Vineland adaptive behavior scale	no
	Social communication questionnaire	no
	Social Responsiveness Scale	no
	Child behavior checklist	no
	Frequency of eye contact	yes
	Accuracy of facial emotion expression	no

Table 2.2: Outcome Measures of Randomized Controlled Trials - Part II

References	Outcome Measures	Significant effect in favour of the robotic group
Zheng et al., 2020	Screening Tool for Autism in Toddlers and Young children	no
	Average prompt level that a participant needed to hit a target in a trial	no
	Target hit rate	no
So et al., 2018a	Number of gestures recognized Training Scenarios	yes
	Number of gestures recognized Non training scenario	yes
	Number of gestures recognized Non training scenario with human	yes
	Number of gestures produced Training Scenarios	yes
	Number of gestures produced Non training scenario	yes
	Number of gestures produced Non training scenario with human	no
So et al, 2018b	Number of correctly produced gestures traning scenario	yes
	Number of correctly produced gestures non training scenario	yes
	Number of appropriate gestures	no
	Number of gestures with verbal imitation	yes
So et al., 2019	Number of correct recognized gestures	no
	Number of produced gestures	no
So et al., 2020	ESCS Initiation of Joint Attention	yes
	ESCS Response to Joint Attention	no
	SPA Other directed functional play	yes
	SPA Self directed functional play	no
	SPA Symbolic play	no
	Social Responsiveness Scale	yes

Meta-analysis

Five RCTs reported the full numerical values, which we combined in a meta-analysis. We started by classifying the outcome measures of the RCTs according to the International Classification of Functioning, Disability and Health (ICF) framework [122]. The ICF is a manual produced by the World Health Organization which provides a standardized language and a conceptual basis for the definition and measurement of health and disability. When classifying the health of an individual, it considers the presence or absence of specific criteria and their corresponding intensities [123]. For each outcome measure reported in the studies, we matched it to the criterion in the ICF framework that most closely aligned with its description. This approach allowed us to group metrics with varying names but similar focuses, thereby reducing heterogeneity introduced by combining different types of studies.

Two meta-analyses were constructed, considering the areas relevant to our study: one regarding the [d720] Complex interpersonal interactions and another on the [d160] Focusing attention. The first criterion refers to the capability of sustaining a structured interaction with others, in a contextually and socially appropriate manner. The second criterion refers to the ability to sustain attention, filtering out all disturbing noises [122].

For both meta-analyses, the Standardized Mean Difference (SMD) was chosen to analyse the treatment effects. This summary statistic enables the combination of various outcome measures that were grouped based on the same criterion. Basically, it expresses the size of the intervention effect of each study, relative to the variability observed in that specific study, which is summarized by Equation 2.1,

$$SMD^i = \frac{\overline{X_E^i} - \overline{X_C^i}}{S_{within}^i} J^i \tag{2.1}$$

where $\overline{X_E^i}$,$\overline{X_C^i}$ are the estimated means of the experimental and control group respectively, J^i is a correction factor introduced by Hedges [124] to compensate for the underestimation of this statistic. S_{within} is the within groups standard deviation pooled across the two groups, given for each study i by Equation 2.2,

$$S_{within}^i = \sqrt{\frac{(n_C^i - 1)^2(S_C^i)^2 + (n_E^i - 1)^2(S_E^i)^2}{n_C^i + n_E^i - 2}} \tag{2.2}$$

where n_C,n_E are the total number of participants in the Control and Experimental group and S_C, S_E are the standard deviations of the Control and Experimental group. In this way, the intervention effect of each study is expressed in units of standard deviation. The overall effect is given by a weighted average of each study's intervention effect, in which the weights consider the sample size and the standard deviation of the study. For further details, consult [125]. For our statistical analysis, we used the Cochrane review writing program RevMan [126], considering the confidence interval of 95%.

In the first meta-analysis, we combined the results of four studies, resulting in a sample of 96 children. Figure 2.7 (a) shows a tendency in favour of the experimental (robotic) group (Standardized Mean Difference 0.22; Confidence Interval [-0.52; 0.97]), but this tendency is not statistically significant. In the second meta-analysis, just two studies had a similar definition of outcome measures (target hit rate and frequency of eye contact), joining 36 children. Again, there is a tendency towards the experimental group (Standardized Mean Difference 0.18; Confidence Interval [-0.49; 0.84]) but not statistically significant.

The design, execution and analysis of a randomized controlled trial can lead to bias, yet the extension to which this influences the results is unclear. For this reason, assessing the risk of bias in each study included in a meta-analysis is standard practice. For our analysis, we use the RoB2 tool from Cochrane [127] in which seven domains are classified as ”High risk of bias”, ”Some concerns” or ”Low risk of bias” [128]:

- Bias arising from the randomization process: assesses whether the allocation sequence (e.g., the order in which children were assigned to control and experimental groups) was random and adequately concealed. It also checks for baseline differences between the two groups that may suggest problems in the randomization process;
- Bias due to deviations from intended interventions: examines whether the patients, caregivers and people delivering the intervention were aware of the assigned intervention during the trial and whether there were other collateral interventions;
- Bias due to missing outcome data: considers whether outcome data is available for all, or nearly all the participants that took part in the randomization process;
- Bias in measurement of the outcome: evaluates whether the outcome measurement was inappropriate or assessed differently between the two groups. In addition, it verifies if the outcome assessors were aware of the intervention;
- Bias in selection of the reported result: investigates whether an analysis plan existed prior to the unblinding the outcome data and whether the numerical results reported were selected according to multiple analyses or from different outcome measures.

Then, an overall risk of bias judgement is obtained for the specific outcome measure, which considers all the individual domains. In this way, each study has its own classification of risk of bias. Studies are classified as follows [128]:

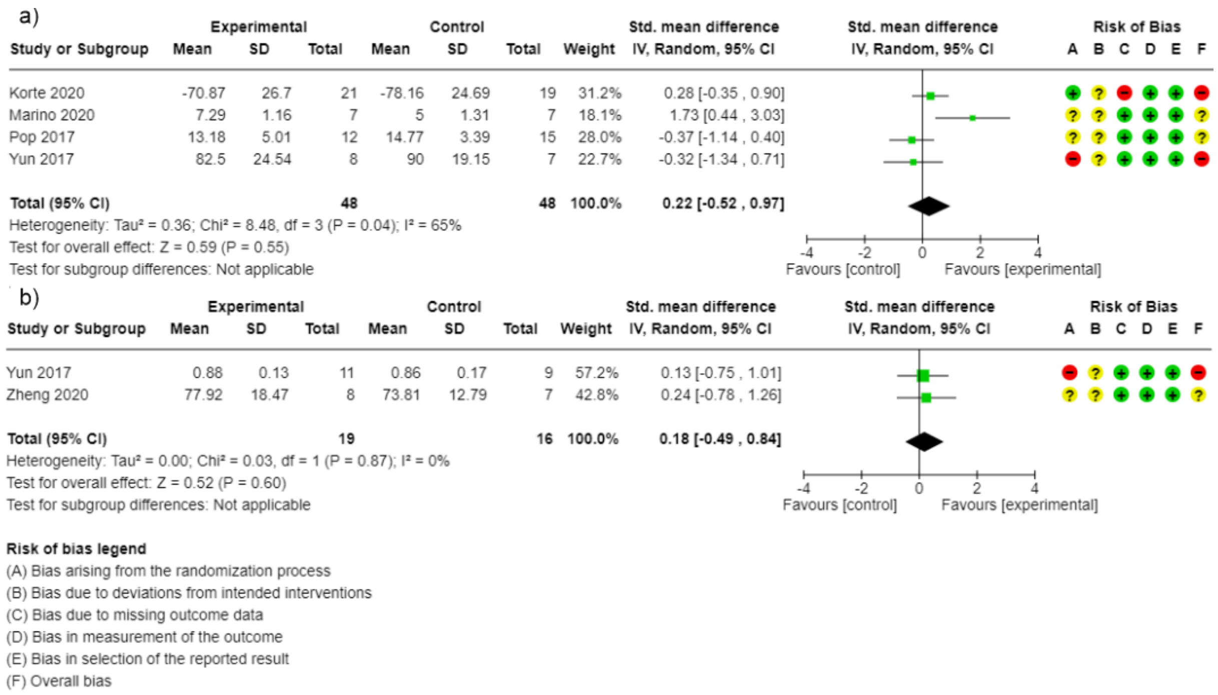
- Low Risk of Bias: All domains are rated as ”Low Risk of Bias.”
- Some Concerns: At least one domain is rated as ”Some Concerns,” but none is rated as ”High Risk of Bias.”
- High Risk of Bias: At least one domain is rated as ”High Risk of Bias.”

In our meta-analysis, out of the five studies, two studies were classified as ”High Risk of Bias” while three studies as ”Some Concerns”, which is a symptom of the lack of high-quality evidence in this field. The more critical criteria were the ones related to the randomization process and the deviations from intended interventions. The first was due to the lack of information regarding the generation of the allocation sequence and how it was concealed by the participants. The second was related to the fact that the participants were not restricted from taking other collateral interventions. Although this constitutes a Risk of Bias, it is a mandatory recommendation of the Ethical Committees since children cannot stop their rehabilitation process for therapies whose effect is unknown.

2.2.4 Technical limitations and clinical domain constraints

In summary, the current principal limitations of the studies presented in the literature are three: (i) the short duration of the studies; (ii) the low number of participants; (iii) the lack of scientific evidence essential for clinical translation, caused by few quantitative measures, which are necessary for the comparison between different works, and the reduced number of randomised controlled trials. Moreover, the few available clinical trial results are limited by ASD heterogeneity, suggesting the need to overcome the rigidity of the ”one-fits-all” control systems to cope with specific children.

Taking into account the main issues collected in the state of the art, this thesis initiated from a master thesis in which a robotic therapy was developed for children with ASD [5]. Through a strong collaboration with the clinicians, we decided to implement protocols for robotic mirroring therapy to train imitation skills



a)

Study or Subgroup	Experimental Mean	Experimental SD	Experimental Total	Control Mean	Control SD	Control Total	Weight	Std. mean difference IV, Random, 95% CI	A	B	C	D	E	F
Korte 2020	-70.87	26.7	21	-78.16	24.69	19	31.2%	0.28 [-0.35 , 0.90]	+	?	−	+	+	−
Marino 2020	7.29	1.16	7	5	1.31	7	18.1%	1.73 [0.44 , 3.03]	?	?	+	+	+	?
Pop 2017	13.18	5.01	12	14.77	3.39	15	28.0%	-0.37 [-1.14 , 0.40]	?	?	+	+	+	?
Yun 2017	82.5	24.54	8	90	19.15	7	22.7%	-0.32 [-1.34 , 0.71]	−	?	+	+	+	−
Total (95% CI)			**48**			**48**	**100.0%**	**0.22 [-0.52 , 0.97]**						

Heterogeneity: Tau² = 0.36; Chi² = 8.48, df = 3 (P = 0.04); I² = 65%
Test for overall effect: Z = 0.59 (P = 0.55)
Test for subgroup differences: Not applicable

b)

Study or Subgroup	Experimental Mean	Experimental SD	Experimental Total	Control Mean	Control SD	Control Total	Weight	Std. mean difference IV, Random, 95% CI	A	B	C	D	E	F
Yun 2017	0.88	0.13	11	0.86	0.17	9	57.2%	0.13 [-0.75 , 1.01]	−	?	+	+	+	−
Zheng 2020	77.92	18.47	8	73.81	12.79	7	42.8%	0.24 [-0.78 , 1.26]	?	?	+	+	+	?
Total (95% CI)			**19**			**16**	**100.0%**	**0.18 [-0.49 , 0.84]**						

Heterogeneity: Tau² = 0.00; Chi² = 0.03, df = 1 (P = 0.87); I² = 0%
Test for overall effect: Z = 0.52 (P = 0.60)
Test for subgroup differences: Not applicable

Risk of bias legend
(A) Bias arising from the randomization process
(B) Bias due to deviations from intended interventions
(C) Bias due to missing outcome data
(D) Bias in measurement of the outcome
(E) Bias in selection of the reported result
(F) Overall bias

Figure 2.7: Meta-analysis with the Randomized Controlled Trials selected. In a), it is presented the analysis of the ICF criteria d720 - Complex interpersonal interactions and in b) the analysis of the ICF criteria d160.

that are usually impaired in ASD children and are fundamental for developing other skills. Moreover, with the clinicians, the main requirements for the system were established:

- non-intrusive sensors: the sensors used should be non-intrusive, not disturbing the child and the therapy, given the hypersensitivity of the children with ASD.
- a triadic interaction: the therapist should always be present in the room and should participate in the activities that the robot proposes to the child
- holistic training: the protocol should train not just motor, but also social and cognitive skills.

Regarding the quantitative measures, motion and gaze measures are the ones of interest for the current thesis. We chose these measures because movements and gestures were the main focus of our therapy in all versions of our protocol, and the gaze measures provided an indirect way to describe the children's attention during the therapies which is important to evaluate their impact (see Chapter 1).

2.3 Technical Challenges

The intrinsic characteristics of our therapy and our main goal of obtaining quantitative measures generate different technological challenges. These challenges can be grouped in three sections whose current solutions in literature are presented in the next subsections.

Starting from the non-intrusiveness requirement, it puts a limitation in terms of the sensors that can be used. The gold standard for motion capture is the optoelectronic system [129]. This system is constituted by several high-resolution cameras, designed to detect and track retroreflective markers. However, it requires the positioning of the markers which can be intrusive for children with ASD. A direct substitute is the Inertial Measurement Units system, which have the same problem. That is why, in robotic-assisted therapies, cameras are used as main sensors. They are either integrated into the robot [40, 97] or independent [130, 21, 44, 131, 132]. Specifically for imitation training and motion measurement, depth cameras, like Microsoft Kinect, are chosen [42, 95, 67], since they can give the 3D joint positions (skeleton) of each identified person. However, the triadic free interaction between child and therapist creates a cluttered environment with several types of occlusions, hardening the reconstruction of the skeletons. There are many frames in which the skeletons are not reconstructed or are represented as a combination of the skeleton of the child and the skeleton of the therapist. Current methods for pose estimation and their respective limitations are shown in Section 2.3.1

After choosing the motion capture system, the main problem is the action accuracy definition, how to establish if a movement/action is correct or not. One of the main problems is the individuation of the action, the definition of the beginning and end [133]. Therefore, the aim is to identify a method for measuring gesture accuracy that is independent of the individual, their position, and any occlusions. We describe the methods already present in the literature in Section 2.3.2.

On the other hand, for the gaze measurement, there are two parts: eye-tracking, which consists of following the eyes' position frame-by-frame and the gaze-estimation in which the direction of the gaze in 3D is calculated and tracked. Initially, gaze measurement was done recurring to mounted systems [134], which we excluded due to the non-intrusive requirement established by our clinical partners. The gold standard nowadays is screen-based eye tracker such as EyeLink or Tobii. They are non-intrusive systems in which a screen presents the stimulus, and through image processing algorithms, the gaze direction is extracted. However, their working range is limited, being incompatible with our free interaction scenario [135, 136]. In Section 2.3.3, we delineate the state of the art regarding non-intrusive systems for gaze measurement.

2.3.1 Pose Measurement Methods

Given the preference for non-intrusive devices, depth cameras have been extensively used for robotic-assisted Therapies. One of the most popular is Microsoft Kinect being employed in several studies of Autism Spectrum Disorder [87, 88, 76]. Its working principle, called time-of-flight is based on the difference between the times of emission and reception of an infra-red ray after reflecting in a given object[137]. By knowing the velocity of the ray, the distance of the object can be extracted.

Kinect has its specific skeleton reconstruction method that produces a skeleton for each detected person. The skeleton contains 25 joints that are illustrated in Figure 2.8 and that are named keypoints. It has two main limitations. On the one hand, its accuracy varies with the depth [138]. On the other hand, the Kinect is highly sensitive to occlusions, namely self-occlusions [139].

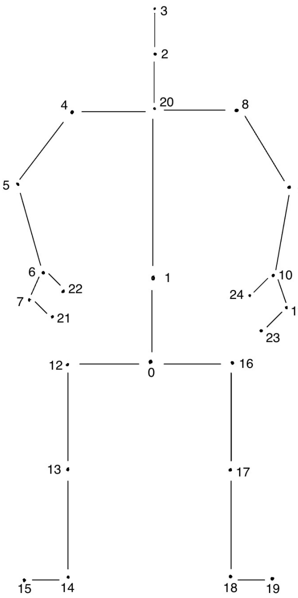


Figure 2.8: Skeleton format of the 25 joints Microsoft Kinect identifies.

Alternatively, monocular algorithms have gained more expression lately, focusing on this kind of occlusions. These algorithms guess the 3D poses of people present in 2D images. Two generic approaches are used: top-down or bottom-up [140]. In the bottom-up approach, possible joints are identified in the scenario and then they are grouped according to the different people [140]. In the top-down approach, each person is detected and then the 3D pose of each person is reconstructed. An example of a top-down approach is the Coherent Reconstruction of Multiple Humans (CRMH) [2]. Figure 2.9 shows the main blocks that constitute this system. In the first part, the bounding box of each person is predicted considering just a 2D image. After, in the reconstruction block, for each bounding box, the pose of the person is constructed. The authors assume a full perspective camera centred at each bounding box. Inspired by the work of [141], passing each bounding box image through a neural network, they obtain the camera parameters as well as several parameters that will be the input for a 3D model of the human body, the Skinned Multi-Person Linear Model (SMPL) [142]. This model trained in thousands of 3D body scans, departs from a Template mesh that is deformed through pose and shape parameters to obtain a final description of the person's pose. These parameters (pose and shape) are the ones guessed by the neural network described previously. Finally using the camera parameters, and a pre-set focal length f, each person i is put in the global scene position according to the translation vector in Equation (2.3).

$$t_i = \begin{bmatrix} \dfrac{d_i(x_i\alpha_i + c_{i,x} - \dfrac{w}{2})}{f} \\ \dfrac{d_i(y_i\alpha_i + c_{i,y} - \dfrac{h}{2})}{f} \\ \\ d_i \end{bmatrix}, \tag{2.3}$$

where $c_{i,x}$ and $c_{i,y}$ are the x and y coordinates of the bounding box centre, w is the width of the image, h is the height of the image, $\alpha_i = max(x_{max} - x_{min}, y_{max} - y_{min})$ is the size of the bounding box, f is the focal length, and d_i is the depth of each person calculated with:

$$d_i = \frac{\mu f}{s_i \alpha_i}, \tag{2.4}$$

where $\mu = 2$ is a multiplicative factor defined by the authors and s_i is an intrinsic parameter predicted by the model.

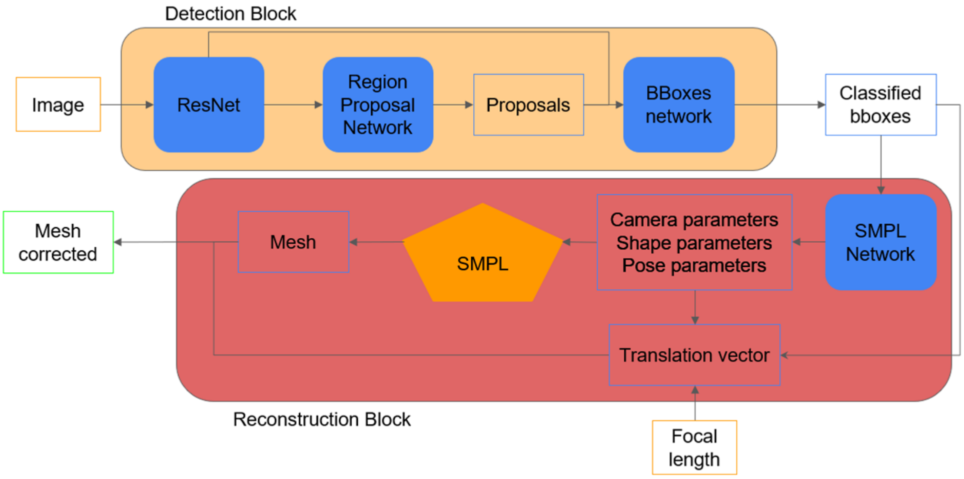


Figure 2.9: Block diagram illustrating the CRMH model process. The model takes an RGB image as input, and then detects individuals, providing bounding boxes around them. Following this, each individual is reconstructed through their skeletons, utilizing an additional input of focal length and the SMPL model. Ultimately, the process culminates in a 3D scene, delivering an accurate mesh representation of the individuals [2].

For obtaining the final skeleton from the mesh M given by the SMPL model, a regressor $J(\theta, \beta)$ is used according to Equation (2.5).

$$\text{3D skeleton} = J \times M \tag{2.5}$$

For training the SMPL-Parameters Neural Network, the authors used three different losses: a reprojection loss that minimizes the distance between the 2D ground truth keypoints and the projection of the 3D joints to the image; an interpenetration loss that penalises overlapping positions of the subjects; a depth ordering aware loss that uses instance segmentation annotation datasets to identify which person each pixel belongs to, guaranteeing that the depth of each person is estimated in the correct order. The largest disadvantage of this method comes from the SMPL model which was trained only with body scans of adults.

An alternative work that includes child body scans is the Bird-Eye View Model [3]. Contrary to CRMH, this method uses a bottom-up approach, considering all people in a single forward pass. For each image, three types of maps are predicted: Bodycenter heatmaps, Localization offset maps and Mesh Feature maps. Bodycenter heatmaps represent the probability of each pixel being a body centre. Localization offset maps integrate the information of the correct translation of each body centre. Two maps are produced for each of the two types mentioned before, representing the front view and a bird-eye view. The bird-eye view map allows an easier reasoning about the depth. The front and bird-eye view maps are expanded in depth and height and combined to generate a full 3D representation of the new Center and Offset maps, which correspond respectively to the Bodycenter and Localization maps. Then, for each person, a body centre is chosen from the Center map and put in the right location through the Offset map [3].

On the other hand, Mesh Feature maps contain the parameters that will be used by an adaptive version of the SMPL model (SMPL-A). In this model, if an age-related parameter is above a certain threshold, the SMPL model is used, if it is below, the Skinned Multi-Infant Linear model, a model trained specifically with children's body scans, is used [143]. The method overview is shown in Figure 2.10.

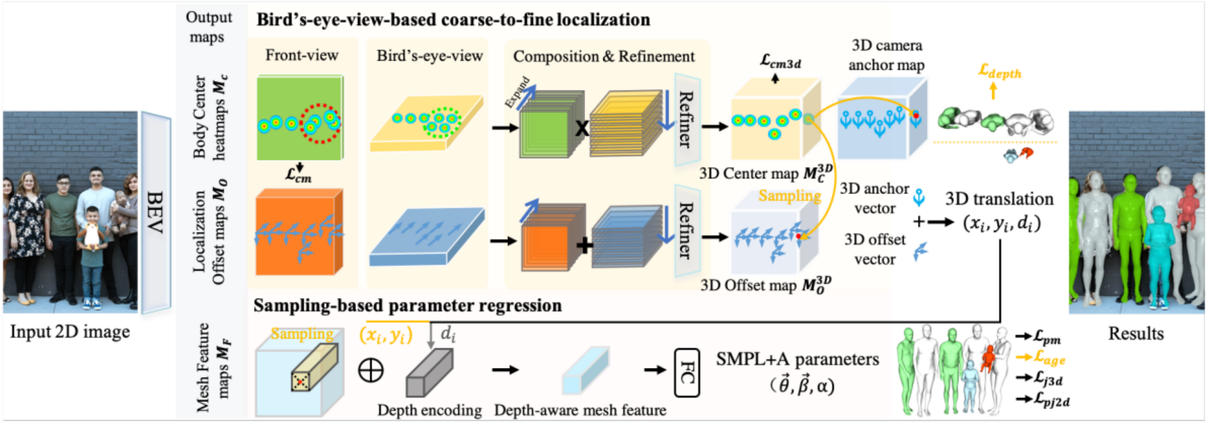


Figure 2.10: Overview of the BEV framework (extracted from [3]). Given an RGB image, BEV first estimates the 3D translation of all people in the scene via compositing the front view and the bird's-eye-view predictions. Then guided by the 3D translation, we sample the mesh feature of each person to regress their age-aware SMPL+A parameters [3].

Although the BEV and CRMH have been compared in terms of 2D reprojection in a standard dataset (Relative Human Dataset - see [3]), a comparison of the 3D reconstruction in real therapeutic scenarios is missing. Moreover, an evaluation of the computational time to verify the possibility of replacing completely the depth camera should be made.

2.3.2 Action Analysis from Video

Action recognition systems are generally constituted by two parts: Activity detection and Activity classification. Activity detection consists of defining the start and the end of the gesture. A possibility can be to define a fixed window in which it is considered that the action should be detected. An example is [144] which uses sliding windows with a fixed size to detect the presence of the gestures. This method's disadvantage is connected with the variability of starting moments of the gestures executed by the children, often different from the predefined window. [133] solved this problem using a binary classifier for each gesture and its respective score. They defined the action beginning when the sum has surpassed a given threshold and the action ending when the maximum sum has been reached. They tested this method in a dataset of actions (like a tennis serve or draw rick), using a Support Vector Machine Classifier, achieving an accuracy of 96%. Instead, [145] proposed the Kinetic Energy Method as a gesture descriptor, in which the increase would be associated with the beginning of the gesture and the decrease with the end of the gesture. They calculated the Kinetic energy according to the classical mechanical definition. The segmentation of the gesture was not done but they use this descriptor to recognize and identify the gesture inside a Dynamic Time Warping recognition system.

Regarding activity recognition, there are systems based on rules in which it is verified that the movement measurement is according to some predefined rules. For example, it is verified if the angles of the subject are between certain thresholds or not. However, these strict definitions are very sensitive to small variations of the chosen measure. In our case, small variations of the angle can prevent gesture identification. Machine learning methods can be used as alternatives such as Dynamic Time Warping [146, 147] or Hidden Markov Models [146, 148], or Artificial Neural Networks that have gained more popularity due to their performance in terms of accuracy.

As mentioned before since in robotic-assisted therapies the main sensors are cameras, we are particularly interested in computer vision algorithms for image processing. Given the pattern recognition capabilities of Convolutional Neural Networks (CNNs), they have been specifically utilized for gesture recognition. CNNs' greater advantage is the capability of feature learning, surpassing manual feature extraction processes required by other algorithms [149]. Through unprocessing images, CNNs can create feature extraction classifiers automatically. In this field, Residual Neural Networks (ResNet) have a central role due to their shortcut connections that allow them to train deep networks while solving the vanishing gradient problem. The shortcut connections skip the training of one or more layers, thus the gradient does not propagate layer by layer. An example of this specific type of Neural Network is given by Pham et al. [6], in which skeleton sequences are transformed into 3D arrays and consequently into images that are then classified by ResNet. Other works on the usage of more generic CNNs are from [150], [151], [152] and [153]. While the first two, [150, 151] use a strategy similar to Pham et al., converting the information of the motion of the skeletons into images that are then evaluated by the CNN, [152] prefer to apply a CNN directly on raw 3D coordinates, and [153] use 3D CNNs associate with convolutional Long Short-Term Memory networks. Most of these works achieve accuracies above 80% (only [153] reports an accuracy below) proving the effective performance of CNNs for gesture recognition.

However, effective implementation requires suitable datasets. Most datasets in the literature focus on gestures involving a wide range of motion [150, 154, 153], which differ significantly from those needed for our protocol. The existing datasets for small gestures were either captured using multiple cameras [155] or with a single camera at close range [156], conditions that are not ideal for therapies with children with ASD. In pilot studies in therapy settings, using multiple cameras can be challenging, as they increase setup time in sessions that typically last less than five minutes. Additionally, during regular therapy sessions, children often move around the room, making it impractical to use short-distance cameras that require specific proximity.

2.3.3 Gaze Measurement Methods

Previous works on attention analysis of children with ASD show a preference for qualitative assessment of attention. An operator manually evaluates the child's attention while watching the session in person or through a video [17, 82]. This type of work is tiresome and prevents comparing different robotic therapies. Instead, quantitative assessment of attention has become more common recently [47, 112, 108] and usually includes three parts: data acquisition, gaze estimation, and comparison of the gaze's estimation with the position of the targets-of-interest (named Area-of-interest (AOI)).

Regarding data acquisition, for the reasons presented in the other sections, most works choose non-intrusive devices, mainly cameras, for measuring ASD children's attention [47]. Usually, multiple cameras are synchronised to have a full assessment of child gaze [112, 108]. The setup complexity of these structures is not reasonable for the clinical environment. This complexity is justified by the characteristics of gaze pose estimators available in the literature. An example of a gaze estimator is OpenFace [157], that works exclusively when all the feature marks of the face are visible. Several times a proxy for gaze estimation is head pose estimation, such as WHENet, an algorithm that predicts the head pose in 360º [158]. It uses a convolutional backbone associated with multi-loss approaches, in which the loss functions are adapted to the wide range estimation. Another example of head estimator is RT-Gene [159]. This algorithm is based on Multi-Task Cascaded Convolutional Neural Networks, trained with information from a depth camera and motion sensors, allowing the prediction of a head direction.

Instead, Gaze360 is a specific gaze estimator that solves the 360º problem of gaze estimation, even when the face markers are not completely visible. This capability comes from its training on a large dataset collected both indoors and outdoors, featuring a wide range of head poses and distances [4].

Moreover, this algorithm uses Long Short-Term Memory networks, which means that for each frame, its inputs are the images of three previous frames and three forward frames. In this way, if the gaze is not visible in a specific moment, it can still be estimated based on the other frames. An overview of the full architecture is shown in Figure 2.11. The final output of this network is not just the gaze direction in spherical coordinates (azimuth and elevation) but also a measure of confidence that corresponds to the expected error bounds from the gaze prediction. Since the gaze direction is given in spherical coordinates with the reference frame centred on the eyes of each person, when the subject looks directly at the camera, the system outputs $0rad$ for both azimuth and elevation.

In all these algorithms, the eye tracking phase, described in Section 2.3, is replaced by a face tracking phase. In the WHENet algorithm this is done through the Yolo object detector [160]. In RT-Gene, it is the convolutional network that directly obtains face landmarks. In Gaze360, the 3D mesh estimator Densepose [161] is applied to segment the face.

Yolo combines two contemporaneous processes that start from an arbitrary grid: one that detects the object contours and another that classifies each portion of the arbitrary grid with an object label and a probability factor. Instead, Densepose obtains a 3D mesh from an image and it was trained with 5 million images with the corresponding 3D representation. The human body model used was again SMPL (Section 2.3.1). These differences between the face detectors cause large computation time differences in the gaze estimators, which may preclude the online use.

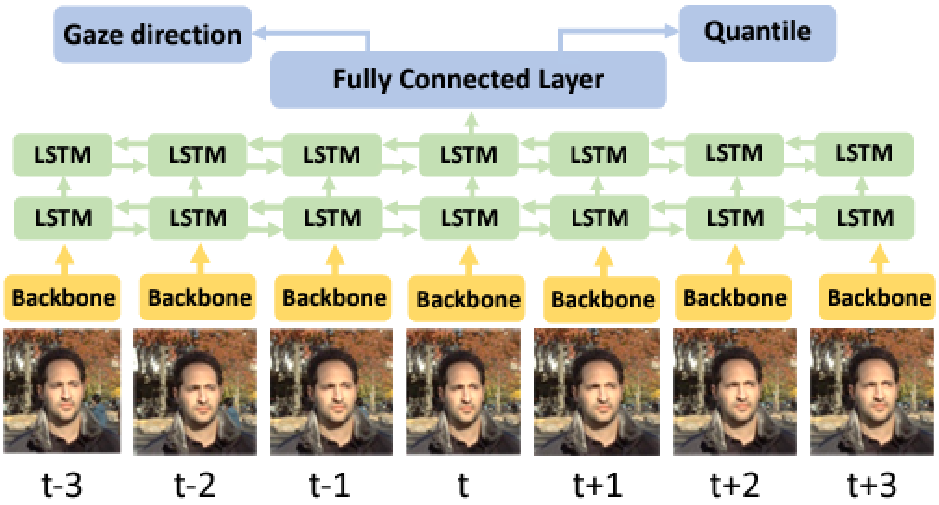


Figure 2.11: Full architecture of Gaze360. Each image passes through a backbone network (ImageNet - pretrained ResNet18), which produces a representation that goes to a bidirectional Long Short-Term Memory network. This network outputs a gaze direction and a quartile that represents the confidence in the gaze direction measurement [4].

Another characteristic of the scenarios for attention assessment is that they are constrained, with the child staying in a fixed position while interacting with the robot [78, 112, 108]. These constraints allow a better performance of the gaze pose estimators and an easier description of the targets that the child should look at. In this way, the gaze direction, usually in azimuth and elevation angles, can be compared to the angles that define the targets, such as the robot or specific points in the room (posters or other persons). However, these constraints were not considered feasible by our clinical partners to a real motor, cognitive and social therapy with the children. To our knowledge, the only work in an unconstrained scenario was the one proposed by [162]. They applied OpenPose and Gaze360 to determine the eye contact between a child and a therapist during a therapy session. The authors used the therapist as the only target of attention and did a 2D analysis of the attention. However, in our work, we want to include at least two targets, the robot and the therapist and profit from our 3D pose estimator, doing a 3D attention analysis.

2.4 Conclusions

It becomes clear that although robots have been used to assist therapy sessions in autism for a long time, the heterogeneity of the disorder makes the realization of statistically significant studies harder. This is one of the main drawbacks of the implementation in regular clinical practice. Even in Randomized Controlled Trials, conclusions can just be taken after the division of the initial group of children into several subgroups.

Moreover, there is an interest in having more automatic robots that adapt to children's characteristics. However, since robots are machines, they need automatic quantitative measures to react.

Extracting these metrics is challenging because gold standard methods for motion and gaze measurement often require wearable devices or the child needs to be fixed in certain positions. Therefore, these sensors are both intrusive and oblige a constrained scenario, being incompatible with the characteristics required for a rehabilitation session with autistic children. The main alternative sensors are cameras that lack in terms of information in depth. Nevertheless, more and more methods have been developed to extract the maximum information from the images acquired by them.

These methods are mainly focused on deep learning. Thus, they are highly dependent on the data used for their training. Most of them have never been used in children and even less in children with ASD. This will be one of the main gaps this thesis will try to fill, to arrive at our ultimate goal of developing new quantitative measures and flexible protocols for robotic therapies.

Chapter 3

Clinical Protocols

We explore the several protocols created for this thesis in an iterative process to incorporate the requirements of clinical specialists. The clinical staff initially included the Pediatric Neuropsychiatry team from Fondazione Don Carlo Gnocchi (FDG) in Milan Italy, a specialized centre on the rehabilitation of children with neurodevelopmental disorders. The Pediatric Neuropsychiatry team has a unit dedicated to new technologies for rehabilitation with whom we collaborated. After, we expanded this work to Portugal with the Associação Portuguesa para o Autismo e as Perturbações do Desenvolvimento (APPDA), who have an internal unit for adult rehabilitation and provide different kinds of therapies for children in schools. In a subsequent phase, we also joined the Centro de Apoio ao Neurodesenvolvimento (CADIn) in Portugal, which explores different therapies for children with neurodevelopmental disorders.

Departing from *the basic mirroring protocol*, four different protocols were created: the hierarchical protocol, the sensorial protocol, the scale protocol and the bingo protocol. Common characteristics are the robot's presence in a triadic interaction with the therapist and the child. The setup always included a Microsoft Kinect camera (nowon called Kinect) to register the whole session, a computer responsible for the control of the robot and sometimes a tablet, easier to carry by the therapist and that simply mirrored the interface present in the computer.

3.1 The basic mirroring protocol

Initially, our system had two sub-protocols: *Robot Coach* and *Adult Coach*. Both child and therapist participated in the clinical protocols, and their starting positions formed a triangle with the robot, shown in Figure 3.1 (a). Both protocols were designed as turn-taking games with the first round initiated by a different agent.

In the *Robot Coach* protocol, the NAO robot led the interaction (Figure 3.1(b)):

1. the robot showed a movement, then gave a "go" signal, pointing towards the adult and asking him to repeat the movement.

2. the robot had some time to start processing the recorded data, and then it mirrored the adult.

3. the robot gave the "go" signal to the child, who executed the gesture while NAO was mirroring.

4. If the movement was performed correctly, NAO gave positive vocal feedback ("Bravo" or "Grande", i.e. "Great") and its LEDs turned green. Otherwise, if the adult or the child did not finish the exercise within 20 seconds, the LEDs on NAO eyes became red. This time limit was determined from clinical experience.

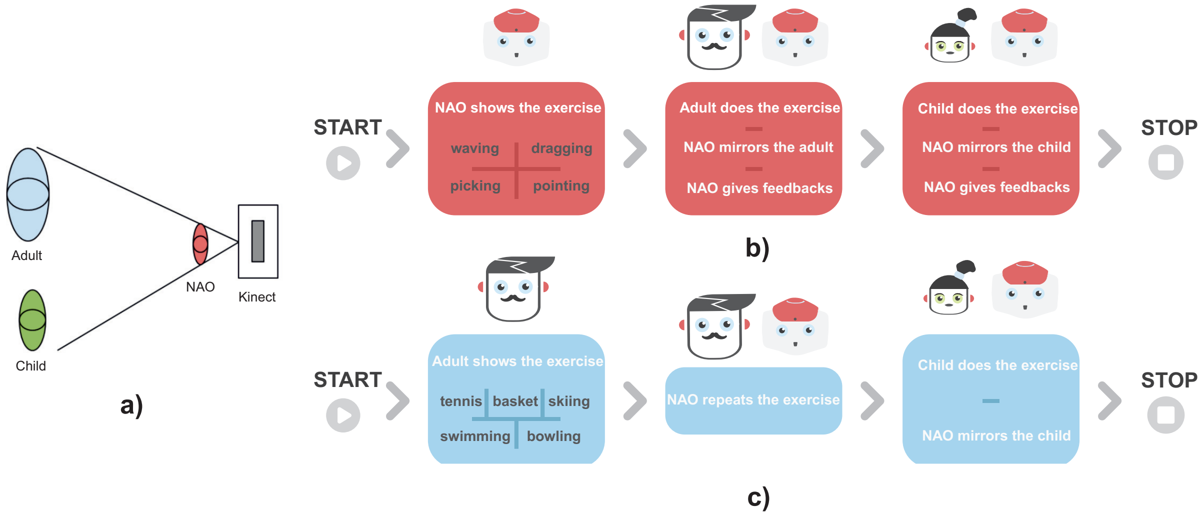


Figure 3.1: Diagram of the triadic interaction (adult-child-robot), illustrating the geometry between the participants and the perception system (Kinect) used during a protocol session (a) and block diagrams of the two games protocols. (b) *Robot Coach* protocol: first the NAO shows the exercise, then the adult repeats it while the robot is mirroring, and finally the child does the exercise, mirrored by the robot. (c) *Adult Coach* protocol: the adult leads the game, showing the exercise, which is then repeated by the robot and finally by the child while the robot is mirroring. Adapted from [5].

There were four exercises incorporated into a story aimed at training both cognitive and social skills. In this narrative, NAO was portrayed as a being from another planet. In the first exercise, *waving*, NAO introduced itself by saying, ”Hi, I am NAO!” while performing a greeting gesture. Next, in the *dragging* exercise, it moved the clouds away by simultaneously sweeping both arms from the top left to the bottom right. Then, during the *picking* exercise, it sequentially picked stars from the sky by moving each arm in turn. Finally, in the *pointing* exercise, it pointed to its planet. Each of these three exercises was accompanied by a phrase that the robot articulated while performing the corresponding movement, similar to the *waving* exercise.

For each movement, positive feedback was given whenever the current pose joint angles p_c reached the target angular position p_t in each subphase of the movement, for the most significant degrees of freedom, with a certain range of variability, α (Equation 3.1):

$$||p_t - p_c||_\infty \leqslant \alpha \tag{3.1}$$

For the completion of the exercise and reception of the final vocal feedback, the subject had to accomplish 6 subphases in the case of the waving (3 inward rotations and 3 outward rotations), 2 subphases in the case of the dragging and 4 in the case of the picking, as described in Table 3.1. The pointing exercise contained one single subphase. The target poses and the range of variability were found using a previous database of 28 neurotypical adults.

In the *Adult Coach* protocol, the adult was the master:

1. the adult demonstrated the movement for a certain time interval (7s), asking NAO to replicate the movement;
2. NAO replicated the movement;
3. the child performed the movement, while NAO mirrored him/her (Figure 3.1 (c))

No feedback was given by the robot, because the adult, e.g. the therapist, was supposed to provide it, leading the interaction. In this case, the exercises chosen were sports that were already worked

Table 3.1: Target poses and ranges of variability (α) used in the feedback system of the *Robot Coach* protocol.

Movements	Target Pose	Final Target Angles (rad)				α (rad)
		Elbow Roll				
Waving	Inward Rotation	1.19				0.3
	Outward Rotation	0.29				
		Shoulder Pitch		Shoulder Roll		
		Right	Left	Right	Left	
Dragging	Hands up	-0.54	-0.63	-0.54	0.10	0.3
	Hands down	1.2	1.3	-0.2	0.42	
		Shoulder Pitch		Shoulder Roll		
Picking	Right Hand up	-0.87		0.14		0.3
	Right Hand down	1.37		0.14		
	Left Hand up	-0.91		-0.37		
	Left Hand Down	1.34		-0.16		
		Shoulder Pitch				
Pointing	Hand up	-0.96				0.2

in regular therapies: basketball (basket), tennis, bowling, skiing and swimming. For the basket, the movement was throwing up a ball with both hands above the head; for tennis, hitting a ball laterally; for bowling, throwing a ball forward with one arm; for skiing, sliding the arms;for swimming, breaststroke movements.

In this protocol, we used a Kinect camera which recorded the scene and captured the 2D and 3D joints positions of the child and the adult. The interface for the choice of the exercise was done in a computer. During the protocol, child and adult were identified by their pose: the child was the person whose left hand was more at the left of the Kinect; the therapist was the other person.

We have tested these two protocols with two primary school children with autism. The two children had different levels of the disorder, thus different needs of support. We verified that the child with a lower level of autism (the one who needed less support) reacted better than the other child. This different reaction was attributed to the high complexity of the exercises for the child with a higher level of autism. The results of this initial study are presented in [5].

3.2 The hierarchical protocol

Departing from the conclusions in the last section, our new protocol included two familiarisation levels for the children to understand the setup before passing to the training levels. The construction of several levels allowed the therapist to adapt the protocol to the child's behaviour during each session. Another novelty of this protocol was the focus on intransitive gestures. These gestures convey socio-communicative intent, and their recognition is particularly difficult for children with Autism Spectrum Disorder (ASD) compared to gestures involving objects [13]. Moreover in addition to the Kinect and the computer, the therapist used a tablet which replicated the interface present in the computer. The tablet

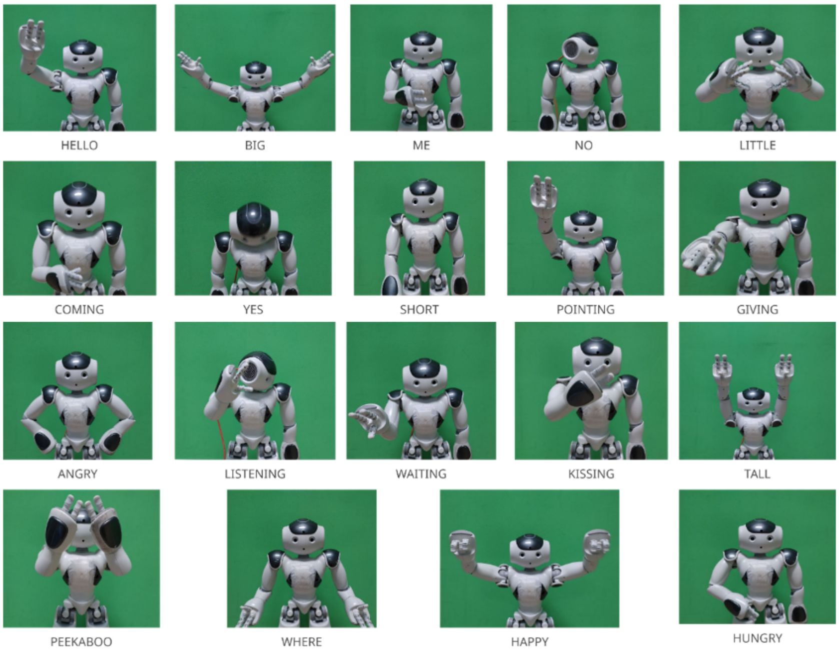


Figure 3.2: Robot NAO performing the 19 gestures of the robotic-assisted therapy

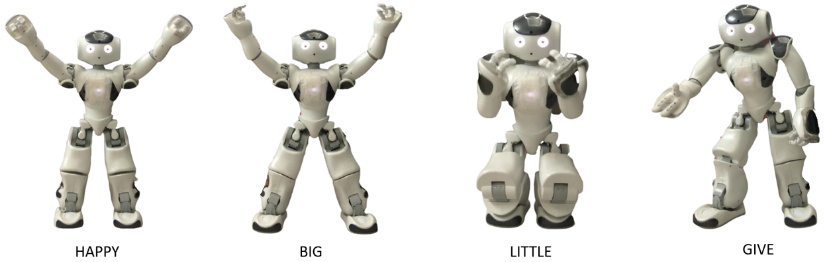


Figure 3.3: Total body gestures included in the Hierarchical Protocol.

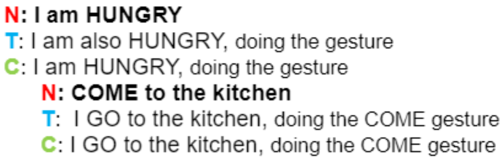
was easier to handle and facilitate the contact between child and therapist. In this protocol the therapist worn a piece of cloth with a certain colour that distinguish her/him from the rest of the scene and was tracked during the whole session.

3.2.1 The five levels

We divided the new protocol into the following five levels:

Level 1 - Acquaintance: This level represents the first contact of the child with the robot. From the moment the child enters the therapy room, the robot shows its functionalities using simple phrases, movements of upraising and sitting and its different lighting modes. We also included some songs to engage the child (e.g.: ”If you're happy and you know it”), a piano keyboard activated by the child's touch on the sensors of NAO, and a first introduction to the gestures that are part of the training levels. In this level, the robot presents the gesture associated with a phrase (e.g. the *Hello* gesture is accompanied by the sentence 'Hi, I am NAO'). The gestures chosen were: *Hello*, *Big*, *Me*, *No*, *Little*, *Coming*, *Yes*, *Short*, *Pointing*, *Giving*, *Angry*, *Listening*, *Waiting*, *Kissing*, *Tall*, *Peekaboo*, *Where*, *Happy*, *Hungry* (Figure 3.2).

All gestures were initially implemented using just the upper limbs since stability constraints should be considered when dealing with total body gestures. Nevertheless, the gestures *Big*, *Little*, *Giving*, *Happy* were also tested with the lower limbs to enhance their communicative meaning [163] (Figure 3.3).



N: I am HUNGRY
T: I am also HUNGRY, doing the gesture
C: I am HUNGRY, doing the gesture
    N: COME to the kitchen
    T: I GO to the kitchen, doing the COME gesture
    C: I GO to the kitchen, doing the COME gesture

Figure 3.4: Example of dialogue in Level 4 (N: NAO; T: Therapist; C: Child).

Level 2 - Mirroring: During Level 2, the child learns how the mirroring system works. The therapist uses a red shirt, which is recognized by the Kinect and allows the correct identification of the therapist and child skeletons. The robot mirrors just the upper limbs of the child. During this level, the therapist performs several dances that are imitated by the child, who sees his/her direct reflection on the robot.

Level 3 - Single Gesture Training: Level 3 is the first level of training. In this level, the two training modes designed previously (*Robot Coach* and *Adult Coach*) are applied with the new gestures and the new phrases established between the robot, the child and the therapist. In the *Robot Coach* Protocol, the robot gives feedback to the therapist and the child if they executed the gesture correctly. The therapist also has a button in the interface to reward the child in other moments. Contrarily to the basic mirroring protocol we did not give any feedback for a negative performance.

Level 4 - Contextualized Gestural Training: In Level 4, the gestures learned are implemented in specific scenarios known to the child. The two training modes were integrated, and 5 scenarios were created: kitchen, bedroom, school, train, and beach. The scenarios were shown on the computer, tablet or on the wall through a projector. In general two gestures are used for each scenario creating a dialogue between the robot, child and therapist. An example of dialogue in the Kitchen scenario, for the *Robot Coach* Protocol, is presented in Figure 3.4. In this level, the robot also gives feedback if the gesture is correctly performed.

Level 5 - Generalizing Gestural Training: This final level is characterised by a new training mode, the *Child coach*. In this case, the initiative should depart from the child after re-watching the scenarios of Level 4. It is expected that the child reproduces some of the gestures learned. The robot has predefined phrases, established by the therapist, that would be executed with the respective gestures. This level should train the capability of the child to generalise the gestures learned while maintaining the triadic interaction between the child, the robot and the therapist.

This protocol was tested first with neurotypical children and Adults in Polimi, and then with 3 Adults with Autism in APPDA. Subsequently, two pilot studies were run, one in FDG with 10 children with ASD and one in APPDA with 6 children with ASD. Ultimately, a randomized controlled trial was run in CADIn with 11 children with ASD in the experimental group and 11 in the control group. The protocol of the control group was the same as the experimental group but the robot did not participate, although it was present in the room. In the control protocol, all actions were performed by the therapist.

3.3 The sensorial protocol

The motivation to develop this protocol was to enrich the variety of stimuli involved. Some of these stimuli were already included in the hierarchical protocol during Level 1 of the familiarization phase. These proved essential for certain children in the pilot study of FDG who never progressed beyond this level. Therefore, they were incorporated and expanded in the new protocol. The overview of the sensorial protocol is presented in Figure 3.5. The main difference from the hierarchical protocol is its single level, allowing the therapist to choose different types of stimuli independently. The several modalities included:

light, sound or movement. We describe each of them in detail in the next paragraphs.

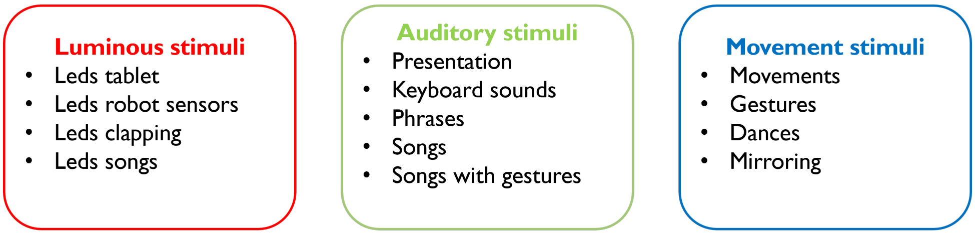


Figure 3.5: The sensorial protocol with its three parts: luminous stimuli, auditory stimuli and movement stimuli.

Luminous stimuli

For luminous stimuli, we take advantage of the LEDs present on the eyes of the robot. First, these LEDs can be activated through an interface (*Leds tablet*) as in Level 1 of the previous protocol. These LEDs present a white colour. In another game, the robot sensors of the hands, feet and head can be used to activate different colours on the eyes of the robot (*LEDs robot sensors*). In this way, the child can explore the different body parts of NAO. Other new functionality involved clapping to activate the LEDs (*Leds clapping*) with the same colour. NAO's microphone detects a peak in the sound signal, activating the LEDs. When the child's engagement should be increased, the LEDs are activated repeatedly with well-known songs, decided by the therapist (*Leds songs*). This last game also tests the tolerance of the children to two simultaneous stimuli (luminous and auditory).

Auditory stimuli

Regarding the auditory stimuli, first, the robot presents itself as in Level 1 of the previous protocol (*Presentation*). Then, similarly to the luminous stimuli, the child can explore the body parts of the robot while it produces a different sound for each sensor of the robot (*Keyboard sounds*). There is also an input field where the therapist can write what the robot will say at every moment (*Phrases*) to facilitate the system's adaptability, namely the communication with the child. In another functionality, to engage the children, but without giving extra stimuli, the songs chosen for the luminous stimuli are here presented alone (*Songs*). The last functionality involves two stimuli together: auditory and movement stimuli. In this case, the same songs are presented with specific gestures designed by the therapist (*Songs with gestures*). The gestures chosen have a semantic meaning as the gestures of Level 3 of the previous protocol, although they are completely different from the preceding gestures. The therapist can stop the robot through a head sensor, allowing the child to complete the song.

Movement stimuli

For the last type of stimuli, as in the other stimuli, the robot presents simple movements (*Movements*), like standing up and down, walking forward, etc. Then the gestures of the previous Level 3 are presented. From the hierarchical protocol, we selected the gestures considered more significant by the therapists: *Hello*, *Pointing*, *Yes*, *No*, *Giving*, *Happy*, and *Angry*. The gestures do not require a turn-taking game or mirroring, making it easier for the children to replicate them. The mirroring was transposed for a specific part of the system (*Mirroring*), in which the robot mirrors purely the movement of the child, as in Level 2 of the hierarchical protocol. Alternatively, the robot presents a song with the respective gesture, and after each stanza, the music stops and the robot mirrors the child's gestures. The song continues after the therapist activates the head sensor of the robot. The songs and gestures are the same as the previous stimuli. An additional point, just for the engagement of the child was the introduction of a simple dance with movements that do not have a semantic meaning (*Dances*).

At the time of the conclusion of this thesis, this protocol is being tested in a Randomized Controlled Trial with 20 children with ASD.

3.4 The scale protocol

We have previously mentioned subjectivity as the main problem of current clinical scales, since they always depend on an operator to classify them. Nevertheless, they are the main instrument of clinicians and therefore essential for the integration of a technology in the clinical practice. For the establishment of a clinical outcome measure, the Early Social Communication Scale (ESCS) was the clinical measure more related to the goals of our protocols, especially to the development of joint attention. From the 17 tasks of the ESCS, we selected the three that evaluated directly Joint Attention: the object spectacle task, the gaze following task and the book presentation task [32]. We included these three tasks in a new robotic protocol so that our evaluation protocol complied with our therapeutic scenario.

For this protocol, a table was used similarly to the ESCS setup and the therapist and the child sat according to Figure 3.6. The robot was put in front of both without occluding the faces of the participants from the camera that registered the whole scene. The therapist starts by exploring the robot with the child. After the robot presents itself, the object spectacle task begins. In this case, as in the ESCS, the robot points to a mechanical toy placed on the table asking the child if he/she wants to play. Then the therapist activates the mechanical toy letting the child play freely during a period. After this period, the therapist activates the robot again, turning on its LEDs and the robot tells the child to give back the mechanical toy to the therapist. Next, there is the gaze following task. In this task, first, the robot turns its head to different posters present in the room. The posters are displayed on three different walls as represented in Figure 3.6. Secondly, the robot while pointing also calls the child's attention with a simple phrase ”Look how beautiful it is!”. This task is repeated with the three different posters. Following, in the book presentation task, the therapist shows two books and the robot points to her asking the child ”What book do you want?”. The child chooses one of the books and explores it with the therapist.

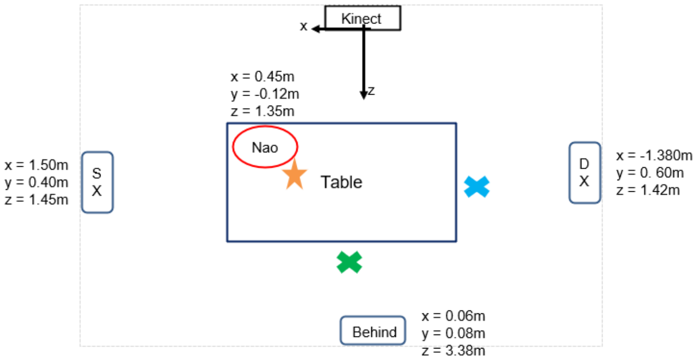


Figure 3.6: The scale protocol setup. The orange star represents the mechanical toy; DX (right), SX (left) and Behind are the three posters on the wall. The blue x represents the therapist, while the green one, the child.

In addition to the tasks based on the ESCS, we developed a song task and a body recognition task. These tasks were included to evaluate the child's attention in different types of play and interaction. In the first, the child interacts with the robot while singing and dancing to the song ”If you're happy and you know it". In the second, the robot asks the child where a body part is and then points to it. In the end, the robot gives positive feedback to the child and says goodbye. A full overview of the protocol is represented in Figure 3.7.

From this protocol, we just tested the gaze following task with 4 different neurotypical adult pairs. The therapist role was always performed by the same person.

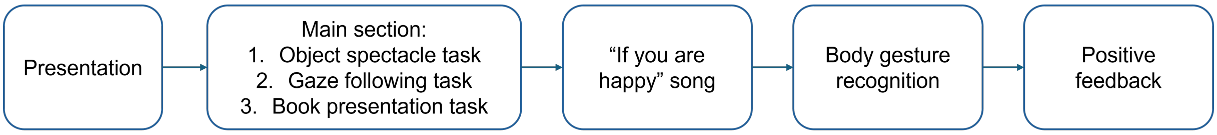


Figure 3.7: The scale protocol's overview

3.5 The bingo protocol

The bingo protocol was created as a different path from the experience with the hierarchical protocol in CADIn Portugal. The main idea was to integrate the robotic sessions directly with the usual therapeutic sessions of children in this institution. First, we reduced the number of gestures of the protocol to the more essential ones: *Big*, *Little*, *Me*, *Hello*, *Giving*, *Pointing*, *Yes*, and *No*. As in the sensorial protocol we have created a single-level protocol divided into moments. The flow of the protocol was chosen by the therapists and personalised for each child. A general overview is provided in Figure 3.8.

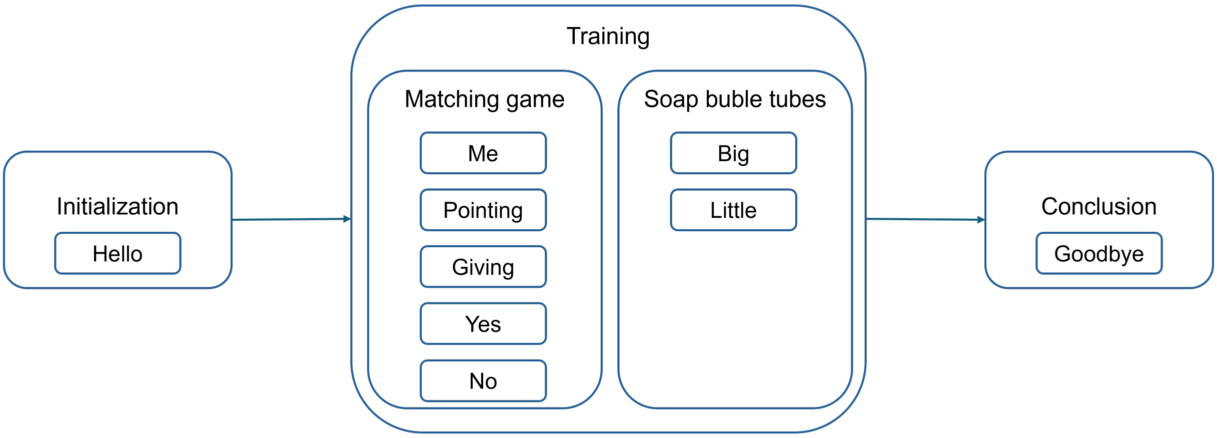


Figure 3.8: Block Diagram illustrating the flow and structure of the bingo protocol, emphasizing the key gestures and stages of interaction.

The protocol started with an Initiation phase in which the robot greets the child and the therapist while doing the corresponding gesture: *Hello*. After, two different scenarios were designed. For the first scenario, three boards were printed, representing three themes: food, animals and cartoon characters. Each board was accompanied by corresponding cards that needed to be matched with the images on the board, resembling a bingo game. An example of a board is shown in Figure 3.9.

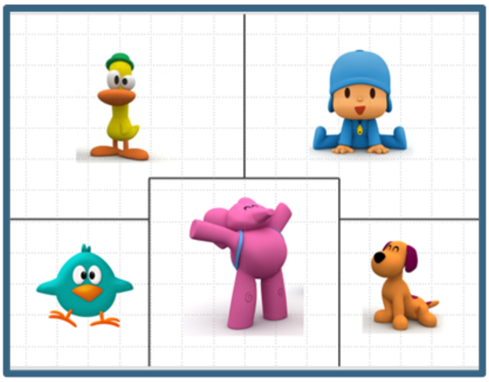
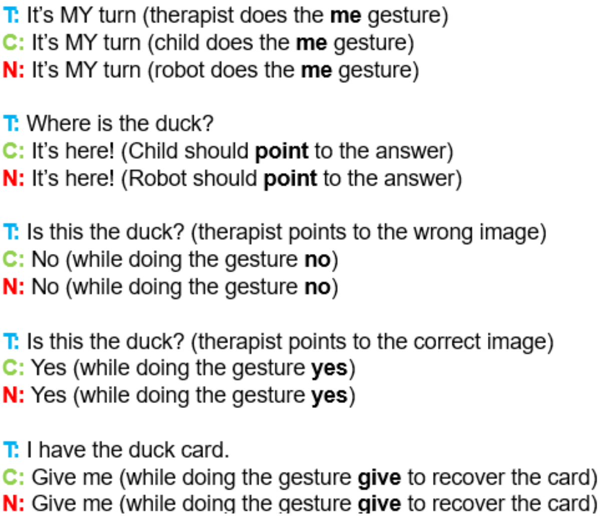


Figure 3.9: Illustrative Paper with Cartoon Theme. This figure showcases the selection of a theme suitable for the child in the Bingo Robotic Game. The therapist has laid out an illustrative paper on the ground, featuring engaging cartoons that correspond to the chosen theme, which could be food, animals, or cartoon characters.

Following the choice of the theme, the action departs always from the therapist who performs a gesture followed by the robot and the child. All the gestures are integrated into the bingo game. A possible succession of gestures and phrases said by the therapist is presented in Figure 3.10



T: It's MY turn (therapist does the **me** gesture)
C: It's MY turn (child does the **me** gesture)
N: It's MY turn (robot does the **me** gesture)

T: Where is the duck?
C: It's here! (Child should **point** to the answer)
N: It's here! (Robot should **point** to the answer)

T: Is this the duck? (therapist points to the wrong image)
C: No (while doing the gesture **no**)
N: No (while doing the gesture **no**)

T: Is this the duck? (therapist points to the correct image)
C: Yes (while doing the gesture **yes**)
N: Yes (while doing the gesture **yes**)

T: I have the duck card.
C: Give me (while doing the gesture **give** to recover the card)
N: Give me (while doing the gesture **give** to recover the card)

Figure 3.10: Example of dialogue in Bingo robot (N: NAO; T: Therapist; C: Child).

Neither the presence nor the order of the gestures is mandatory and the therapist can adapt it to each child.

In the second scenario, instead of a bingo board, two bubble soap tubes of different sizes are used by the therapists to train the gestures big and small. The therapist asks the child and robot which type of bubble soap they want and they should answer with the respective gesture. After this game, the robot concludes the session by saying goodbye to the child. At the time of the conclusion of this thesis, this protocol is being tested with 5 children with ASD.

3.6 Conclusions

Five protocols are presented in this chapter. The last three were developed almost in parallel and derived from the hierarchical protocol. Common characteristics are the camera, the robot and the triadic interaction between the participants. A general conclusion is that therapists from both countries, after testing the hierarchical protocol, preferred a single-level protocol to adapt faster to the children's mood. The type of gestures was also simplified with respect to the hierarchical protocol and they were integrated inside more contextualized activities like the dances in the sensorial protocol or the cards in the case of the bingo protocol.

Essential for the development and improvement of the protocols were the metrics and respective results which will be shown in the next chapters. In total, 11 therapists and 33 children were involved in the tests.

Chapter 4

Action Analysis from video

In this and the next two chapters, we describe the main algorithms developed during this thesis, focusing on overcoming the limitations found during the literature review. We concentrated on the action and attention performance measures since they were the most important to our several protocols.

Relatively to the action performance, the imitation is present in several of our protocols therefore we started by defining a mirroring measure to evaluate the performance of the participants (Section 4.1). After, our protocols changed focus to intransitive gestures and, consequently, our measures. Especially, for the hierarchical protocol, a gesture recognition system was required to provide immediate feedback to the children. We have developed this system first offline and then online (Section 4.2). For each metric we present the respective results in neurotypical children and children with Autism Spectrum Disorder (ASD).

4.1 Mirroring measure

In the basic mirroring protocol, we wanted a performance measure to understand the differences on the execution of the exercises by the different children. Moreover, we were interested in an indicator of the complexity of the exercises to choose the best exercises for each child. Therefore, we decided to address the following research questions:

1. Which protocols are better for mirroring?/Which exercises are more complex?
2. How can the robot evaluate the movements executed by a person?
3. Which metrics are better for inter-subject overall comparison over time? (Which movements are more standard among subjects and allow their comparison? Which performance metrics can be used to evaluate a subject continuously?)

We have used the basic mirroring protocol and the two developed subprotocols (*Robot Coach* and *Adult Coach*) and extracted new quantitative measures to evaluate the level of difficulty, repeatability and mirroring of the movements chosen.

4.1.1 *Robot Coach* Protocol

As seen in Section 3.1, in the first protocol there were four exercises related with the story of the robot: *Waving*, *Dragging*, *Picking*, *Pointing*. For the calculation of the performance metrics, we considered the 3D coordinates extracted by the Kinect and the most important control angles given to the robot in each exercise, presented in Table 4.1. All performance metrics aimed to evaluate the difficulty of execution, quantifying the exercises' complexity, related to our first research question. In addition this protocol allowed us to test our feedback system, which was related with our second research question.

Table 4.1: Control angles of each exercise of the *Robot Coach* protocol.

Movements	Control angles
Waving	Elbow Roll
Dragging	Shoulder Pitch and Shoulder Roll
Picking	Shoulder Pitch and Shoulder Roll
Pointing	Shoulder Pitch

We established four performance metrics:

- latency: time interval between the robot "go" signal and the time instant when the person starts the exercise. We used the keypoints of the hand and the wrist to compute the exercise's starting instant.
- duration: exercise overall duration, calculated from the angles. Just exercises concluded in less than 20 seconds were considered for this computation.
- number of failures: number of uncompleted exercises, all the exercises that had a duration larger than 20 seconds.
- number of attempts: number of trials a person should do before reaching the target pose (Table 3.1).
- first positive feedback time: time interval until receiving the first feedback.

We calculated these last three outcome measures using the angles given to the robot for the mirroring.

4.1.2 *Adult Coach* Protocol

Unlike from the *Robot Coach* protocol, in the *Adult Coach* protocol, just the child is mirrored online by the robot. Moreover, child and adult could repeat the same movement several times. To segment the movement repetitions, after selecting the most significant joint for each exercise (e.g. shoulder roll for tennis), we found the frames in which the signal was above a calculated baseline, had a peak in between and returned under the baseline value. This baseline was determined for each exercise, considering a percentage of the range between minimum and maximum angle of the movement. For each exercise, this empirically defined percentage was between 20% and 55%.

After extracting the movement repetitions, we needed measures that reflected the standardisation of the exercises and their *mirroringability* to answer to our first and third research questions. Since the movements were more complex in this protocol, we considered features for both the single-subject repetitions or paired adult-child repetitions. Paired adult-child repetitions reflected that in this protocol the mirroring started from the adult and then was done by the child.

From the single-subject repetition, we calculated both the amplitude and half-width duration of each segmented signal for the adult and the child and normalized them by the maximum value for each sport to understand better their variability. From the paired repetitions, we calculated the correlation between paired subject movements and extracted the maximum correlation and respective lag. In summary, we had three sets of features (Figure 4.1): the children set, the adult set and the paired set represented in the following equation.

$$\begin{aligned} C_j &= \{R_1^{C_j}, ..., R_{N_j}^{C_j}\} \\ A_k &= \{R_1^{A_k}, ..., R_{N_k}^{A_k}\} \\ P_{k_j} &= \{R_1^{C_j} R_1^{A_k}, R_1^{C_j} R_2^{A_k}, ..., R_{N_j}^{C_j} R_{N_k}^{A_k}\} \end{aligned} \tag{4.1}$$

where N_j is the number of repetitions R of the movement for child C_j, N_k is the number of repetitions R of the movement for adult A_k and P_{k_j} are the paired repetitions.

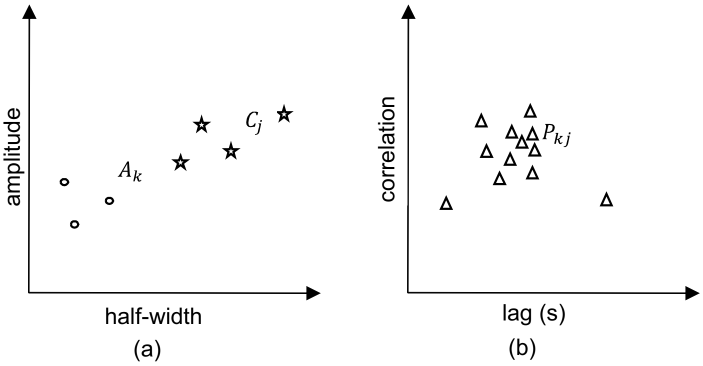


Figure 4.1: Example of a set of observations using (a) single subject analysis and (b) paired analysis. In (a), circles and stars indicate the data points for one example adult and child, respectively; in (b) triangles indicate data points for one example pair.

After the computation of the features, we defined three quality criteria for each exercise:

- Movement variability - evaluation of the variability of each exercise for the group of children and for the group of adults: we constructed two new sets $C = \bigcup_{j \in N_j} C_j$ and $A = \bigcup_{k \in N_k} A_k$ with all the repetitions of the children and the adults respectively. We calculated the mean distance of the elements of each set to its centroid to have a measure of variability. We named this metric the *intraset distance*. The distance chosen was the Euclidean distance. The lower this distance, the less variable the set and the more repeatable the exercise among the categories that have tried it.

- Repeatability among pairs - evaluation of the repeatability of each exercise considering the pairs adult-child: Different sets were composed using the paired repetitions of the children and adults in the same session (P_{jj}). Again the *intraset distance* was calculated for each set. The mean of this last metric across all sets represented the repeatability of each exercise. A lower mean was related to lower intraset distances, meaning less variable exercises among the different pairs.

- *Mirroringability* - assessment of the ease of mirroring another person for each exercise: we constructed the sets with paired repetitions of the child and several repetitions of each adult, P_{kj}. Then, we calculated the *intraset distance* between the child and respective adult and compared it with the mean of the *intraset distances* between the child and all the other adults. When the lowest value of *intraset distance* was between the child and respective adult, they had a lower variability of correlation and lag between them compared to the other combinations of the same child and the other adults. They were mirroring each other since the child's movement was more similar to his/her respective adult than to the other adults.

4.1.3 Results

The results presented regard acquisitions done with 30 individuals, 14 neurotypical children (between 5 and 10 years old) and 14 adults close to each child (usually the parent), naive to the whole system. The remaining two individuals were a 5-year-old girl diagnosed with ASD and her therapist, experienced with technology-mediated treatments and already familiar with the system. This last acquisition allowed the assessment of our system in a clinical setting.

Robot Coach Protocol

The results for the first protocol are presented in Tables 4.2 to 4.4. Regarding the neurotypical participants, we noticed that children and adults had similar performance metrics. Overall, the standard deviations are large, reflecting the variability of each person participating in the protocol. For the children, the most difficult exercises were the *Waving* and the *Pointing*. For the adults, the most difficult exercise was the *Waving*, justified by the fact that it was the first exercise to be executed, during the whole experiment. Both exercises, *Waving* and *Pointing*, required just one arm. The participants often started with the ipsilateral arm instead of the contralateral arm. In this way, the mirroring exercise was incorrect and the participants took extra time to complete it.

The therapist's performance metrics are much better than the other adult participants. Her lower values compared to the mean of the other participants were expected since she was already familiar with the technology. On the other hand, the ASD child's latency and duration are above the mean of the neurotypical children. The feedback system did not work as expected since it depended on several consecutive target poses. If one failed due to an erroneous pose, there would be postponed feedback.

Table 4.2: Latency times of the Therapist and ASD child and mean latencies of the neurotypical adults and children for the exercises in the *Robot Coach* protocol. The zero values represent situations in which the subject was already in the starting position at the onset of the analysis.

	Mean±SD Adults	Therapist	Mean±SD Children	ASD child
Waving (s)	4.45±5.10	0.00	3.76±3.77	31.25
Dragging (s)	0.79±0.91	0.00	1.18±1.67	3.57
Picking (s)	1.06±1.43	0.26	1.69±1.20	11.20
Pointing (s)	1.21±1.39	0.00	4.26±4.22	5.08

Table 4.3: First positive feedback times (FPFT) and Number of attempts until the first feedback of the Therapist, ASD child, neurotypical adults and children for the exercises in the *Robot Coach* protocol.

	Mean±SD Adults	Therapist	Mean±SD Children	ASD child
Waving				
FPFT (s)	5.37±4.64	0.30	5.32±3.91	6.31
# of attempts	2.93±3.43	1	2.79±1.89	5
Dragging				
FPFT (s)	2.21±1.15	0.00	2.22±1.89	3.98
# of attempts	1.79±1.25	1	2.71±4.07	2
Picking				
FPFT (s)	0.89±1.29	1.05	2.33±1.37	11.54
# of attempts	1.50±1.02	1	1.36±0.74	6
Pointing				
FPFT (s)	6.64±3.00	3.02	10.86±6.96	8.12
# of attempts	1.36±0.63	1	2.79±1.85	3

Table 4.4: Duration of successful trials and number of failures for the Therapist, ASD child, neurotypical adults and children in the exercises of the *Robot Coach* protocol. The empty cells are exercises not completed in the preset time.

	Total/ Mean±SD Adults	Therapist	Total/ Mean±SD Children	ASD child
Waving				
# failures	7	0	6	1
Duration(s)	12.72±5.94	2.75	10.27±3.84	-
Dragging				
# failures	0	0	3	1
Duration(s)	6.79±4.25	3.57	6.50±3.24	8.71
Picking				
# failures	5	0	3	1
Duration(s)	9.81±3.90	6.77	12.08±3.89	-

Adult Coach Protocol

In the *Adult Coach* protocol, children and adult movements are different by design. In this protocol, the adult first executes the exercise, then the robot repeats the exercise, replicating the adult's movements. In the child's turn, the robot mirrors the child's exercise simultaneously.

Table 4.5 shows that the children's movement variability was always larger than in the adults. The most variable movements were the tennis, the bowling and the basket for the adults and the neurotypical children. Similarly, in the ASD child/therapist pair, the child variability was always larger than the therapist except for the ski exercise.

Table 4.5: Movement variability: *Intraset distance* in the children set, adults set, therapist and ASD child sets for each sport of the *Adult Coach* protocol.

	Tennis	Swimming	Bowling	Ski	Basket
Adult set	0.17	0.16	0.27	0.14	0.26
Child set	0.30	0.19	0.27	0.21	0.28
Therapist	0.16	0.18	0.00	0.09	0.05
ASD child	0.26	0.24	0.15	0.08	0.10

The most repeatable exercises among pairs were tennis, swimming and basket (Table 4.6). The ski was the exercise in which the execution of the adult and the respective child differed more, both in the neurotypical pairs and in the ASD child/therapist pair. The ski was also the most difficult to mirror exercise according to Table 4.7, with the intraset distance between a child and the respective adult being larger than the intraset distance of the same child with the other adults for the several children. This difference was significant according to the Wilcoxon Signed test with $p < 0.05$. This movement included the hyperextension of the arm, being the keypoint of the arm occluded from the Kinect. The robot could not mirror correctly, so the children adapted the movement accordingly by slowing down the movements or by putting their arms in front of the trunk.

Table 4.6: Repeatability among pairs: Mean of the *intraset distances* for neurotypical subjects and *intraset distance* for ASD child-therapist pair for each type of exercise in the *Adult Coach* protocol.

	Tennis	Swimming	Bowling	Ski	Basket
Mean of intraset distance neurotypical	0.07	0.10	0.40	0.53	0.04
Pair therapist ASD child intraset distance	0.07	0.03	0.20	0.23	0.06

Tennis and basket, already the most repeatable exercises (Table 4.6), were also the most 'mirroringable' exercises, with significantly ($p < 0.05$) lower distances to the respective adult, than to the other adults (Table 4.7). In the clinical experiment, the similarity of the movements of the child and the therapist became evident through the computed outcome measures. For all exercises, the intraset distance between the ASD child and therapist was always smaller than the mean of the intraset distances for the pairs of the ASD child with the other adults.

Table 4.7: Mirroringability: For each sport in the *Adult Coach* protocol, *intraset distances* of the pair child-respective adult (left column) compared to the mean of *intraset distances* for the pairs child-other adults (right column). The coloured cells mark the lower distances, being green when verified by the couple child-respective adult (mirroring verified), and red, otherwise.

	Tennis		Swimming		Bowling		Ski		Basket	
	Resp adult	Other adults	Resp adult	Other adults	Resp adult	Other adults	Resp adult	Other adults	Resp adult	Other adults
Child1	0.03	0.04	0.004	0.06	0.66	0.21	0.67	0.34	0.03	0.15
Child2	0.38	0.39	0.004	0.02	1.10	1.05	0.52	0.45	0.02	0.02
Child3	0.04	0.05	0.14	0.18	0.02	0.08	1.41	0.77	0.00	0.00
Child4	0.06	0.06	0.01	0.06	0.52	0.58	0.56	0.54	0.01	0.05
Child5	0.12	0.04	0.00	0.07	0.08	0.07	0.50	0.12	0.07	0.09
Child6	0.00	0.00	0.01	0.04	0.94	0.14	0.07	0.14	0.03	0.03
Child7	0.02	0.05	0.09	0.03	0.06	0.60	0.20	0.26	0.12	0.04
Child8	0.02	0.01	0.16	0.25	0.70	0.48	0.80	0.43	0.02	0.05
Child9	0.02	0.03	0.22	0.16	0.00	0.34	0.79	0.09	0.06	0.04
Child10	0.01	0.01	0.02	0.07	0.03	0.05	0.57	0.46	0.00	0.02
Child11	0.00	0.01	0.11	0.06	0.04	0.15	0.54	0.27	0.07	0.08
Child12	0.10	0.10	0.27	0.21	0.13	0.13	0.02	0.16	0.04	0.06
Child13	0.00	0.01	0.16	0.11	0.57	0.48	0.34	0.24	0.02	0.03
Child14	0.01	0.01	0.17	0.07	0.71	0.86	0.41	0.67	0.06	0.09
ASD child	0.07	0.18	0.03	0.14	0.20	0.52	0.23	0.24	0.06	0.19

4.1.4 Discussion

This pilot study demonstrated the feasibility of a mirroring robotic coach protocol for turn-taking games between children and adults. Additionally, we introduced several performance and quality metrics, which show different values for the child with ASD compared to neurotypical children. This suggests that these metrics could be useful for assessing the progress of children with ASD during therapy.

Furthermore, we concluded that developing a new feedback system that recognizes the gestures as a whole and does not depend on their subphases is an important next step, especially when moving to intransitive gestures. In the Adult Coach protocol, we demonstrated that the correct imitation of the gestures depends on the characteristics of the robot. Thus, in these protocols, it is important to balance the movement complexity and the technical characteristics of the robot.

4.2 Gesture recognition system

As seen in Chapter 2, there is a lack of gesture recognition systems for small and similar gestures as the ones included in our hierarchical protocol. Moreover, through section 4.1, we verified how the lack of a gesture detection algorithm made the recognition part difficult. Thus, our first goal was to find an activity detection method that did not imply any prior knowledge. After, our goal was to construct a gesture recognition system specifically for our type of gestures, the intransitive gestures shown in Figure 3.2. In this section we will use the terms activity and gesture as synonyms.

4.2.1 Activity Detection

For the activity detection, two methods were tested: a classical one with the definition of a sliding window and a new one associated with the energy of the gesture, which we called the Kinetic Parameter Method, based on the work of [145]. This parameter is calculated at each frame, according to Equation (4.2), where p_f are the coordinates of each joint (keypoint) at the frame f, n is the total number of joints, and ΔT is the time interval between two successive frames. The Kinetic Parameter mean μ and standard deviation σ are essential for defining the segmentation points.

$$KP_f = \frac{1}{2}\sum_{j=1}^{n}\frac{(p_f^j - p_{f-1}^j)^2}{\Delta T} \tag{4.2}$$

Thus, the starting point is associated with the moment when the Kinetic Parameter surpasses $thresh_{start} = \mu + 0.3\sigma$ and the ending point is when the Kinetic Parameter is less than $thresh_{stop} = \mu - 0.3\sigma$. Using a validation dataset, the factor of 0.3 was chosen by trial and error to maximise the accuracy of the segmented data.

4.2.2 Activity recognition

After identifying the starting point, following the pipeline described in [6], the keypoints are filtered with a median filter and adjusted to a reference frame centred on the hip centre keypoint. In this way, the recognition is insensitive to the subject translation in relation to the camera. Moreover, the coordinates were normalized according to the subject trunk length to be independent of the subject's height.

Since this pipeline is based on image recognition from a Convolutional Neural Network, the keypoints are stacked in a 3D matrix until the stopping point (dimension = $K \times F \times N$, where $K = 25$ is the number of keypoints, F is the number of frames between the starting and stopping points and $N = 3$ is the number of coordinates) to form an image. A Min-Max normalization is done to consider the volume of the movement. Thus, the 3D matrix is normalized by the minimum and maximum coordinates of the movement.

Afterwards, a Savitzky-Golay smoothing is applied to improve the quality of the data. As in [6], we decided to use a five point filter. Subsequently, the keypoints were rearranged in order to generate more recognizable patterns. First, we grouped the keypoints according to the body parts: trunk; right arm; left arm; right leg; left leg. The assignment of each keypoint to the body parts is presented in Table 4.8. Three combinations were tested:

1. Trunk - upper limbs - lower limbs
2. Trunk - lower limbs - upper limbs
3. Upper limbs - trunk - lower limbs

We did not consider any difference between left or right cause a mirroring process was applied to the selected dataset. We excluded some keypoints cause they did not contribute to the recognition process, namely the head and feet keypoints. The head keypoints were moved in several gestures and even when the person was not doing any gesture. The feet keypoints were very noisy, not always correctly recognized and that is why we decided to eliminate them.

After, the 3D matrix was multiplied by 255, being converted to an RGB image, in which each coordinate was associated with a colour channel ($x - R$, $y - G$, $z - B$). Thus, each image row corresponded to the keypoints and the columns to the time, as shown in Figure 4.2. We called this image an RGB Pose Feature.

Table 4.8: Index of the keypoints associated to each different body segment. For index meaning please refer to Figure 2.8.

Body segment	Keypoint indexes
Trunk	0, 1, 4, 8, 20
Right Arm	9, 10, 11, 23, 24
Left Arm	5, 6, 7, 21, 22
Right Leg	16, 17, 18, 19
Left leg	12, 13, 14, 15

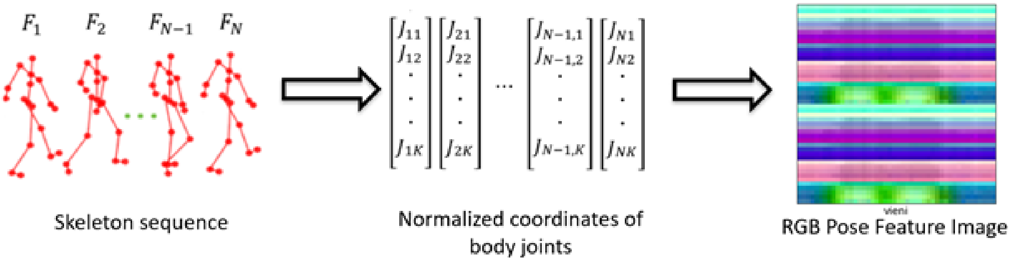

F_1
F_2
F_{N-1}
F_N
Skeleton sequence
J_{11} J_{12} J_{1K}
J_{21} J_{22} J_{2K}
$J_{N-1,1}$ $J_{N-1,2}$ $J_{N-1,K}$
J_{N1} J_{N2} J_{NK}
Normalized coordinates of body joints
vieni
RGB Pose Feature Image

Figure 4.2: Main steps of the data processing starting from raw keypoints to obtain an RGB pose feature image. In particular, N refers to the number of frames and K is the number of keypoints considered (23 in this work) [6].

Then, the data matrices were interpolated along the temporal dimension with a Nearest Neighbour resampling filter to have inputs with the same dimensions. This is essential for the training of Convolutional Neural Networks in which inputs cannot have different dimensions. Finally, a Contrast Limited Adaptive Histogram Equalization was applied to the image to enhance the contrast. In this way, the Neural Network could recognise the patterns better.

As in [6], we used Convolutional Neural Networks and, more specifically, the ResNet architecture. Our model was formed by three ResNet blocks and one Softmax layer. The network received an image as input, and outputs a vector with the probability of each gesture (dimension 17). The chosen gesture should correspond to the entry with the highest probability but that was not the final output of the system since there were several false positives. Therefore, if the highest probability was higher than a given threshold, the vector of probabilities was added to a buffer. When this buffer was completed, its mean probabilities were calculated, and the maximum among these probabilities was the final output of the system. If the recognised gesture was equal to the gesture selected by the therapist, the robot produced a positive feedback (”Bravo!”), if not, a new frame was added, and a new prediction was obtained. We repeated this procedure until the gesture was correct or until the end of a predefined time interval (10s). The threshold and the buffer size were the main hyperparameters of this part. Figure 4.3 illustrates the block diagram with all the several steps.

4.2.3 Results

Two different datasets were created and used for training and testing the gesture recognition system: the neurotypical participants’ dataset and the retrospective ASD dataset. The first dataset took inspiration from the hierarchical protocol and consisted of acquisitions performed with 16 neurotypical subjects aged between 23 and 27 years old. The protocol included a triadic interaction between the robot and two people, similar to the previously presented protocols. The robot performed each of the 17 gestures, and each person repeated the gesture two times at his/her turn. In total, 32 samples were obtained for each gesture, producing a balanced dataset.

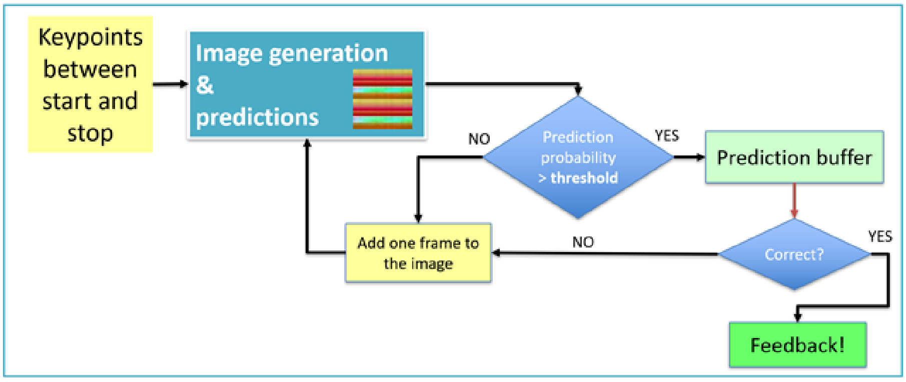

Keypoints between start and stop
Image generation & predictions
Prediction probability > threshold
NO
YES
Prediction buffer
Add one frame to the image
NO
Correct?
YES
Feedback!

Figure 4.3: Block diagram of the online activity recognition system. After extracting the keypoints between the start and end points, the RGB pose feature images are generated and passed through the ResNet architecture. If the prediction probability exceeds a threshold, the probabilities from the last softmax layer are stored in a buffer. Once the buffer is full, if the gesture with the highest probability matches the one presented by the therapist, the robot provides feedback; otherwise, a new frame is added.

The retrospective ASD dataset consisted of previous sessions with 5 children with ASD (age 4.9 ± 0.74 years old) done with the hierarchical protocol in Fondazione Don Carlo Gnocchi (FDG) within a Pilot Study. As described in Chapter 3, the hierarchical protocol was constituted by several levels, in which the first two were of familiarization with the setup and the last three were the levels of training of intransitive gestures. The number of sessions done by each child was different depending on the moment in which they started Level 3. Each child also performed a different number of gestures, constituting an unbalanced dataset. A final real-time acquisition was done with one ASD child included in the same Pilot Study.

Activity Recognition performance

For the recognition of the gesture performed and to avoid false positives, we collected each network's prediction in a buffer. Only when we had the same prediction six times, we considered it the recognized gesture. The number six was defined from a trial and error procedure. The threshold to include the prediction in the buffer (to mark each prediction as a correct prediction) was tuned for each gesture. Initially, this threshold was set to 0.85. The mean predictions obtained in 10 attempts by neurotypical subjects were analysed for different gestures. Based on the analysis presented in Table 4.9, in the gestures *short*, *listening*, *small*, the threshold was set at 0.75 and it was increased to 0.99, in the gesture *give*.

Table 4.9: Mean prediction probabilities in output from the Softmax layer of ten gestural trials of neurotypical subjects.

Gestures	Mean±SD	Gestures	Mean±SD
tall	0.98±0.01	where	0.96±0.08
angry	0.97±0.10	hungry	0.99±0.05
listening	0.83±0.22	happy	0.87±0.20
waiting	0.87±0.12	big	0.96±0.07
kissing	0.97±0.05	me	0.95±0.10
short	0.75±0.28	small	0.76±0.08
hello	0.95±0.11	pointing	0.99±0.02
peekaboo	0.93±0.10	come	0.99±0.04
giving	0.99±0.08		

After establishing the threshold, we analysed just the performance of the activity recognition method (Figure 4.4). We have used the Neurotypical dataset for training the Neural Network. After segmenting the gestures, we constructed the training set with 12 people, the validation set (for the hyperparameters tuning) with 2 people and the test set with another 2 people. The maximum accuracy achieved in the test set was 96%. Analysing the confusion matrix, the most misclassified gestures were *waiting* and *kissing*. These gestures were associated with the occlusion of certain keypoints of the Kinect, which justified the decrease in performance. . Our accuracy is close to that reported in [6], which achieved 99.9%, but on a dataset with a greater variety of gestures and, in particular, smaller gestures.

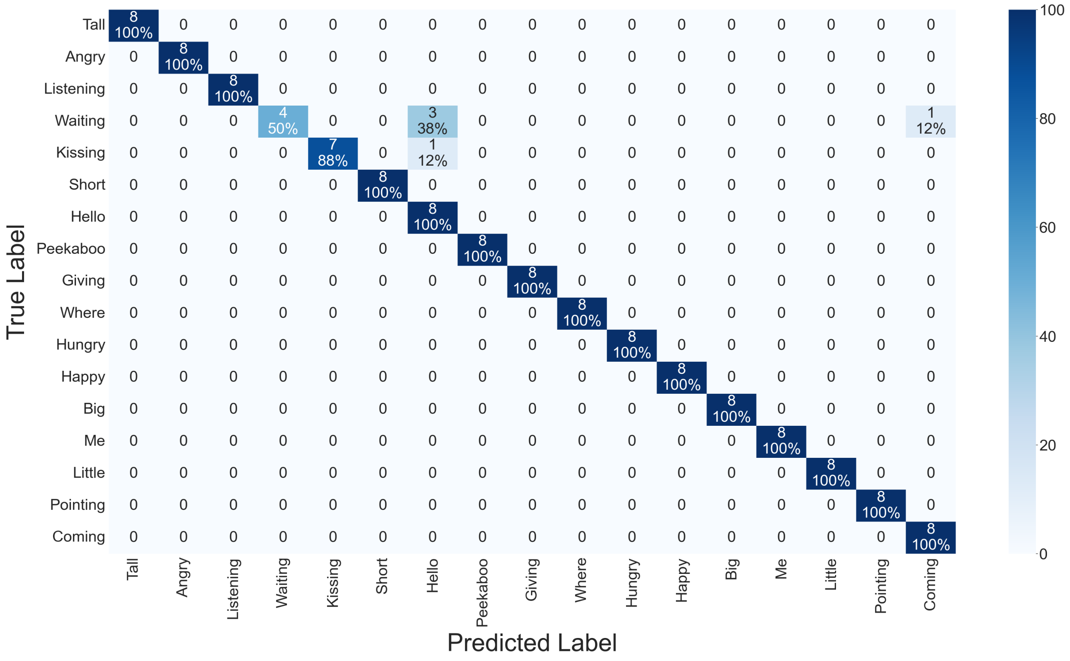


Figure 4.4: Confusion matrix obtained by applying the activity recognition model on the Neurotypical dataset. These results were good since the accuracy of 96% was much larger than the chance level (1/17 gestures × 100=6%).

Comparison of the Activity Detection methods

We have tested the protocol online with three neurotypical subjects to evaluate the complete recognition system. Each gesture was repeated eight times by the subjects. The accuracy obtained was 92%, very close to the offline one, and the confusion matrix was similar to the one obtained offline. One of the principal misclassified gestures was *waiting*.

In order to understand the importance of activity detection, we applied the online method to our target population with the retrospective ASD Dataset, using the Sliding Window method and the Kinetic Parameter Method. We gave the algorithm each frame separately, to simulate the online scenario. The results are reported in Table 4.10.

As previously explained, this dataset is unbalanced since not all the children performed the same number of gesture. Thus, we compared the two methods regarding the F1-score. The Kinetic Parameter presented an improvement of 19% specifically in this metric, but all the other metrics also increased compared to the Sliding Window method. The precision was always higher than the recall, which is better for this type of protocol, where a lower number of false positives is more important than a lower number of false negatives. The feedback should only happen when the children do the gestures correctly. In

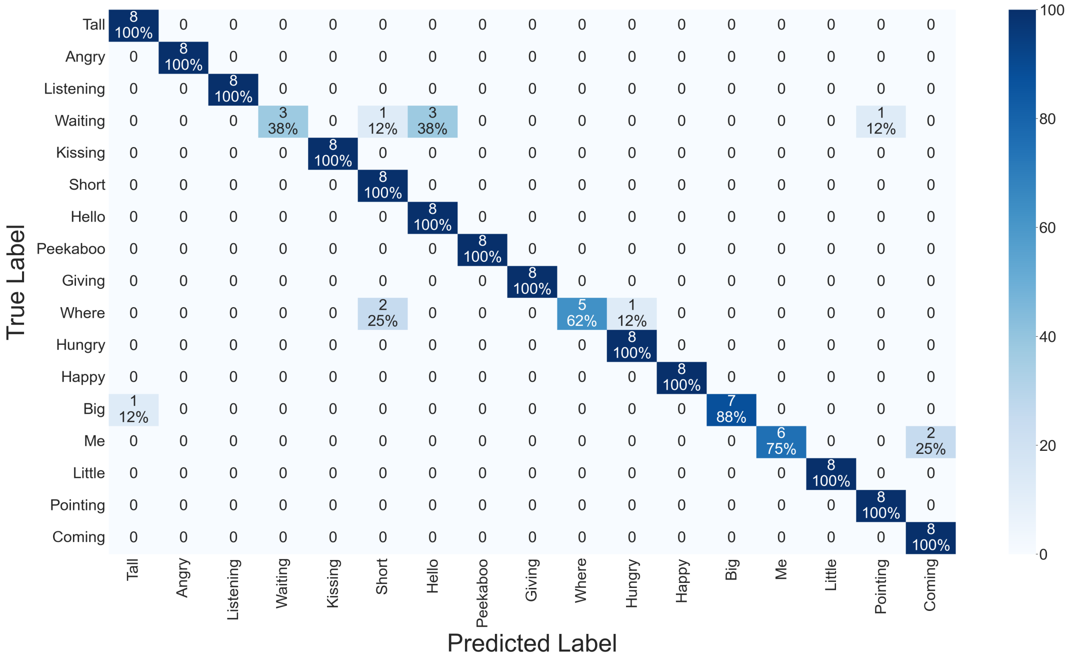


Figure 4.5: Confusion matrix obtained with the complete online method with 3 neurotypical subjects and 8 repetitions.The small decrease of accuracy in relation to the offline method was expected since it depends on the activity detection part. Nevertheless the accuracy 92% is still above of the chance level (6%).

case of false negatives, the therapist can always give positive feedback using a button in the interface.

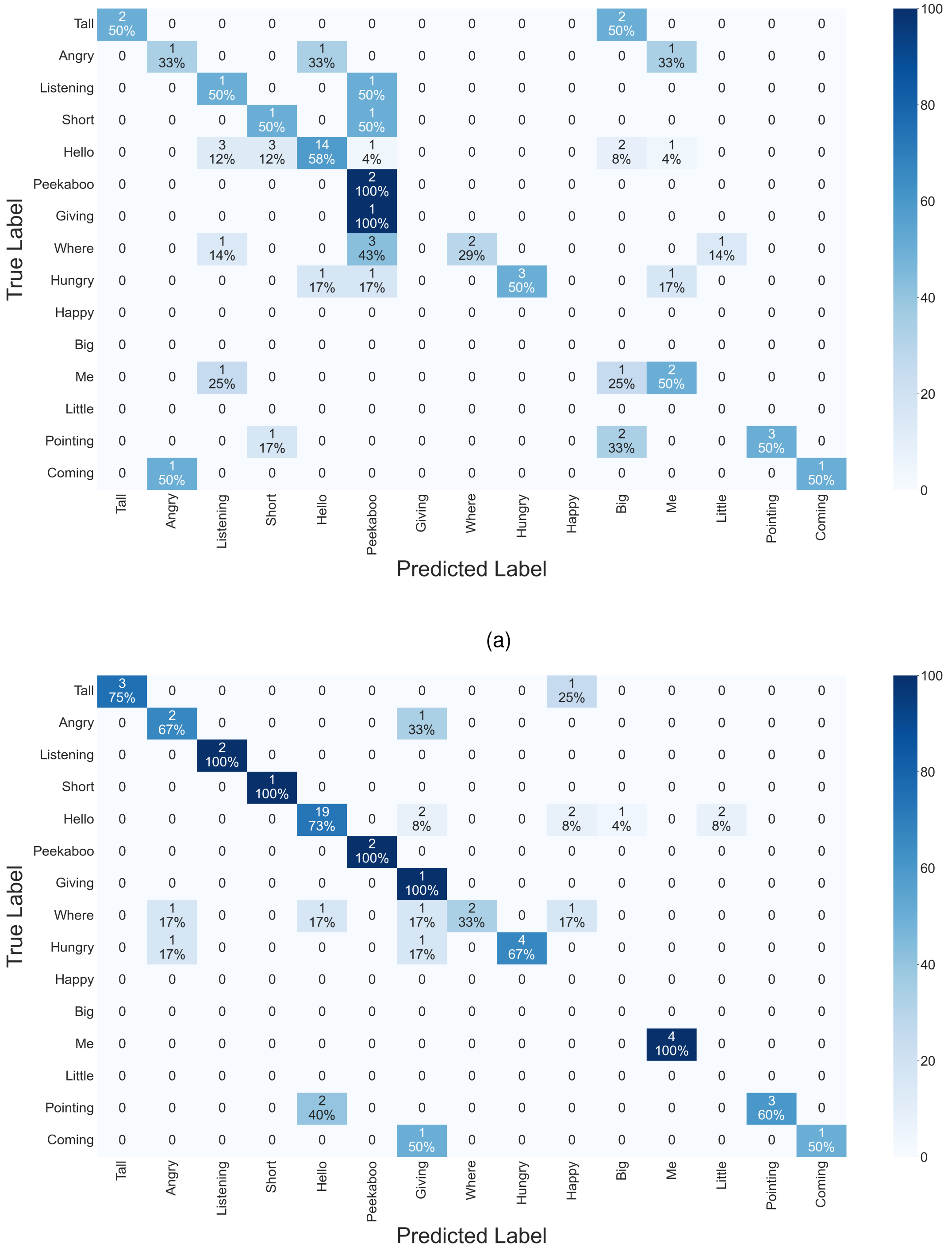
Table 4.10: Comparison between results on ASD children obtained with two segmentation methods: with Sliding Window[1] and with the Kinetic Parameter. Bold indicates the main measure analysed, F1-score, since the dataset is unbalanced.

	Sliding Window	Kinetic Parameter
Accuracy	49%	70%
Precision	80%	91%
Recall	49%	70%
F1-score	**57%**	**76%**

The confusion matrix in Figure 4.6a for the retrospective ASD dataset shows a considerably larger number of misclassified gestures compared to the confusion matrix of the neurotypical dataset (Figure 4.5). This misclassification was expected since the Neurotypical dataset consisted only of adults, and children have a more variable dynamic when performing the gestures, as concluded by our first contribution in Section 4.1.3. Comparing the two methods (Figure 4.6a and Figure 4.6b), the Kinetic Parameter considerably improved the recognized gestures reducing the number of misclassifications, improving the accuracy and the F1-score (Table 4.10).

Segmentation Quality by human expert validation

In order to further evaluate the importance of the activity detection part, we have asked an external assessor to classify if the segmentation was done correctly or not in 63 gestures. 38 gestures (60%) were correctly segmented. From these, 82% were correctly classified, reinforcing the importance of correct detection for correct recognition of the gesture. Moreover, we have manually inspected the gestures to verify the starting and stopping instants. Then, we calculated the delay between the algorithm-proposed



(a)

(b)

Figure 4.6: Confusion Matrices of the online application on the *Retrospective ASD children dataset* without (a) and with (b) the Kinetic Parameter algorithm. The improvement in terms of the accuracy is noticed by the reduction of the number of misclassified gestures.

instants and the real instants (Table 4.11). We concluded that correct detection and recognition of the gestures are especially dependent on starting instant gesture identification. The delta between the perceived incorrect and correct gestures was much higher for the starting instant than for the stopping instant. Thus, correct and incorrect gestures have similar delays in terms of the stopping instant identification, while they differ for the starting instant identification.

Table 4.11: Mean delays between expected and predicted starting and stopping instants for the gestures classified as correct and incorrect by the blinded assessor.

	Correct	Incorrect	Delta
Start (s)	1.11±1.59	3.1±3.18	1.99
Stop (s)	3.57±3.18	4.43±4.03	0.86

Real-time Online Recognition during Therapy

At the time of our study, just one child of our Pilot Study in Fondazione Don Carlo Gnocchi (FDG) was currently doing Level 3, where the recognition system was used. The child executed 11 gestures with an emphasis on total body movements. The accuracy obtained was 55% which was considerably lower than the one obtained by the retrospective analysis but still above the chance level (1/17 gestures×100=6%). Most gestures were misclassified with the gesture *happy*. During the therapy, the child sat before performing the gesture and then stood up, an action associated with the gesture *happy*. Nevertheless, the total body gestures were more engaging to the child than the ones with just the upper limbs.

4.2.4 Discussion

With this work, we produced a new recognition system able to identify small and similar gestures using a single model, contrarily to [6] that had a model for each type of gesture. Key innovations of this study, such as defining parameters like keypoint order and probability thresholds, proved to be important for accurate recognition. Moreover, adding a new activity detection method improved the results obtained with the previous dataset of ASD children.

However, these results were inferior to the ones obtained with the neurotypical dataset. A solution could be to expand the training dataset to include more children. First, the impact on the performance metrics of adding neurotypical children could be studied. Then, the same analysis could be conducted for the inclusion of children with ASD.

Furthermore, an important next step is solving the issue raised during the clinical acquisition, in which several gestures were misclassified with happy, since the child was always sitting and rising. A solution could pass by considering other variables, such as the jerk and the acceleration, that would help distinguish a rising movement from a gestural starting.

4.3 Conclusion

We presented several metrics for the evaluation of the actions performed during a robotic-assisted therapy. Through the first mirroring measure, we concluded that a feedback system just based on the angles was not able to obtain the gist of the movement and failed sometimes. With the implementation of the neural network, this problem was solved probably because the feature maps provided a more complete representation of the gestures.

The segmentation of the movements revealed to be the hardest part in both measures. In the mirroring metric, the parameters for segmentation were established through trial and error. In the gesture metric, after using the sliding window, we established the new kinetic energy parameter, which was effective, but further work is required to improve its robustness in real therapeutic scenarios. At that stage, we could apply our initial performance metrics—such as latency and feedback time—to our new dataset of children with ASD.

Nevertheless, these metrics have shown the difference in performance between children and adults and in the case of the mirroring measure, they are able to identify the difference between neurotypical children and children with ASD, enlarging the possible application to other rehabilitation fields in Autism. A comparison with clinical scales should be explored for implementation in clinical practice.

Chapter 5

Attention biomarker

After the exploration of the action performance measures we focused on assessing attention indicators. Attention was one of the main measures targeted by our clinical staff, since, as previously explained, children with ASD usually present some deficits in this capability. Specifically, joint attention is one of the main capabilities trained in the rehabilitation therapies of children with ASD due to its importance in learning.

Therefore, first, we created a system that analysed the children's attention during each session from the recordings registered with the camera (Section 5.1). Second, with the final goal of including this system into the adaptive control of the robot, we studied methods for transforming this offline version to an online one (Section 5.2). However, when changing the offline method to online, we verified a decrease in the system's accuracy, caused by a non-linear offset in the gaze estimator. To deal with this offset and since we also wanted to establish an outcome measure based on a clinical scale, we constructed a neural network for the automatic classification of the subjects' attention (Section 5.3). As in the previous chapter, the results and the respective discussion are presented for each contribution.

5.1 Attention classification system

For this work, we had two goals: (i) construct a robust attention system classifier for a triadic unconstrained scenario and (ii) find assessment measures that could be interpreted by therapists. The attention classification system used the subject's gaze and defined Area-of-interest (AOI) around the targets. In this triadic scenario, each person's primary targets were the other individual present and the robot. Additional targets were included as needed, depending on the protocol. While most estimators provide both azimuth and elevation angles, we focused on the azimuth angle here, as it was sufficient to distinguish most of our targets.

Overall, we have followed four different steps:

1. compare two 360º estimators present in the literature, Gaze360 and WHENet, in a constrained environment with one neurotypical adult;
2. process the clinical sessions with the chosen estimator and prepare the respective skeletons;
3. analyse the scene geometry and define the AOI around each target (the AOI corresponds to a range of angles to look to a certain target- robot, therapist, etc.);
4. classify the gaze of each frame of each subject as 'looking at the respective target', if it was between the established range of AOI or 'looking elsewhere', otherwise.

A full overview is shown in Figure 5.1.

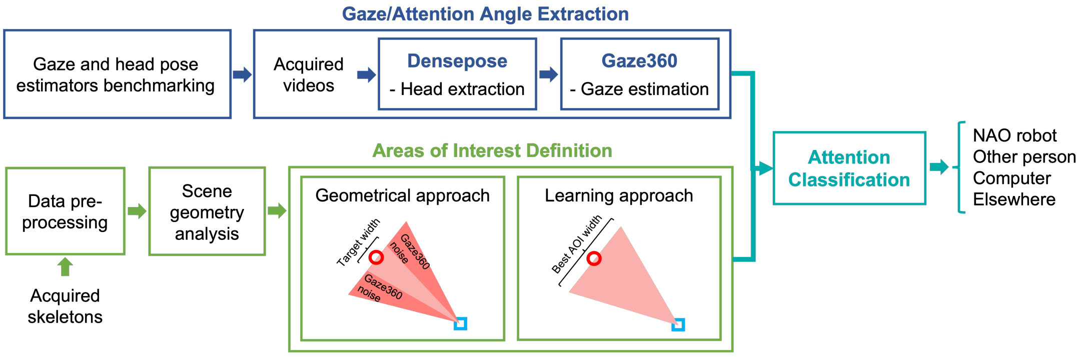


Figure 5.1: Full overview of the attention classification system. The system is composed of two main blocks: the attention angle extraction and the Areas of Interest (AOIs) definition (Sections 5.1.2 and 5.1.3). After these two blocks, we compared the extracted angle with the defined AOIs to classify the attention (Section 5.1.4) into one of four targets: NAO robot, Other person, Computer, Elsewhere. For the choice of the attention estimator, we did an initial benchmark of gaze and head pose estimators (Section 5.1.1). For the definition of the AOIs, an initial data preprocessing (Section 5.1.2), followed by the analysis of the scene geometry (Section 5.1.3) was required.

5.1.1 Benchmark of the 360º Estimators

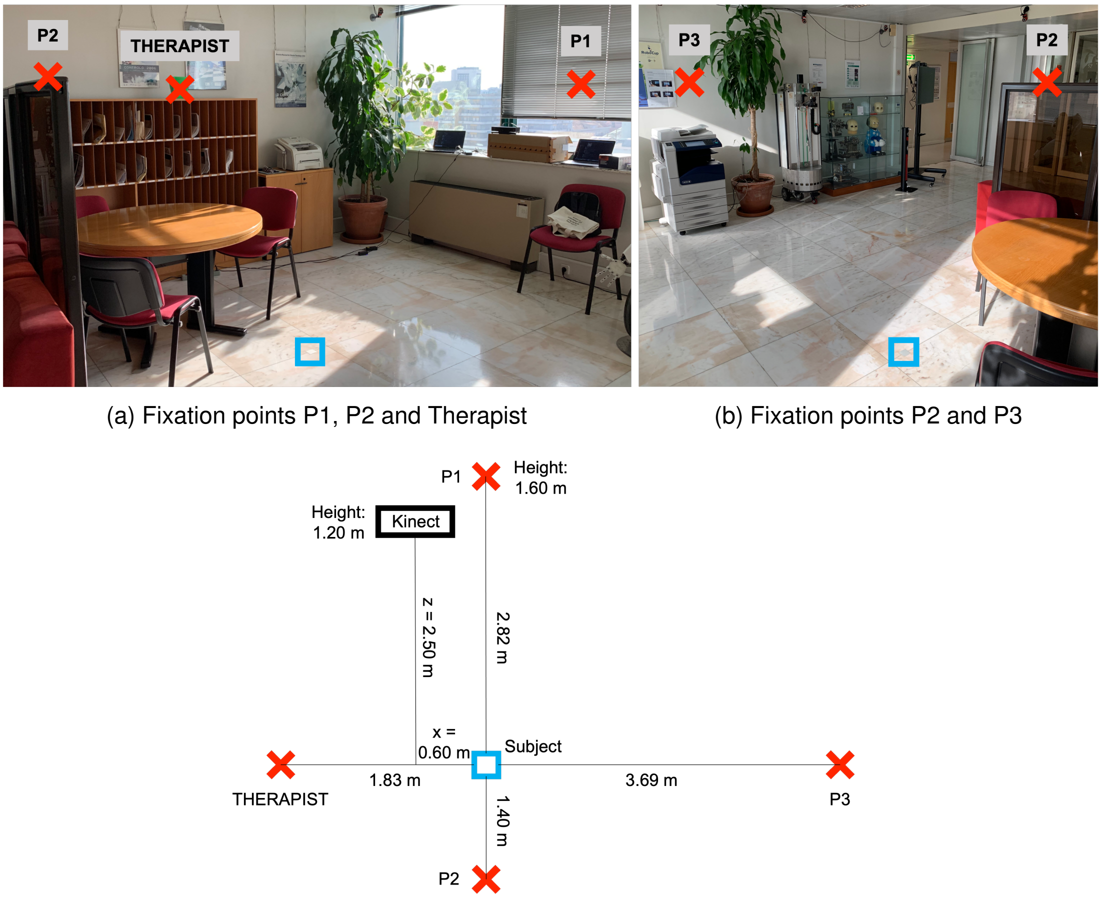
For the benchmarking, we selected Gaze360 and WHENet from the literature for their capability of estimating gaze even when the facial features are not visible. Three experiments were conducted to cover the full field of view and the distance conditions present in the clinical studies. In the three experiments, the subject was in front of a camera at 2.5 m, and several targets surrounded him/her at known locations, as ground truth. Every 10 seconds, after a sound cue, the subject shifted his/her gaze to a different target.

The first experiment was designed to test the robustness to occlusions. Therefore there were four points around the subject including one in the back. The points were all at eye level to minimize the noise and the subject was instructed to move her eyes, head and body over each fixation point (Figure 5.2). This experiment was performed four times.

In the second experiment, the four points were all in front of the subject and one of the points represented the robot. The subject was told to move just the head and the eyes (Figure 5.3). The subject repeated this experiment three times.

In the third experiment, there were three fixation points, all at eye level and in front of the subject and the subject should move just the eyes (Figure 5.4). As the previous experiment, this experiment was performed three times.

In the end, for each experiment, we calculated the Root mean squared error (RMSE) between the ground truth and the signal predicted by each estimator. In Table 5.1 are presented the results of each experiment. From experiment 1, WHENet has a better accuracy than Gaze360, especially for estimating the point in the back. Instead, in experiment 2, when all the points are in front of the subjects, WHENet and Gaze360 present a similar performance. However, there is an accuracy decrease in the WHENet in the third experiment, as expected since the subject is not moving the head and just the eyes. Overall, Gaze 360 has a good performance when at least the profile facial features are visible, a condition that happens during the therapies where all the targets are in front of the subject. On the other side, WHENet predicts well when the facial features are not visible but it does not follow the gaze, since it is a head estimator. Therefore, we chose Gaze360 for our attention classification system.

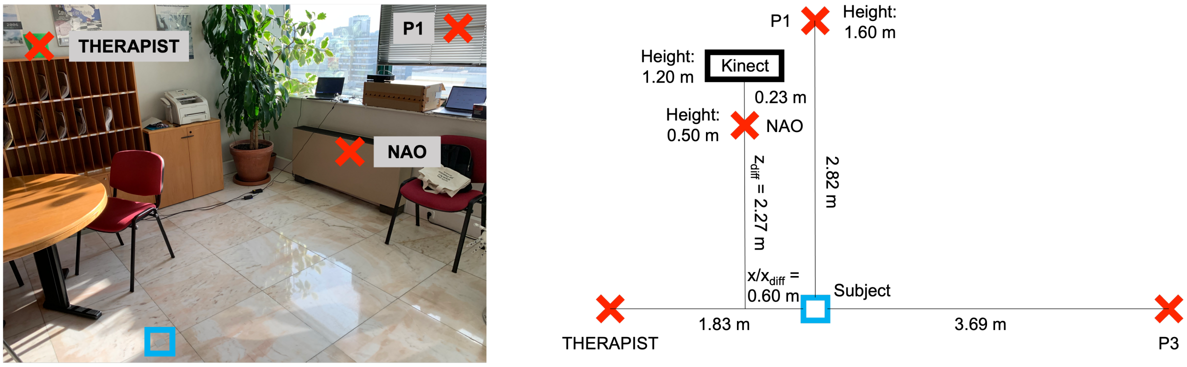


(a) Fixation points P1, P2 and Therapist

(b) Fixation points P2 and P3

(c) Scheme view from above (not scaled)

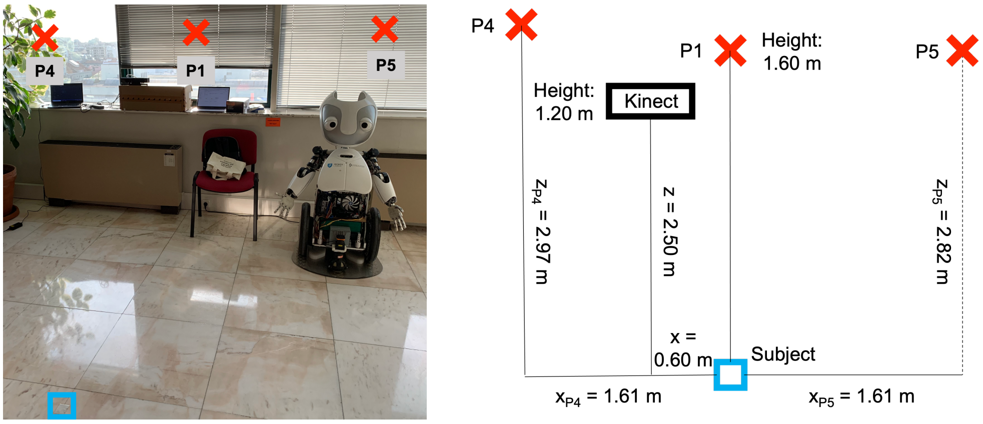
Figure 5.2: Long Distance Benchmarking Setup: Experiment 1. The red crosses represent the 4 fixation points and the blue square represents the subject.



(a) Fixation points 1, Therapist and NAO

(b) Scheme view from above (not scaled)

Figure 5.3: Long Distance Benchmarking Setup: Experiment 2. The red crosses represent the 4 fixation points and the blue square represents the subject.



(a) Fixation points 1, 4 and 5

(b) Scheme view from above (not scaled)

Figure 5.4: Long Distance Benchmarking Setup: Experiment 3. The red crosses represent the 3 fixation points and the blue square represents the subject.

Table 5.1: Average RMSE of WHENet and Gaze360 azimuth estimates (rad). Experiment 3 focused only on the movement of the eyes and it was the experiment in which Gaze360 presented better results. Bold indicates the lowest RMSE per experiment. In Experiment 2, no value is marked due to their similarity.

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

5.1.2 Gaze Estimation and Data Processing

After integrating Gaze360 into our attention system to capture azimuth and elevation angles, we established Areas of Interest (AOIs) based on the targets' positions. The individuals' positions were determined using data extracted from the Kinect sensor. However, this setup presents significant challenges for such cameras: children are often lying on the ground with half of their bodies obscured, and therapists may hug children to soothe them, among other scenarios. Consequently, detecting both the child and therapist bodies is sometimes impossible. Therefore, we remove all the frames in which the two skeletons were not detected. As our analysis focused on the interactions between the child, therapist, and robot, ensuring the presence of both the child's and therapist's skeletons was essential for accurate gaze analysis. Then, we did a linear interpolation to compensate for the cases in which a considerable amount of Kinect's data was lost. The results of such interpolation are shown in Table 5.2. To reduce the number of wrong skeletons created by the interpolation, we assigned each Kinect skeleton with the bounding boxes produced by the Densepose. The frames in which this correspondence did not happen were discarded. We studied the effect of augmenting the Densepose bounding boxes by $p \in \{25\%, 50\%, 75\%\}$ to eliminate more or fewer skeletons and, consequently, frames.

Table 5.2: Percentage of lost data for the children who attended that therapy session: with and without data interpolation. The red cells represent the session in which more than 2/3 of the data was lost (%).

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

5.1.3 Areas-of-Interest construction

For each target, we have constructed an AOI, considering the range of angles that each participant's gaze should have if he/she looked at the target. The targets chosen were three: robot, other person (child or therapist) and computer. We included the computer because it was a distraction during Level 3 and a focus of attention during Level 4 since the scenarios were presented there.

Using the scene's geometry, our AOIs were defined by the position of each target and a certain line of sight with a width w that established the AOI's size. The line was set perpendicular to the standard gaze direction (α_{target}), shown in Figure 5.5

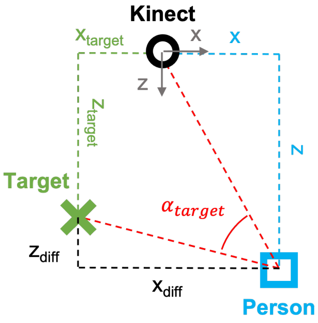


Figure 5.5: Standard angle (α_{target}) representation. The green cross represents the target, while the blue square represents the person from whom the standard angle is calculated. The origin of the reference frame is located in the centre of the Kinect and the coordinates of the person and the target are represented by (x, y) and (x_{target}, y_{target}), respectively. The variables x_{diff} and z_{diff} denote the differences between the 2D coordinates of the person and the target.

α_{target} represented the angle that the subject's gaze did when he/she looked directly at the target, and we denominated from now on standard angle. Therefore, this gaze direction depends on the position of the person and the target. When the person and target were on the same side of the camera, and the person was closer to the camera in the x and z direction, the standard angle was given by Equation (5.1). In this equation, x and z are 2D positions of the head keypoint given by the Kinect and x_{diff} and z_{diff} are the differences between the 2D coordinates of the person and the target.

$$\begin{aligned} \alpha_{target}(t) = \; & n \arctan\left(\frac{x(t)}{z(t)}\right)^{n} + \\ + \arctan\left(\frac{z_{diff}(t)}{x_{diff}(t)}\right)^{n} & + n\frac{x(t)}{|x(t)|}\frac{\pi}{2} \quad \text{, with} \end{aligned} \tag{5.1}$$

$$\begin{cases} n = 1, \text{if} \arctan\left(\frac{z(t)}{x(t)}\right) \geq \arctan\left(\frac{z_{target}(t)}{x_{target}(t)}\right) \\ n = -1, \text{if} \arctan\left(\frac{z(t)}{x(t)}\right) < \arctan\left(\frac{z_{target}(t)}{x_{target}(t)}\right). \end{cases}$$

For all the other positions of the person and of the target, the standard angles was given by Equation (5.2).

$$\alpha_{target}(t) = -\arctan\left(\frac{z(t)}{x(t)}\right) + \arctan\left(\frac{z_{diff}(t)}{x_{diff}(t)}\right). \tag{5.2}$$

After establishing the line of sight direction, we tested two approaches for the definition of the width w:

- *Geometrical approach*: we established a width for each target, using their real dimensions and adding the Gaze360 estimated noise. Thus, for the robot, we considered its torso width plus the length of its open arms ($27.5 + 2 \times 31.1 = 89.7cm$) since it moved the arms considerably during the gestures' production. For the child's and therapist's width, we experimented to adopt the same width for both groups or differentiate between groups. In the first case, we used the average shoulder width of a female adult ($43.26cm$) [164] or associated it with the average child's shoulder width in the second case ($32.8cm$) [165]. The Gaze360 noise was intrinsic to the estimation of the algorithm and reflected its precision. We estimated it experimentally in the laboratory, in which a subject looked at several fixation points for 10 seconds. Then, the standard deviation between the expected signal and the Gaze360 signal was calculated. The Gaze360 noise was given by the maximum standard deviation ($0.069rad$).

- *Learning approach*: several widths were tested for each target, with ranges between $[0.4, 3.0]m$ for NAO, $[0.4, 2.0]m$ for the Other Person, $[0.4, 1.0]m$ for the Computer. The best widths were the ones in which the final algorithm performed better, optimising the recall and the false positive rate. These scores were calculated by comparing the final classification of the system with the ground truth established by two independent annotators. Moreover, we tested the same widths for all the people involved but also considered two separate groups: neurotypical and Autism Spectrum Disorder (ASD) children. In this case, the group of neurotypical was just formed by one therapist.

With the width of each target, we established the borders of our AOI and transformed them to angles (α_1 and α_2), using again the Equation (5.2) and Equation (5.1) (Figure 5.6). In this way, we directly compared the values estimated by Gaze360 and the ranges of the AOI.

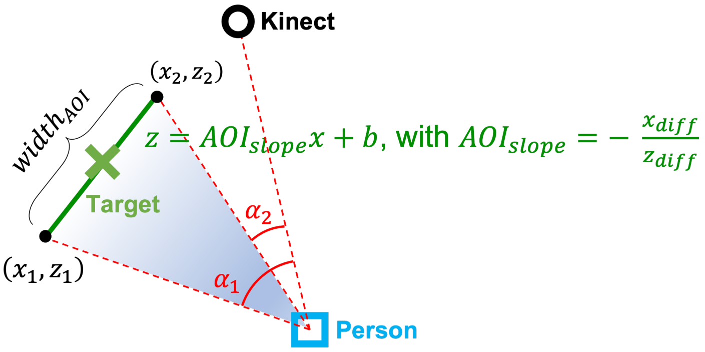


Figure 5.6: Representative top view of an Area-of-interest (AOI). The green cross represents the target, while the blue square represents the analysed person. The dark green line corresponds to the AOI width and the area in blue to the range of angles (receptive field) for looking at the target.

However, sometimes the AOIs of the different targets had some overlap. Two steps were adopted to correct this overlap. If the AOIs overlapped completely we established a priority order according to the scene's geometry and one of the AOIs was disregarded. For example, if the therapist's AOI was in the same gaze direction of the NAO's AOI, this latter was deleted. These priorities were found based on what was more frequent during the sessions: Other Person>NAO>Computer. Instead, if the AOIs overlapped partially, a decision threshold was computed. For each instant, for each target, a Gaussian curve was created $N_i(\mu_i, \sigma_i)$, with $i = \{1, 2\}$. The mean, μ_i, was defined as the mean value of the AOI limits at that instant (Equation (5.3)). The standard deviation, σ_i, was calculated using an empirical rule. $k\sigma$, with $k = \{1, 2, 3\}$, was defined as half of the AOI width (Equation (5.4)). 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.

$$\mu_i = \frac{\alpha_{1_i} + \alpha_{2_i}}{2}. \tag{5.3}$$

$$k\sigma_i = \frac{width_{AOI_i}}{2}. \tag{5.4}$$

The azimuth in which the Gaussians intersected was defined as the decision boundary between the AOIs (Figure 5.7).

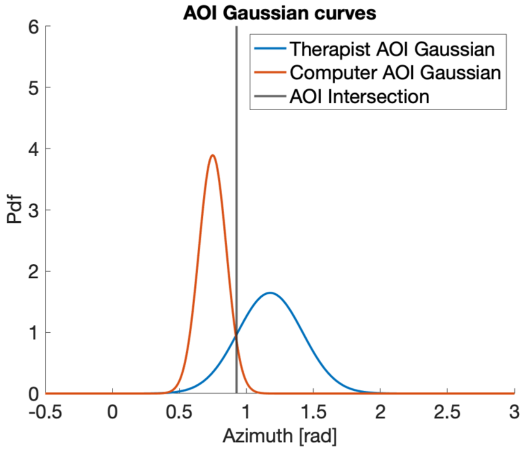


Figure 5.7: Example of decision boundary at a given frame for two overlapping Areas of Interest (AOIs) (computer and therapist). After establishing the probability density function (Pdf) for each AOI, the decision boundary is decided based on the intersection of the two gaussian curves.

5.1.4 Attention classification

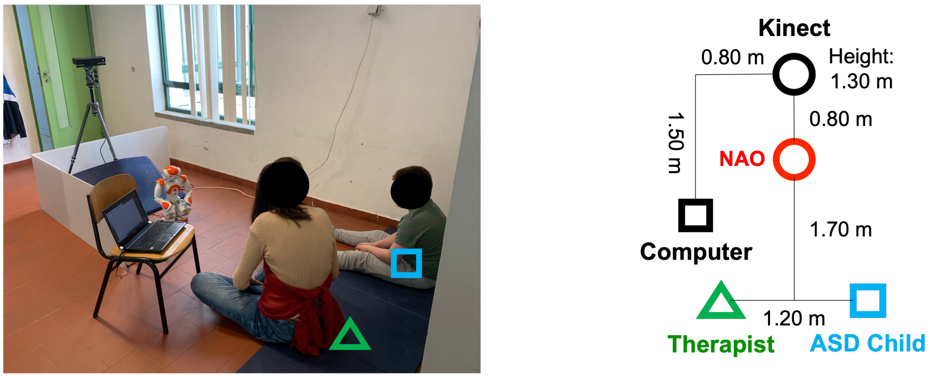
For the correct classification of the attention, we compared the Gaze360 angles with the ranges of AOIs. We then quantized the Gaze360 signal into four levels indicating where the subject was looking: at the other person, the robot, the computer, or elsewhere. We referred to this processed signal as the fixation signal. We considered that a person was looking to a target if the duration exceeds 400 ms, according to [166]. We removed the spurious fixations using a median filter.

After generating the fixation signal, we opted to summarize the information into statistical indicators that would be valuable for the therapists. Specifically, we calculated the Total Fixation Duration (TFD) for each target, which represented the percentage of time during a session that a child spent looking at each target, reflecting his/her attention.

5.1.5 Results

Our attention system was evaluated on robotic sessions done in a primary school in Portugal (Escola Básica Bernardim Ribeiro) between May and July 2021. Associação Portuguesa para o Autismo e as Perturbações do Desenvolvimento (APPDA) develop an intervention program for Autism in this school, and they chose it for our study.

The same therapist tested our protocol with 5 children between 7 and 11 years old. For each child, the number of sessions varied between 2 and 7 sessions according to their school attendance. Four of the five children were diagnosed with Level 3 of autism, while one was diagnosed with Level 1, according to the DSM-V [10].



(a)

(b)

Figure 5.8: Setup representation with the therapist (green triangle), the ASD child (blue square), NAO (red circumference), the computer (black square) and Kinect (black circumference).

The sessions were carried out in the school atrium with the children sitting according to Figure 5.8. Since multiple people passed in this atrium, to identify the child and the therapist, we selected the two skeletons with the largest x coordinate of the left shoulder, which corresponded to the skeletons more to the left in the camera view. The data acquired with the Kinect camera were the video of the complete session and 3D and 2D skeletons of the children and the therapist. We have just analysed the sessions corresponding to Levels 3 and 4 of the protocol. In these levels, the child was already familiar with the robot and the interaction between the child, robot and therapist was more important than in the other levels.

Data division and hyperparameters tuning

As mentioned before, the children participated in a different number of sessions. Depending on the strategy chosen for the widths computation when defining the AOIs, the division of sessions for validating the hyperparameters of the system and testing it was different. In the learning approach, the session chosen for training was session 3 because it had a higher number of subjects. For both geometrical and learning approaches, the session used for validating the parameters was session 6. All the other sessions were used for testing.

As previously mentioned, the labeling was done by two annotators. Since some of our videos contained more than 15,000 frames and we had more than 20 videos, each annotator labeled every three seconds. This approach ensured the inclusion of different fixations and targets. At each frame, the annotator determined whether the child and the therapist were looking at NAO, another person, a computer, or elsewhere.

To assess the system's performance, we only retained the frames in which both annotators agreed, which accounted for 75% of the frames, demonstrating strong agreement between the two annotators. We then compared these labels with the system's estimates and calculated the recognition rate, false positive rate, and false negative rate. The recognition rate was computed by considering both children and therapists together, serving as the overall system accuracy/performance indicator.

The hyperparameters validated were the Densepose bounding boxes ratio ($p \in \{25\%, 50\%, 75\%\}$) that controlled the amount of skeletons discarded (Section 5.1.2) and variable $k \in \{1, 2, 3\}$ that established the standard deviation of the Gaussian curves when two AOIs overlapped (Section 5.1.3). The results are shown in Table 5.3. We can notice that for the parameter k there was no agreement between the two approaches since its best value was $k = 1$ for the geometrical approach and $k = 2$ for the learning approach. Regarding the p parameter, it did not have a considerable impact on the accuracy but larger values of p are associated with better performances, probably because useful and reliable keypoints were kept in these cases. In the learning approach, the definition of the widths by groups performed better than having the same width for both groups, while in the geometrical approach, it was the opposite. Looking specifically at the width values, the therapist's widths in the learning and geometrical approaches are similar (0.4 in the learning approach vs 0.43 in the geometrical approach). Instead, the child's widths are very different (1.8 vs 0.32). Thus, increasing the child width to the same value of the female adult augments the accuracy.

Table 5.3: Attention classification system's accuracy for the different hyperparameters configurations in the validation set (%). Bold indicates the highest accuracy for each approach.

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

Attention classification and comparison with the Therapist feedback

In the end we tested our attention system on the several testing sessions, using the best parameters for each approach:

- Geometrical approach: width was the same for both groups; $k = 1; p = 75\%$
- Learning approach: width was different for each group; $k = 2; p = 50\%$

The results are presented in Table 5.4. Globally, our model seems to generalise well for the several test sessions, presenting high scores in all. Moreover, the learning approach performs better than the geometrical approach in all the testing sessions.

In addition, our best result (82%) outperforms the results of the state-of-the-art (73.5%) [112] where a head pose estimator and a constrained scenario were used, very different from the requirements of our clinical partners. Further comparisons with other works were not performed because they did not provide the system performance metrics.

To gain a deeper understanding of the participants' attention landscape, we plotted the azimuth and corresponding elevation angles for both the children and the therapist across all study sessions (Figure 5.9). Comparing our fixation maps to those of Anzalone et al. [78], we observed similar patterns for both neurotypical children and children with ASD. While the therapist appears to divide their attention across three main targets (NAO, computer, and child), the child tends to focus primarily on the robot and the computer, engaging less with the therapist. These plots seem capable of revealing a visual signature of ASD, which could help assess the children's engagement during therapy sessions. However, further testing with more participants is necessary to confirm these findings.

To understand how our system corresponded to real events during therapy, we compared the classifier's accuracy with the therapist's qualitative feedback at the end of each session (Table 5.5). This analysis showed that the children rated by the therapist as having the highest performance—specifically, Child 10 and Child 19—also demonstrated high accuracy in our attention system. Notably, Child 19, who has Level 1 ASD, exhibited gaze behavior similar to neurotypical children, often turning his/her head to look directly at the target. In this way, the highest system performance metric was registered

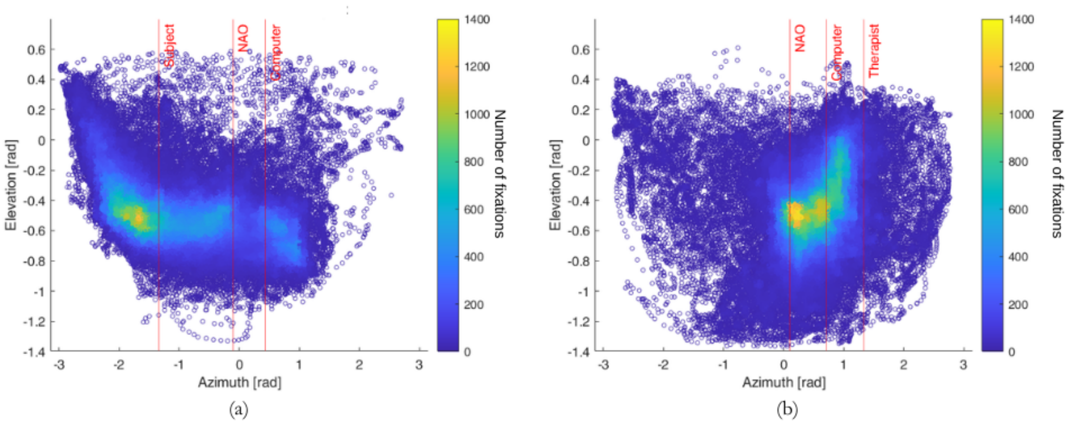


Figure 5.9: Fixation maps of the therapist (a) and the children (b) gaze, considering the fixations of all sessions and all children. The red vertical lines represent the positions of the different targets.

Table 5.4: Accuracy of our attention classification system using the best hyperparameters configuration (%). Session 6 was the validation set for both approaches. Session 3 was the training set on the learning approach, but part of the test set on the geometrical. All remaining sessions were part of the test set for both approaches. Bold indicates the highest accuracy for each session.

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

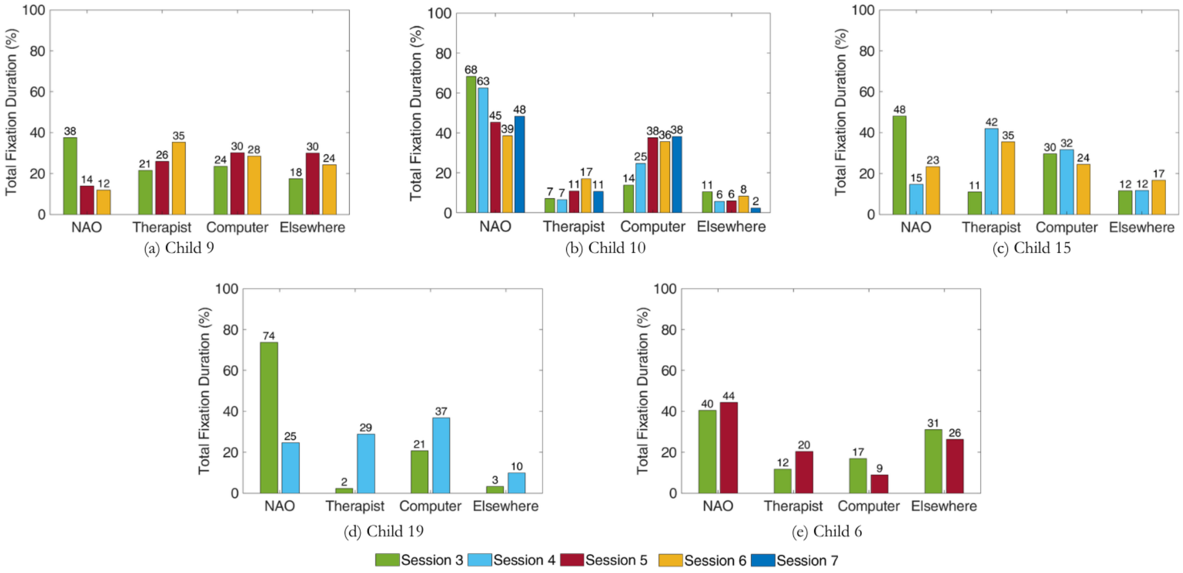
during his/her sessions. In contrast, the children who liked to touch NAO, and consequently moved considerably very close to the Kinect camera, had lower accuracy in our attention system. This accuracy reduction could be justified by the creation of occlusions which led to a wrong estimation/ loss of skeletons.

Table 5.5: Our system attention classification accuracy for each child in each session and therapist's qualitative analysis. Green: Good system accuracy ($> 85\%$) or positive therapist feedback; Yellow: Average system accuracy ($80\% - 85\%$) or neutral therapist feedback; Red: Low system accuracy ($< 80\%$) or negative therapist feedback; P.: Performance; T.: Likes to touch robot (for our attention system, it was a negative characteristic, although it is a positive protocol remark, since the children show interest in the robot).

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

In the end, we calculated the total fixation duration for each child (Figure 5.10), which represents the percentage of time spent fixating each target during a session. For all the children except for child 6, we verified that the attention towards the robot tended to decrease after the first sessions, showing the importance of having a system in which the robot adapts to each child. In addition, generally, the interest in the computer increased. This was expected since after the first session in which the children performed Level 3, the following weeks always included Level 4, in which the children should look directly at the computer. Child 6 did not have this behaviour since in both sessions it performed just Level 3.

Overall, the therapist thought that the figures and the total fixation duration were a good representation of the children's attention. Especially for child 10, the therapist recognised this child as the one with the biggest interest in the robot, reflected in Figure 5.10.



(a) Child 9

(b) Child 10

(c) Child 15

(d) Child 19

(e) Child 6

Figure 5.10: Total Fixation Duration towards the targets and elsewhere along the sessions for Children (a) 9, (b) 10, (c) 15, (d) 19, (e) 6 (%). In Session 3, Level 3 was performed. In the remaining sessions, Level 4 was performed by all children except Child 6 which repeated Level 3.

5.1.6 Discussion

The learning approach achieved better results than the geometrical approach but both models generalized well for the sessions tested, showing the robustness of these approaches. Overall, our attention classification system led to better results in more unconstrained environments than the works described in the literature [112]. The different gaze patterns of the children with ASD and the neurotypical adult were shown in our fixation maps. The children with autism evidenced a lack of interest on the more social stimulus, the therapist, aligning with findings reported in the literature [29].

Furthermore, it showed how an AI algorithm could be applied to produce a metric interpretable by therapists. These metrics can be used for a long-term evaluation of children with ASD during standard and robotic sessions, being a way to comprehend the evolution of these children.

5.2 Online attention classifier

To create a more autonomous robotic system, we aim to integrate the attention classifier into the robot's control loop. This will enable the robot to suggest the next exercises based on the child's attention.

To achieve this, we needed to adapt our attention classifier to operate in real-time. We verified that the step which implied a longer computational load was the one of the face detection. Therefore we explored other face detectors with lower computational time. We selected Yolo and RT-Gene to compare with Densepose, the original face detector of Gaze360. Yolo and RT-Gene are face detectors already used by the head detectors WHENet and RT-Gene respectively. To compare the three algorithms in terms of accuracy and computational time, we considered two types of experiments: one at a short subject-camera distance and the other at a long subject-camera distance.

In the first experiment, we compared the algorithms against the gold standard for eye tracking, the Tobii T60 (Figure 5.11 (a)). Using a chin rest in a fixed position, the subjects looked at 13 different points which appeared separately every two seconds on the Tobii T60 screen. These points are shown in Figure 5.11 (b). We calculated the gaze ground truth by knowing the exact position of the points and the dimensions of the screen. Five different acquisitions were done. Then, we calculated the root mean squared error between the signals obtained with Densepose+Gaze360, Yolo+Gaze360, RT-Gene+Gaze360 and the ground truth.

Regarding the long-distance experiment, we used the data from experiment 2 described in Section 5.1.1 as shown in Figure 5.11 (c) and Figure 5.11 (d). From the room geometry, we established the ground truth and then computed mean squared error with each predicted signal. Additionally, we measured the face detection time to determine which algorithm would be most suitable for our application.

Since these algorithms provided a computational time reduction, we tested if they could effectively be used for the sessions' evaluation. In this scenario, since the computer was replaced by a tablet, we considered just three possible targets: the robot, the other person or elsewhere. As in Section 5.1, two raters labelled the ground truth and the level of agreement was classified using the Cohen's kappa coefficient [167]. This coefficient considers the percentage of agreement weighted by the hypothetical probability of chance agreement, providing a measure of the reliability of the data recoiled. Depending on the value of kappa (κ), the level of agreement can be classified as: almost perfect ($\kappa > 0.9$), strong ($0.8 \leq \kappa \leq 0.9$), moderate ($0.6 \leq \kappa \leq 0.79$), weak ($0.4 \leq \kappa \leq 0.59$), minimal ($0.21 \leq \kappa \leq 0.39$), none ($\kappa \leq 0.20$).

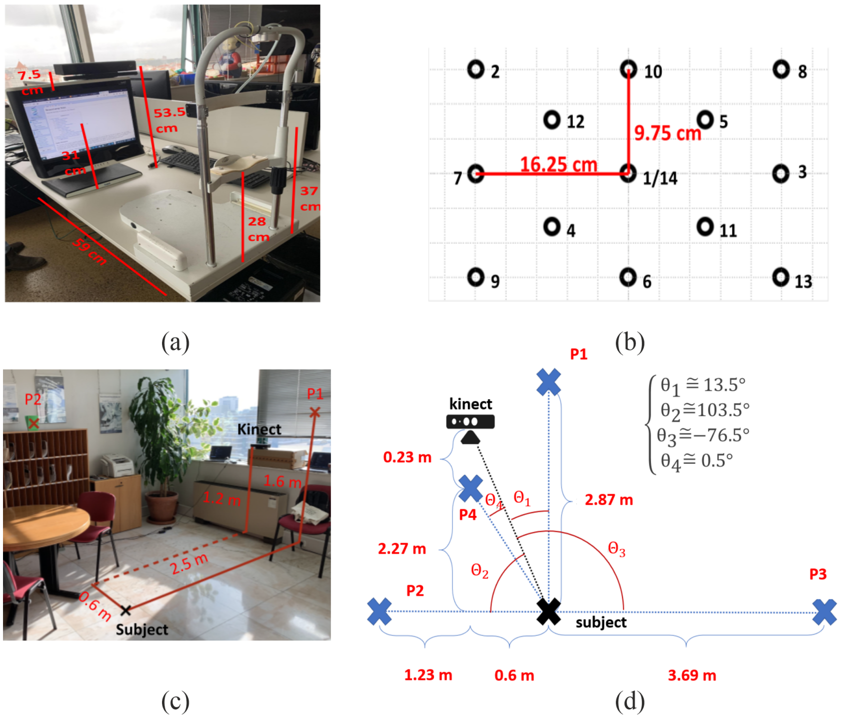
5.2.1 Results

Regarding the conversion of the attention system to an online method, we started by validating different face detection methods, through two experiments, done by one subject. Table 5.6 presents the RMSE results for each signal relative to the ground truth, measured in pixels, as the target points corresponded to elements displayed on the Tobii T60 screen (1280x1064 pixels). We verified that the TobiiT60 as expected for a gold standard provided the best results in terms of accuracy (the lowest RMSE). Densepose which was the algorithm used initially to train Gaze 360 had the best result of the several face detectors followed by Yolo.

Table 5.6: RMSE of short distance validation process (pixel) for the three face detectors tested and comparison with the gold standard Tobii T60.

	Mean±SD (px)
Tobii T60	154±33
DensePose + Gaze360	246±27
YOLO + Gaze360	420±80
RT-Gene + Gaze360	516±131

Similar results were obtained in the long-distance validation process (Table 5.7). In this case, since the targets were in the 3D space, the results are presented in degrees, for the several face detectors.



(a)

(b)

(c)

(d)

Figure 5.11: Benchmark setups for validating different face detection algorithms. (a) and (b) illustrate the short-distance validation process using the Tobii T60, where the points in (b) appeared sequentially as indicated by the numbers on the right, starting and ending at the same point. (c) and (d) depict the long-distance validation process, where the blue crosses represent the points the subject focused on.

As in the short-distance validation process, DensePose was the algorithm with the best mean RMSE, although very close to Yolo. However, Yolo's face detection time is noticeably lower than Densepose, being five times faster. RT-Gene still had the worst results. Therefore, Yolo was chosen as the face detection algorithm.

Table 5.7: RMSE (in degrees) and Face detection time (in seconds) of long-distance acquisitions for the three face detectors tested and comparison with the gold standard Tobii T60.

	RMSE (deg)	Face detection time (s)
	Mean±SD	Mean±SD
DensePose + Gaze360	23.18±2.65	0.335±0.002
YOLO + Gaze360	24.24±1.76	0.064±0.008
RT-Gene + Gaze360	36.31±16.27	0.069±0.001

For the testing with clinical sessions, we used the sessions from two children from the FDG Pilot Study, previously referred in Section 4.2. These sessions took place in November 2020 and January 2021 so in both cases the therapist used a mask and a visor which influenced the perception of the gaze. During the analysed sessions, Child 6, who had a milder degree of Autism compared to Child 7, did three different levels of the hierarchical protocol, while Child 7 only did two different levels. Regarding the ground truth, it was established in 700 frames for each session.

The Cohen's Kappa Coefficient is reported in Table 5.8). Overall, the therapy gaze was easier to classify than the patient gaze, showing larger Cohen's Kappa Coefficients, which was expected from a neurotypical subject when compared with an ASD subject. The patient gaze can be classified as weak agreement for session 2 of Child 6 and session 1 of Child 7, according to [167]. In the same sessions, the therapist gaze had moderate agreement between the two annotators.

In Table 5.9 we present the performance of our attention classifier for each session, each child/therapist

Table 5.8: Inter-rater reliability - Cohen's kappa coefficient. NA represents the session which was not labelled by the second rater.

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

and each target analysed. We can notice a decrease of about 10%-20% compared to Section 5.1.5, which reflects the change of the face detector from DensePose to Yolo. We verified that, in some cases, Yolo was unable to select the face of the subjects, especially when it was not completely visible. On the other hand, the accuracy of the gaze towards the robot tended to be bigger than the accuracy towards the 'other person'. The robot was still while the 'other person' was constantly moving, affecting the accuracy of detection of the Kinect, and consequently, the attention classification system.

Similar to what happened in the APPDA study (Section 5.1.5), Child 6 which had a mild level of autism had better results than Child 7. Child 6 was more collaborative, and maintained some fixed and stable positions, making the detection process easier for the Kinect and Yolo.

Table 5.9: 3D estimation accuracy overview for the two sessions (S1 and S2) of the two children. T represents the therapist, C, the child and R, the robot. The arrow indicates the object of interest, namely T ->C represents the Therapist looking to the Child.

	Child 6 (%)		Child 7 (%)		
	S1	S2	S1	S2	Mean±SD (%)
T -> R	62	77	53	59	63±9
T -> C	57	46	59	49	53±5
C -> R	66	63	65	53	62±5
C -> T	64	58	54	59	59±4

5.2.2 Discussion

Although the validation steps provided encouraging results regarding the replacement of DensePose by Yolo in terms of computational times, the analysis of real clinical sessions shows that the accuracy reduces considerably. Possible causes for the accuracy reduction include the Yolo's difficulty in detecting a face in more extreme scenarios and the smaller number of skeletons detected by the Kinect (or erroneously detected).

Some of the future work will focus on improving the face detection part. A solution could pass by creating a filter that uses previous images in case the new ones are not available. In addition, the Kinect could be included in this process. The skeleton could be used to crop the whole image that would pass directly to the face detector.

5.3 Neural Network for attention classification

From the validation experiments constructed with the face detectors, we concluded that Gaze360+Yolo was the best combination. However, it had a decrease in accuracy compared to the Densepose. Therefore we decided to include more information to improve our accuracy, namely the elevation angle predicted by Gaze360. We followed the same process described in Section 5.1. To understand the advantage of this inclusion, we designed four standardized protocols: the first was based on the scale protocol

(Section 3.4), called S from now on; in the others, two subjects were in two fixed positions and translated in x, y, z corresponding to X,Y,Z protocols.

As we will see in Section 5.3.1, the accuracy improved by a small amount but the precision was low. By looking deeper to the prediction of Gaze360, we verified the presence of offsets, as shown in Figure 5.12. In this figure, the subject was looking at the Kinect, being the expected angle/the standard angle (as explained in Section 5.1.3) 0 degrees. However, the Gaze360 azimuth prediction is around 4 degrees, generating misclassifications in the identification of the subject looking at this target or elsewhere.

Therefore, we designed several protocols for evaluating the different sources of offsets (see [168]). We tested how the offsets changed with time, subject, light, position (subject coordinates), targets, with the signal itself (the Gaze360 prediction) and the respective confidence error. We concluded that the offsets varied in a non-linear way with the subject coordinates, the gaze value and the confidence error angle.

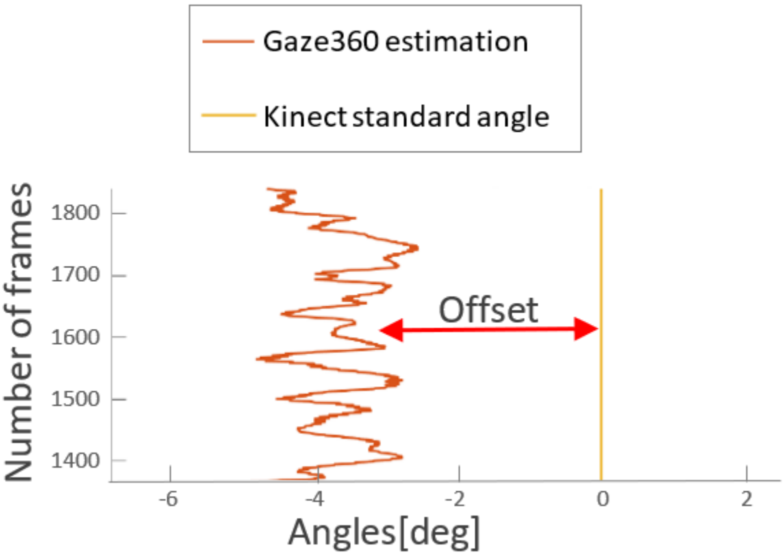


Figure 5.12: Example of offset between predicted Gaze360 and standard angle for Kinect camera in azimuth direction

To correct these offsets and improve the system's accuracy, we developed two methods: an explicit and an implicit. In the explicit method, we assumed that when a subject fixated on a target, the offset for that target would be smaller than for all other targets. Thus, the correct offset (O) at any given moment could be determined by the minimum difference between the Gaze360 estimation (γ) and the standard angles (α) of the various targets (Equation (5.5)). We analysed separately the azimuth and the elevation.

$$O(t) = \min_{\forall target} \gamma(t) - \alpha_{target}(t) \tag{5.5}$$

Thus, for each frame, we added the selected offsets to the lower and upper limits of all the AOIs. Then, we followed the process described in Section 5.1.4 for the establishment of the fixations and comparison with the respective ground truth.

In the implicit method, we designed a neural network to handle the offsets. Given that the subjects were in fixed positions, the neural network could automatically estimate the target they were looking at based on the gaze angle, accounting for the offsets implicitly. As input, we used the gaze angles extracted by the Gaze360, the head coordinates given by Kinect, the bounding boxes pixels extracted by Yolo and two measures of error, the confidence error angle given by Gaze360 and the frame cut according to the bounding boxes pixels. We expected that, if the bounding boxes were not correctly extracted, the image obtained would not contain the face. The output was constituted by 6 targets: 3 posters, NAO, the other person and the Kinect. In this case, we included the Kinect since people spend a considerable amount of time looking at it.

Regarding the architecture, for extracting the best features from the image we used a Pre-Trained EfficientNet, concatenated in two fully connected layers (Figure 5.13). We called the whole architecture CNN. We chose EfficientNet because it was already used in other algorithms for head pose estimation. For the processing of the remaining inputs, we selected a Multi-Layer Perceptron (MLP) with two fully connected layers (Figure 5.14). We tested the importance of the different inputs in three different architectures: MLP alone, CNN alone and MLP+CNN. For all architectures, the loss chosen was given by the error between the predicted and the expected class. In the architecture with the two networks, first, the weights of each network were trained separately and then fine-tuned considering the whole architecture.

We tested two loss functions, mean squared error and cross entropy and different numbers of training epochs from 5 to 100. We also considered the combination of the different inputs in the MLP.

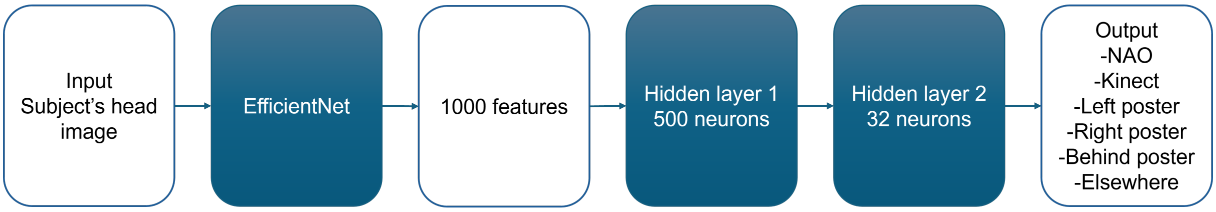


Figure 5.13: Architecture of CNN model.

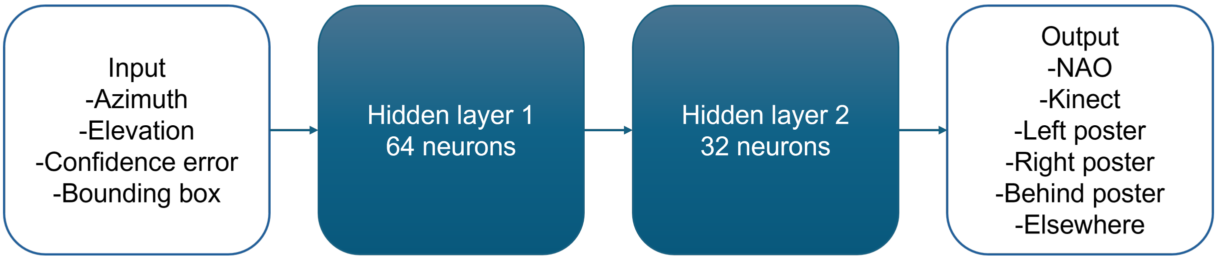


Figure 5.14: Architecture of MLP model.

5.3.1 Results

The protocols S,X,Y, and Z were executed by five different participants in four pairs (neurotypical adults; age: 27 $\pm$ 5 years; 3 females, 2 men). One of the participants was common to all pairs, representing the therapist (Subject 1). For the validation of the hyperparameters, we used the data of Subject 3. Given the low dimension of our sample, we calculated the mean of the cross-validation as the final result, using 4 subjects to train and one to test.

Regarding the first results with the addition of the elevation angle (Table 5.10) both precision and accuracy improved in all subjects, however, it was still not enough for our application.

The explicit method for correcting offsets originated a considerable improvement in the precision. In four of the five subjects, it doubled the precision. The accuracy also increased. However, this methodology failed when: (i) two targets had similar x and y coordinates and their offsets in azimuth and elevation were also similar, being impossible to determine the correct one; (ii) the target was behind the participant, the offset was much larger and could not be calculated in this way.

Table 5.10: Accuracy and precision of the original algorithm and the original algorithm with the elevation computation.

	Original algorithm		Add of elevation computation	
Subject	Precision	Accuracy	Precision	Accuracy
1	0.09	0.74	0.21	0.78
2	0.17	0.76	0.24	0.78
3	0.20	0.77	0.23	0.78
4	0.11	0.74	0.17	0.76
5	0.18	0.76	0.52	0.86
Mean	0.15	0.76	0.28	0.79
SD	0.05	0.01	0.15	0.04

Table 5.11: Accuracy and precision of explicit method solution.

	Explicit method	
Subject	Precision	Accuracy
1	0.58	0.88
2	0.45	0.84
3	0.40	0.83
4	0.44	0.84
5	0.59	0.88
Mean	0.50	0.86
SD	0.04	0.14

Given the wide range of factors influencing the offsets, we developed a neural network model. The training process required 80 epochs for the MLP and 30 epochs for the CNN. While most inputs were beneficial for training the MLP, the head bounding boxes were not. Evaluating the CNN alone yielded lower precision compared to the MLP (66% vs 71%). Consequently, we excluded the CNN from further standalone analysis but combined it with the MLP to explore potential performance improvements.

In Table 5.12, the comparison of the two architectures is shown. Overall, the implicit method had better mean precision than the explicit method. The addition of the CNN was associated with an improvement in accuracy when the validation was done with Subject 1 and Subject 5. Subject 1 was the subject that was common to all pairs, being the subject with the largest amount of data. Therefore, with a shorter dataset, the features provided by the CNN were important to improve the overall result. Subject 5 had the same gender as the subject we used for validating the hyperparameters, thus, better results were expected from this subject. Nevertheless, the sample size should be increased to draw more significant conclusions.

Table 5.12: Accuracy and precision obtained for the Multilayer Perceptron (MLP) architecture and for the combination of the Multilayer Perceptron with the Convolutional Neural Network (CNN) architecture. In the $Subject$ column, additional information about the gender is provided: male (M) and female (F).

	Implicit method			
	MLP		MLP+CNN	
Subject	Precision	Accuracy	Precision	Accuracy
1 F	0.40	0.80	0.48	0.83
2 F	0.60	0.87	0.47	0.82
3 M	0.78	0.93	0.76	0.92
4 F	0.75	0.92	0.67	0.89
5 M	0.64	0.88	0.78	0.93
Mean	0.63	0.88	0.63	0.88
SD	0.15	0.05	0.15	0.05

5.3.2 Discussion

In this section, we demonstrated how a neural network could improve considerably the accuracy and precision of our original algorithm (Gaze360+Yolo). Further study in terms of the network architecture should be done through hyperparameter tuning process, focusing number of hidden layers, number of neurons and activation functions used, which in this work were maintained fixed.

For simplicity purposes, in these protocols, we did not consider the mechanical toy that originally was in the ESCS scale (Section 3.4). A method for real-time targeting of moving objects should be included in the future.

The main limitation of this work is the low data sample, which does not assure the generalization of the results. In the future, this system should be evaluated with neurotypical children and children with ASD to assess its ability to differentiate between the two groups. Comparable metrics to those used in the ESCS scale should be developed to align the automatic measures with established clinical assessments. This approach could transform the system into a quantitative framework for evaluating other pervasive developmental disorders.

5.4 Conclusions

In this chapter, we investigated various strategies for evaluating attention during robotic therapies. We began by exploring an offline approach for gaze estimation, focusing primarily on the Gaze360 algorithm, which at the start of this thesis offered the widest gaze tracking range. The results obtained with children with autism were promising, the attention landscapes (fixation maps) were consistent with the literature, and the calculated metrics aligned well with therapists' assessments. Next, we attempted to adapt the system for online use. During this phase, we observed that Gaze360's performance was influenced by several factors, including lighting conditions and the visibility of facial features, leading to prediction offsets. To address these offsets, we implemented a neural network that achieved satisfactory results when tested with healthy adults.

As discussed in the previous chapter, the neural network solution relied heavily on the studied dataset, which in this case was relatively small, particularly the training set. Future work should include a deeper analysis of the amount of data needed to achieve a more robust solution. Additionally, exploring the feasibility of extending the model to children would be valuable, given that acquiring data from adults is generally easier than from children.

In addition, due to the modular nature of the proposed system, future development may entail adopting novel alternatives to Gaze360. Special focus should be given to the algorithms with fast computational times, to allow a translation towards an online system that could be integrated in the robot controller.

Chapter 6

3D body pose estimation

A significant challenge affecting the performance of the methods discussed in previous chapters was the frequent failure to detect body skeletons due to participants' complex poses and occlusions. Children were often held by therapists, positioned behind the robot, or placed in ways that obscured their limbs. To address this limitation, we explored 3D pose estimation methods based on optical images from the literature. Since most pose estimators are designed for adults, this chapter focuses on adapting these models for children (Section 6.1).

We evaluated our pose estimator against the gold standard and other state-of-the-art methods, both in a constrained scenario (Section 6.1.2), correspondent to our laboratory, and in a clinical therapy scenario (Section 6.1.3). After achieving promising offline results, we extended our investigation to an online implementation of the same algorithm.

6.1 Offline version

We selected the Coherent Reconstruction of Multiple Humans (CRMH) model to estimate the pose of individuals in our clinical sessions due to its specific handling of occlusions, a critical factor in our scenario. The model's key parameter was the focal length, essential for accurately estimating individuals' depth. We determined this parameter through an experiment where a person stood at a known distance, applying Equation (2.4).

Employing this focal length to the clinical sessions' images, we noticed that the child that was at the same depth as the therapist. However, after the processing, he/she was depicted as a small adult at a larger depth. Therefore we calculated a new focal length personalized for each child. Overall, we noticed empirically that the value of focal length varied for people with different heights. Thus, we designed a regression model for estimating the focal length based on the participants' height. The model of CRMH that considered the estimated focal length of each participant was called CRMH-personalized (CRMH-p).

6.1.1 Focal length estimation

The selection of a linear regression model was driven by its applicability both offline and online, thanks to its fast execution times. To build this model, we utilized data from clinical acquisitions of CADIn. Based on the Kinect camera skeletons, we determined the heights of six different therapists. Using frames where the therapists were at known distances, we calculated their respective focal lengths. The selected distances were 2.2m for four therapists, and 2.5m and 3.1m for the other two. These distances reflected the typical depth range where therapists and children spent most of their time during therapy sessions.

To deal with outliers, we used the Random sample consensus (RANSAC) method [169] with the weighted sum of squares as loss function for the decision of the consensus set. The weights were given by the uncertainty associated with each measurement taken for establishing the focal length. This uncertainty was related to the perspective projection and can be described according to Equation (6.1), where h is the person's height, Z is the respective depth, and f denotes the focal length.

$$y = \frac{hf}{Z} \iff \frac{\delta y}{\delta Z} = -\frac{hf}{Z^2} \iff \delta Z = -\frac{Z^2}{hf}\delta y \tag{6.1}$$

The final loss function is explicit in Equation (6.2) where $w_i = \frac{1}{\sigma_i^2}$, $\sigma_i \propto Z_i^2$ and f_i represents the focal length measured for each therapist i and $\hat{f}_i$ the focal length estimated by the regression model.

$$L = \frac{\sum_i w_i (f_i - \hat{f}_i)^2}{\sum_i w_i} \tag{6.2}$$

6.1.2 Controlled Experimental Setting

To test the accuracy of the RANSAC model constructed, we established an experiment in a controlled scenario with three subjects with different heights: 1.42m (a child), 1.62m (an adult) and 1.75m (an adult). We compared our CRMH-p with a motion capture system, the Optitrack system, and the other monocular system present in literature, the Bird-Eye View (BEV) model (Section 2.3.1), specifically designed for the reconstruction of children. We focused our analysis on the coordinates of the skeletons reconstructed by the CRMH-p and BEV and compared them with the Optitrack measurement. The Kinect reconstruction algorithm was not considered in this comparison since the camera interfered with the Optitrack system.

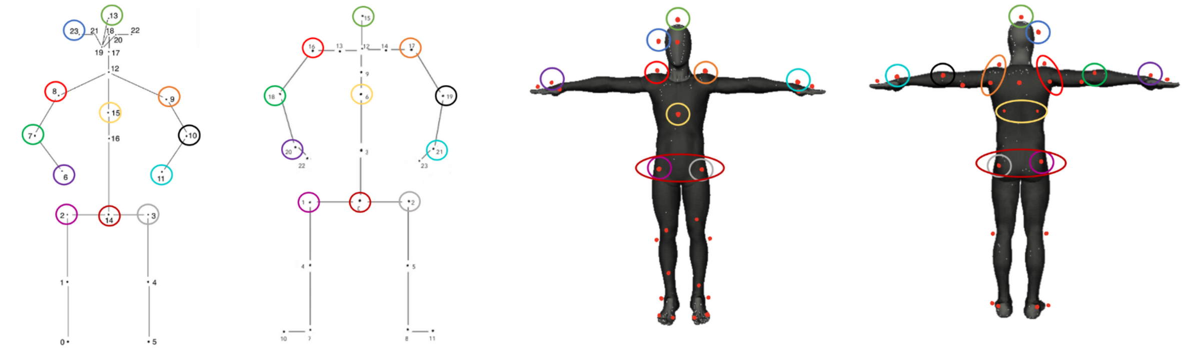
Each participant started at 1.3m with their arms opened to be detected by the Optitrack system. After maintaining this position for 5 seconds, the participant took four different positions walking backwards: 1.9m, 2.2m, 2.5m, and 3.1m.

The metric used was the 3D average Root mean squared error (RMSE) given by calculating the RMSE for each frame (considering selected joints), according to Equation (6.3) and then doing the average across all frames. In Equation (6.3) N represents the total number of joints selected from the skeleton (Figure 6.1), (x_i, y_i, z_i) represents the coordinates of each joint position from the ground truth, and ($\hat{x}_i$, $\hat{y}_i$, $\hat{z}_i$) represents the coordinates of each corresponding estimated joint position from the proposed model (BEV or CRMH-p).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 + (z_i - \hat{z}_i)^2)}, \tag{6.3}$$

In addition, we analysed specifically the z coordinate of the hip joint in the selected models and compared it with the CRMH original and with the gold standard to understand the differences. We focused on this joint because it had the most clear movement.

As a secondary metric, we measured the computational time of each algorithm by evaluating the number of frames processed per second. This assessment helped determine their feasibility for real-time application.



(a) Skeleton from CRMH (b) Skeleton from BEV (c) Optitrack markers' scheme (front view) (d) Optitrack markers' scheme (back view)

Figure 6.1: Skeletons' scheme of the systems. The correspondences between joints used to evaluate are marked by circles of the same colour. Since the systems' skeletons are not a perfect match, for some joints (for example 14 in (a)) we had to use the mean of two associated markers.

6.1.3 Real World Setting

To test the CRMH-p system in the real world, we selected 4 clinical sessions different from the ones used to train the RANSAC model. Each session had a different child and a different therapist. We estimated the focal length of the child and the therapist based on their heights (extracted with the Kinect camera). For the application of the focal length in the model, we identified the therapist based on his/her blue suit. The child was the other person in the room.

We started by evaluating the reconstruction capability of the three tested algorithms (Kinect, CRMH-p and BEV). More precisely we calculated the percentage of frames in which both the skeleton of the child and adult were correctly identified. During the therapy sessions, we could not have a 3D pose ground truth since the therapy rooms do not have an Optitrack system. Therefore, the comparison was made against 2D ground truth silhouettes from Segment Anything algorithm [170]. This algorithm provides segmented masks based on prompts. For each frame, we obtained a silhouette for the child and one for the adult.

For a measure of accuracy, we transformed the 2D skeletons of the three algorithms into silhouettes through a dilation process and compared them with the ones of the Segment Anything algorithm. In this case, we were interested only in the orientation therefore we aligned the centroid of the segmented mask with the ones of the tested algorithms. For the frames that had both silhouettes, we applied the Dice similarity measure (also called Dice coefficient) [171, 172] to compare each dilated mask (Kinect, CRMH-p and BEV) with the ground truth. Moreover, we verified that this similarity measure was maintained in the frames that were correctly reconstructed by CRMH-p and BEV and not by the Kinect.

6.2 Online version

Regarding the implementation in real-time, the main difficulty was the CRMH-p implementation which was done inside a Docker container. The usage of Docker facilitated the interoperability between different working environments which was essential for our analysis but made the communication with our local system harder. This communication is fundamental in a real-time system since we want to associate each calculated skeleton with the respective session's frame.

For this communication, we established and tested two different approaches. First, we used the RabbitMQ message broker. This system allows quick and reliable data exchange between the Kinect and the Docker container. The Kinect sends the frames through RabbitMQ to the Docker, and after processing, the Docker sends the skeleton results back. As an alternative, we constructed a data-shared system through files. The two systems were connected to a directory in which they could access and share files. In addition, we tested two alternatives for the processing, sequential and parallel. In sequential processing, each frame was analysed one at a time, while in parallel processing, multiple frames were sent to different threads and processors and the respective skeletons were obtained.

6.2.1 Evaluation experiments

We tested three different architectures: Sequential Processing with RabbitMQ communication; Parallel Processing with RabbitMQ communication and Sequential Processing with direct communication through mounted directories (Figure 6.2). First, we evaluated the systems regarding the frame rate of the videos processed. We analysed the consistency of each system's performance, the robustness to the different conditions present in therapy and the viability of a real-time application. Second, we calculated the time delay between the capturing of the images by Kinect and the receiving of the skeletons by the docker. In addition, we tried to understand how much was the communication time and the processing time. In the end, after choosing the best system, we analysed the performance when connecting the full system with NAO and the gesture recognition model, during five sessions following the bingo protocol.

As explained in Section 4.2, the gesture recognition system was trained using keypoints extracted from the Kinect. To enable analysis with keypoints from the CRMH-p system, we incorporated an encoder model that converts CRMH-p coordinates into Kinect coordinates. The encoder was constituted by a feed-forward neural network with 72 nodes as input, representing the 3D coordinates of the 24 joints extracted by the CRMH-p model, and 75 nodes as output for the Kinect skeleton representation. The hidden layer was formed by 64 nodes. Regarding the other hyperparameters we used the rectified linear unit as the activation function, the Adam optimizer and the RMSE loss.

6.3 Results

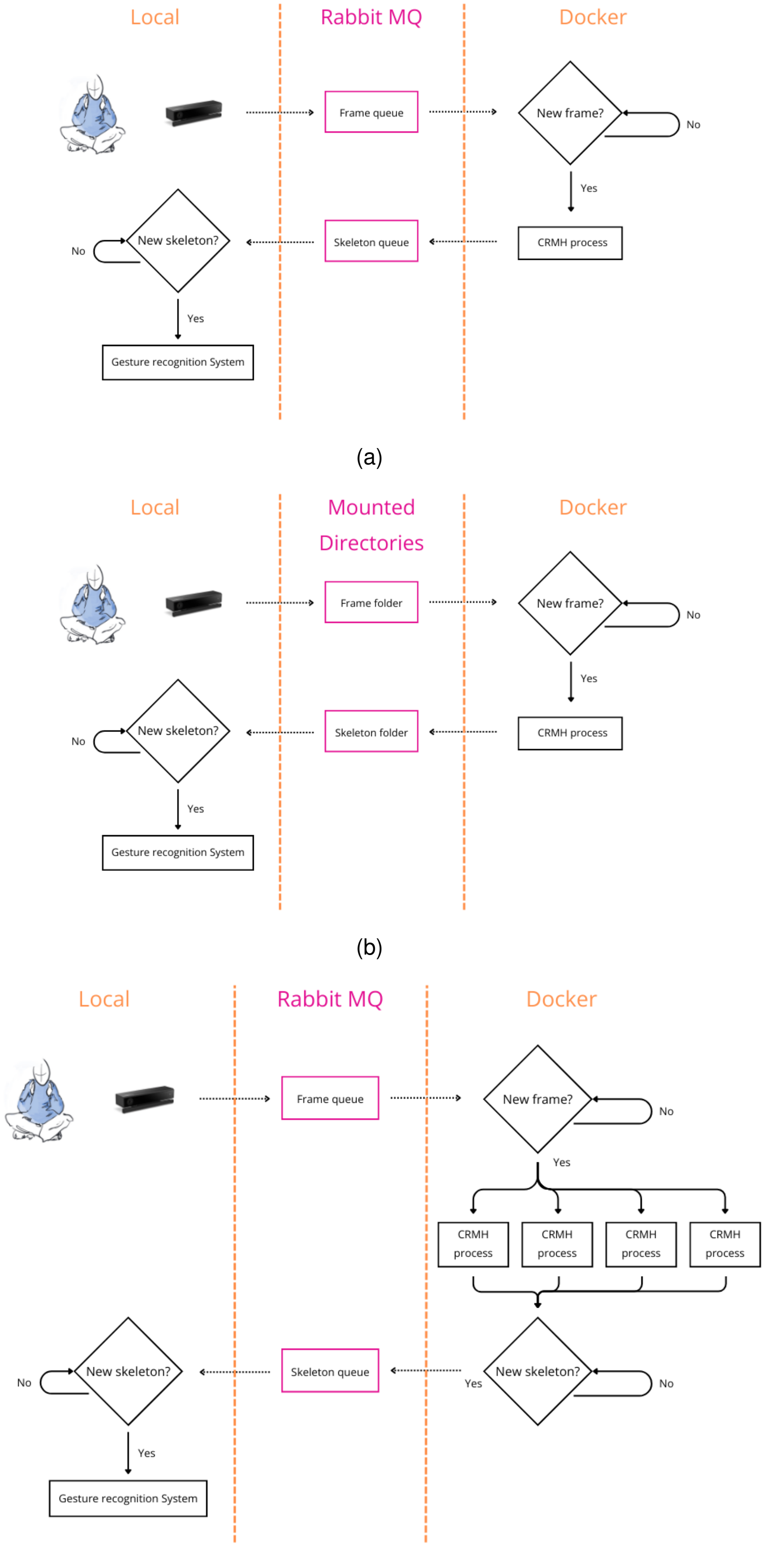
We present the performance results of adapting a 3D body pose estimator to children. First, we establish the model for the focal length estimation, then show the results in a controlled environment and a real world setting. In the end, we demonstrate how an online version of it can be consistent and robust in computational times terms.

6.3.1 Offline version

First, we used the RANSAC method to construct a model to estimate the focal length based on the heights of the subjects (Table 6.1). Accounting for measurement uncertainty in the subjects' heights was crucial to enhance the model's robustness. In fact, the weighted sum of squares strategy produced more consistent models, resulting in a higher determination coefficient R^2 compared to the standard sum of squares approach.

Applying this model to the three participants in the controlled experiment revealed differences in focal length between the adults and the child (Table 6.2). The adults' focal lengths were approximately 400, while the child's was 367. Additionally, the focal length values were correlated with participants' heights, with taller individuals having larger focal lengths.

Moreover, in the controlled scenario, our method considerably improved the performance of CRMH, getting much closer to the gold standard, the Optitrack (Figure 6.3). It is less accurate in the zones of 1.2m and 3.1m but, in the other zones, it follows the ground truth closely. Instead, the current state-



(a)

(b)

(c)

Figure 6.2: Architectures tested for evaluation of the performance and decision on the integration in a real-time system: (a) Sequential processing using RabbitMQ system; (b) Sequential processing using Mounted Directories; (c) Parallel processing using RabbitMQ system. In all cases, the goal was to transmit frames captured by the Kinect to Docker for processing with the CRMH-p algorithm to extract the participants' skeletons.

Table 6.1: Results from the RANSAC model using two different loss functions (sum of squares or weighted sum of squares). Each row corresponds to a linear model ($f = Slope \times height + Intercept$) generated by the RANSAC model with a coefficient of determination R^2. The selected model is underlined in green.

	R^2	Slope	Intercept
Sum of Squares	0.7660	250.51	-4.51
	0.7548	271.56	-38.58
	0.9780	696.14	-750.66
Weighted sum of squares	0.9943	158.20	145.72
	0.9959	164.47	135.23
	0.9959	164.47	135.23

Table 6.2: Height and Focal Lengths of participants in the controlled scenario.

Height (m)	Focal Length
1.41 (Child)	367.13
1.62	401.67
1.75	423.05

of-the-art method, BEV (Section 2.3.1), is more similar to the ground truth in the first zone, when the subject is closer to the camera but presents an offset that seems to increase with the depth.

In a more focused analysis in the working zone (where most of the therapies happen), we observe that our method is better than BEV, regarding both the RMSE and computational time. It is more accurate and faster (Table 6.3). Nevertheless, both methods perform better for the tallest people, probably because this is the average height in the datasets used for training. Overall, the error obtained is below one body width ($0.3m$) that is acceptable to our application [3] and a part of this error is probably related to the different joints representations of the 3D estimators in relation to the ground truth (Figure 6.1).

Table 6.3: 3D RMSE results and processing rate for the CRMH-p and BEV model for each participant in the depth $1.9m < z < 2.5m$ corresponding with the region where therapies take place. The best results for each height are marked in bold.

	RMSE (m)		FPS	
Height (m)	CRMH-p	BEV	CRMH-p	BEV
1.41	**0.20**	0.35	**8.02**	5.25
1.62	**0.23**	0.27	**7.96**	5.24
1.75	**0.14**	0.19	**7.90**	5.20

Concerning the clinical sessions analysed, the focal lengths for the children and therapists are presented in Table 6.4. The therapists had almost two times the focal length of the children, which shows the importance of the regression model estimations. Both methods, BEV and CRMH-p were able to recover a significant number of ground truth skeletons that were lost by the Kinect (Table 6.5). BEV obtained the largest percentage of detection of two skeletons for three of the four children.

Table 6.4: Focal lengths estimations for the children and adults analyzed in the clinical sessions. The names of the children and therapists are represented by a code for anonymization purposes.

		Focal length			Focal length
Children	17	282.45	Therapists	PJ	414.83
	08	290.21		ARG	411.54
	02	274.68		SF	398.38
	11	309.62		MJC	403.32

However, in terms of mean Dice similarity, the CRMH-p had the highest similarity with the ground

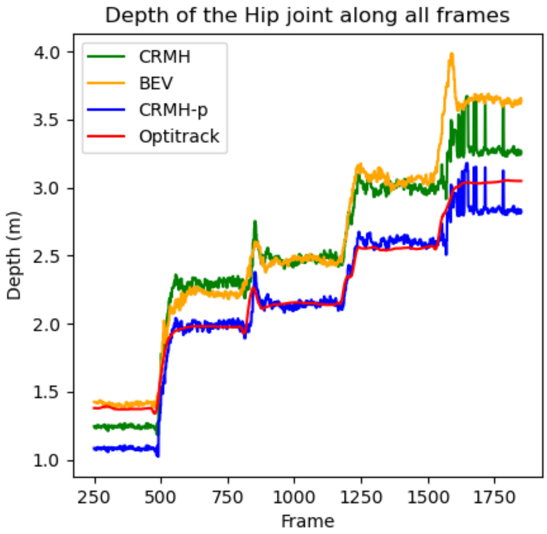


Figure 6.3: Depth values for the child's hip joint using different systems: CRMH, BEV, CRMH-p and Optitrack. The proposed model improved the performance of the original CRMH. In the middle depth range ($1.9m < z < 2.5m$), the accuracy of the proposed model is notably high.

Table 6.5: Percentage of ground truth skeletons in which both skeletons were detected by the Kinect and the proposed model. Each row represents one session done by a therapist (characters' code) and a child (numerical code). The best results for each pair (therapist-child) are marked in bold.

Two skeletons	% Kinect	% CRMH-p	% BEV
MJC 17	12	92	**99**
PJ 08	10	75	**87**
SF 02	57	77	**99**
ARG 11	83	**94**	73

truth in 75% of the analysed cases, as shown in Table 6.6. It is also clear that there is increased difficulty in predicting the child's pose since the similarity metric is almost always larger for the therapist than for the child.

Table 6.6: Mean Dice similarity between the skeletons of the ground truth and the ones of the analysed system (Kinect or CRMH-p) with the centroid correction. Each row represents one session done by a therapist (characters' code) and a child (numerical code). The best results for each child and each therapist are marked in bold.

	Therapist			Child		
	Kinect	CRMH-p	BEV	Kinect	CRMH-p	BEV
MJC 17	0.092	**0.228**	0.189	0.150	**0.165**	0.128
PJ 08	0.232	**0.242**	0.235	**0.227**	0.216	0.198
SF 02	0.225	**0.239**	0.223	0.227	**0.234**	0.192
ARG 11	0.241	**0.245**	0.223	**0.228**	0.227	0.201

Additionally, focusing on the skeletons reconstructed by the CRMH-p but not detected by the Kinect (Table 6.7), their average Dice similarity is comparable to that of the skeletons identified by both methods. This demonstrates the algorithm's ability to generalize effectively, even in scenarios where the Kinect fails.

Table 6.7: Mean Dice similarity for the skeletons identified by the proposed model (CRMH-p) but not by Kinect. Each row represents one session done by a therapist (characters' code) and a child (numerical code).

	CRMH-p	
	Therapist	Child
MJC 17	0.245	0.149
PJ 08	0.236	0.188
SF 02	0.231	0.202
ARG 11	0.231	0.181

6.3.2 Online version

As mentioned before, one of the communication architectures used to transform the 3D pose estimator into an online system was the message broker RabbitMQ. We tested two different processing schemes: sequential and parallel. Parallel processing was expected to be faster than sequential processing, as it handles multiple frames simultaneously. In addition, parallel processing approach was anticipated to introduce a larger initial delay compared to sequential processing, which would later be offset by faster overall processing times.

Our results show that parallel processing outperforms sequential processing in terms of Frames per second (FPS), albeit with a relatively small margin (Table 6.8). The limited gain can be attributed to the computational overhead and communication demands associated with parallel processing. Furthermore, sequential processing demonstrated greater consistency, exhibiting a standard deviation nearly five times lower than that of parallel processing.

Table 6.8: Comparative Frames per second (FPS) analysis between sequential and parallel processing using RabbitMQ communication system.

Session	Sequential (FPS)	Parallel (FPS)
1	7.51	8.75
2	7.64	8.23
3	7.62	8.97
Mean	7.59	8.65
SD	0.07	0.38

Focusing on the delay (Table 6.9), both the mean communication and the processing delays were larger in the parallel processing, but particularly important is the standard deviation of the communication delay which is almost ten times larger than the one of the sequential processing. The larger communication delay in the parallel processing can be caused by the managing of simultaneous frames that do not happen in the sequential processing. Given the need for reliability in our sessions, we decided to choose the sequential processing approach.

Table 6.9: Comparison of communication, processing, and total delay times across sequential and parallel systems using RabbitMQ communication system.

	Sequential		Parallel	
	Mean (s)	SD (s)	Mean (s)	SD (s)
Communication	0.027	0.018	0.063	0.112
Processing	0.104	0.015	0.163	0.057
Total Delay	0.130	0.012	0.226	0.143

Regarding the comparison between RabbitMQ and the Mounted Directories with sequential processing as the processing system, RabbitMQ outperformed the other architecture in the three analysed sessions (Table 6.8). In addition, the mean delay times and their standard deviations were smaller than in the Mounted Directories architecture (Table 6.11). Both architectures use sequential processing, thus they have similar processing time delays. In contrast, the communication time delay of the mounted directories is nearly twice that of RabbitMQ. Probably this increase in the time delay is due to the opening, writing, reading and closing processes inherent to this type of architecture, that are not present in RabbitMQ. Therefore, we chose the RabbitMQ architecture.

Table 6.10: Comparative Frames per second (FPS) analysis between RabbitMQ and Mounted Directories techniques using sequential processing

Session	RabbitMQ (FPS)	Mounted Directories (FPS)
1	7.51	6.85
2	7.64	7.06
3	7.62	6.73
Mean	7.59	6.88
SD	0.07	0.17

Table 6.11: Comparison of communication, processing, and total delay times between RabbitMQ and Mounted Directories techniques using sequential processing.

	RabbitMQ		Mounted Directories	
	Mean (s)	SD (s)	Mean (s)	SD (s)
Communication	0.027	0.018	0.041	0.042
Processing	0.104	0.015	0.109	0.0398
Total Delay	0.130	0.012	0.150	0.031

As the final step, we recorded five sessions following the bingo protocol. Table 6.12 and Table 6.13 summarise the measurements of this last experiment reflecting the practicality of this approach in real time. Comparing with Table 6.8 and Table 6.11, we notice that the framing rate decreases and both the communication and processing time delays increase. These results were expected since the gesture recognition system was running during these sessions. Nevertheless, the mean total delay time is 0.198 s, which still guarantees the therapeutics' session rhythm, and delivering timely feedback and interactions, based on our experience.

Table 6.12: Frames per second (FPS) analysis of the full system using sequential processing and RabbitMQ communication technique.

Session	Full System (FPS)
1	5.05
2	4.81
3	4.92
4	5.11
5	5.01
Mean	4.98
SD	0.12

6.4 Discussion and conclusions

Our protocols primarily focused on movement, with a particular emphasis on gestures, and measuring these in a non-intrusive manner was always a top priority. However, we soon realized that the Microsoft Kinect, the sensor we had chosen, frequently lost track of skeletons due to the intrinsic characteristics of our therapy. This prompted us to search for an alternative solution, like 3D models extracted directed from optical cameras.

Table 6.13: Communication, processing, and total delay times for the full system using sequential processing and RabbitMQ communication.

	Mean (s)	SD (s)
Communication	0.063	0.007
Processing	0.135	0.009
Total Delay	0.198	0.013

Since 3D pose estimators tailored for children are rare in the existing literature, we personalised an existing 3D pose estimator for adults to cope with children by manipulating a parameter, the focal length. The focal length was estimated through a linear model. As a main advantage, this linear model was just dependent on the person's height and allowed an implementation in real time of the same pose estimator. Moreover, it recovered a considerable number of skeletons in clinical sessions already registered with Kinect and presented good similarity metrics in relation to the ground truth both in controlled scenarios and clinical sessions. In controlled scenarios, it is characterized by an error lower than $0.3m$ when compared with the Optitrack system and it is always lower than the recent model BEV. In the analysed clinical sessions, the mean Dice similarity in relation to the ground truth is very close to the values of the Kinect and these results are maintained in the frames that are not predicted by Kinect.

For the online pose estimation, two main therapy requirements guided the choice of our architecture and processing system: reliability and velocity. Therefore we used RabbitMQ for the communication between the camera and the Docker container where our pose estimator ran and we chose to analyse each frame sequentially due to the increased consistency in relation to the Parallel processing. This approach achieved an average frame rate of 4.98 FPS in a scenario resembling the therapy sessions, enabling timely feedback and interactions.

Future work will focus on implementing this model in real clinical sessions to test it in scenarios with considerable occlusions. It will also be crucial to assess whether processing time decreases when the gesture recognition algorithm operates in prediction mode. A comprehensive evaluation of the new gesture algorithm, along with its associated encoder, should be conducted to identify potential errors arising from the conversion of CRMH skeletons to Kinect-type skeletons. Ultimately, a gesture recognition model capable of directly utilizing CRMH-p skeletons should be developed.

Chapter 7

Clinical Studies

We describe the two main clinical studies carried out during this thesis and their respective results. We decided to explore these studies since they had a larger number of participants and testing weeks than the other studies described in Chapter 3. The first is a Pilot Study done in Fondazione Don Carlo Gnocchi (FDG) in Milan in 2020, in which we present just the clinical results. The second is a Randomized Controlled Trial that occurred in Lisbon at Centro de Apoio ao Neurodesenvolvimento (CADIn) in 2022, in which we present the results of the clinical scales used and the results of our developed metric, the Total Fixation Duration (TFD). In both studies, the hierarchical protocol (Section 3.2) was chosen.

7.1 A Pilot Study

The Pilot Study at FDG was our first pilot study with more than two children. It started in February 2020 but was interrupted in March 2020 due to the global Covid-19 pandemic. It restarted in November 2020 and lasted till September 2021. We included children with a diagnosis of ASD according to the Diagnostic and Statistical Manual of Mental Disorders-5th edition (DSM-V) and confirmed with the Autism Diagnostic Interview-revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) (Section 2.1), less than 6 years old at the time of the enrolment. Children with preterm birth, pregnancy complications or perinatal injury history, major facial peculiar characteristics, malformations or neuro-radiologic alterations, epileptic syndromes, known congenital infections, metabolic or genetic diseases were excluded. In the end, 14 children were enrolled.

From the initial sample of 14 children, there were three dropouts due to familiar problems. One of the children did not undergo the full evaluation process so he/she was excluded from the final sample. Thus, the final sample included 10 children, 80% males, and a median[1st quartile;3rd quartile] age of 52[51.25; 65] months. All children had developmental delay. Each of these children performed the hierarchical protocol during 14 weeks, in weekly sessions. Each session had an approximate duration of 20-30 min depending on the child's engagement. The passage from one level to another inside the hierarchical protocol was determined by the therapist according to the performance of the children in each of the levels and not pre-defined by a number of weeks. Participants were not restricted from receiving other interventions during their involvement. The study was approved by the Ethical Committee of FDG (number 6_25/07/2019) and all parents signed an informed consent.

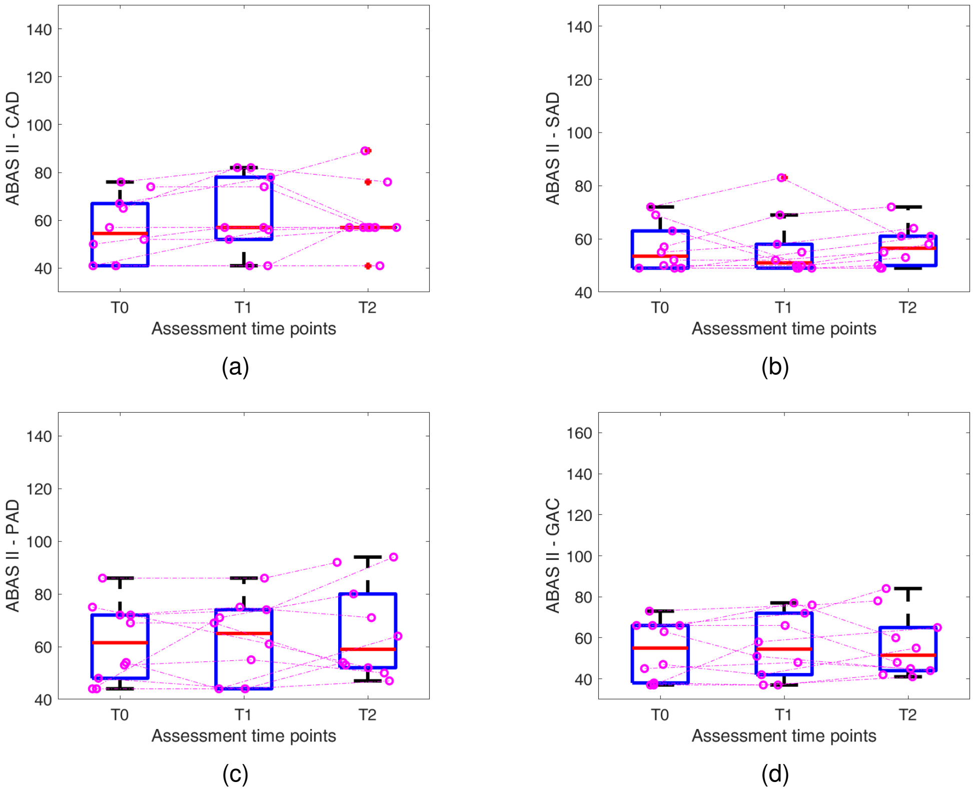
The children were evaluated in three different moments: T0 (baseline), T1 (end of the intervention protocol) and T2 (6 months after the end). The outcome measures used were:

- Griffith's – III Edition: a clinical scale for assessment of overall level of development in children younger than 6 years old. It is constituted by 5 subscales (Foundations of Learning, Language and Communication, Eye-hand coordination, Personal-Social-Emotional, Gross motor). We considered the scores of the Language and Communication subscale and the General Quotient.

- ABAS-II: a questionnaire filled by the caregiver to assess adaptive skills in everyday life; it comprises three domains, Social (SAD), Conceptual (CAD), and Practical (PAD) Adaptive Domains. The scores of these domains are combined to create the General Adaptive Composite Score (GAC).

- MacArthur-Bates Communicative Development Inventories – Words and Gestures (MB-CDI) (Italian adaptation): a questionnaire filled by the caregivers, in which we just considered the part related to gestures (number of gestures, that contains: (A) first communicative gestures, (B) games and routines, (C) actions with objects and imitation of the adult, (D) games of pretending and (F) games of pretending with objects.)

All these scales are scored based on the number of items each child successfully completes, with higher scores indicating greater developmental capabilities. The results for the several scales are shown in Figure 7.1 and Figure 7.2. Visual analysis revealed no time-related patterns in the scores of the children who participated in the Pilot Study. Nonetheless, we conducted a statistical test to quantitatively assess potential longitudinal differences across the pre-, post-, and follow-up phases. Given the small sample size, we used the non-parametric Friedman test. However, no significant differences were found.



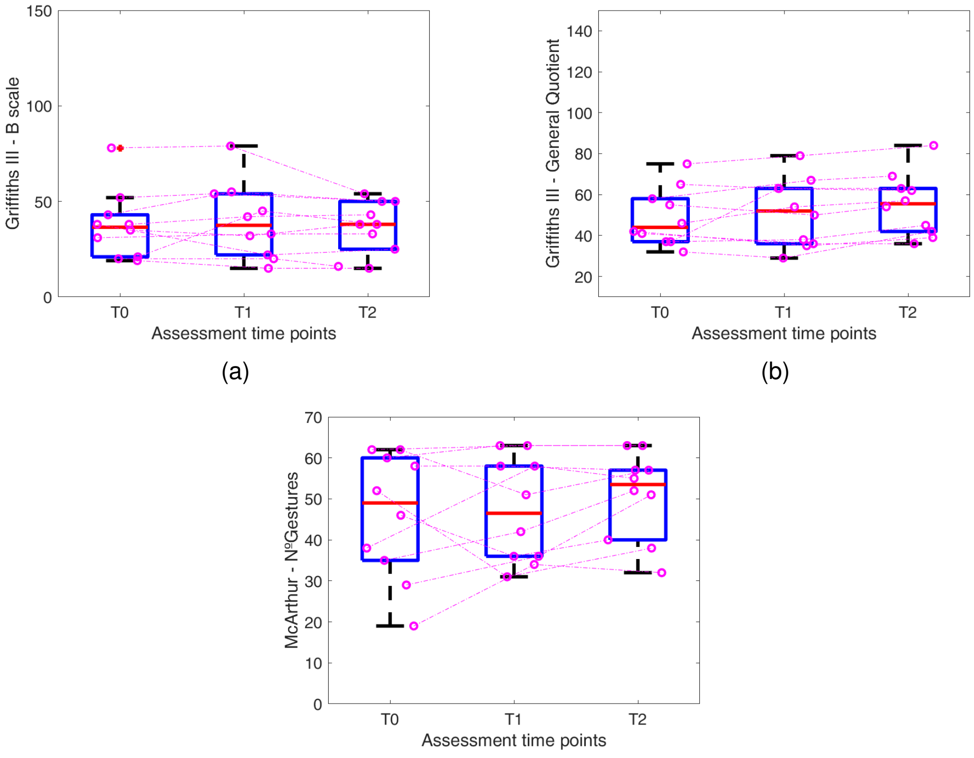
(a)

(b)

(c)

(d)

Figure 7.1: ABAS-II Scores at different time points [T0 (baseline)-T1 (end of the intervention protocol)-T2 (6 months after the end)] for the (a) Conceptual Adaptive Domain (CAD), (b) Social Adaptive Domain (SAD), (c) Practical Adaptive Domain (PAD), (d) General Adaptive Composite Score (GAC).



(a)

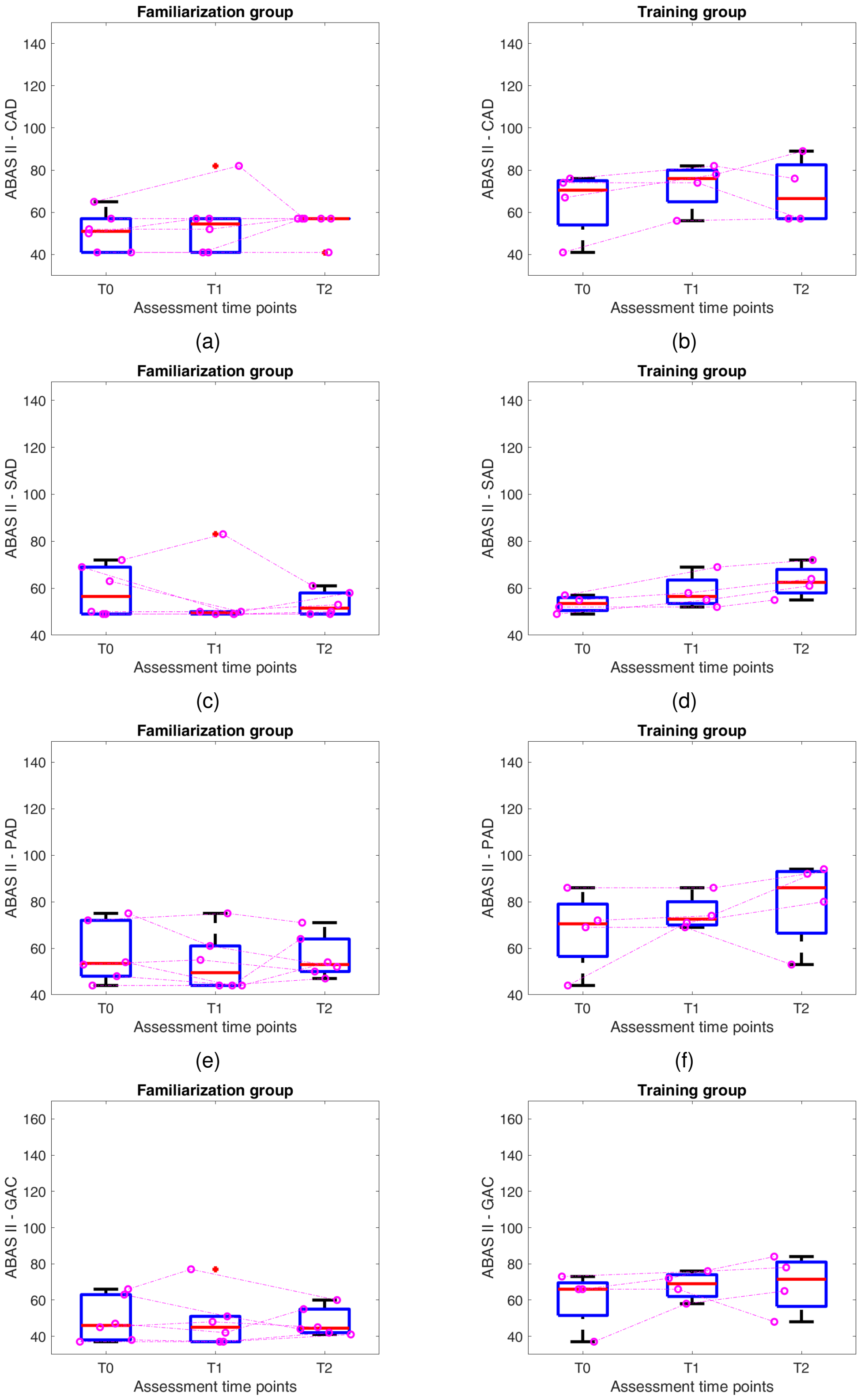
(b)

(c)

Figure 7.2: Griffith-III scores at different time points (T0-T1-T2) for (a) the Communication Scale (B scale), (b) the General Quotient and (c) the number of Gestures performed in the MacArthur's scale.

Following other works in robotics for autism [22, 21], we chose to divide our sample into groups to deal with the heterogeneity intrinsic in Autism as described in Section 2.2.3. This resulted in two groups: the one that progressed to the training levels (training group), and the one that remained in the familiarization levels (familiarization group). This implied a further reduction of each sample size (4 in the training group and 6 in the familiarization group). The results are shown in Figure 7.3 and Figure 7.4. Through a non-parametric unpaired test, the Mann-Whitney U Test, we verified that at T0 the two groups were significantly different in terms of developmental and adaptive skills like ABAS-II and the Griffiths-III ($p < 0.05$) but not in age and level of autism. The figures clearly show that the training group generally had higher scores than the familiarization group at T0, indicating differences in developmental capabilities and supporting our decision to separate the groups to reduce heterogeneity.

Furthermore when analysing the results and the respective medians of the two groups, for all the scores of the training group, the median improved, while in the familiarization group just the CAD domain of ABAS-II and the communication scale of Griffiths-III have improved. Within the training group, the Friedman test revealed a significant result in the SAD domain of ABAS-II ($p = 0.022$), indicating longitudinal differences across the pre-, post-, and follow-up phases (T0 vs T1 vs T2). However, further analysis using the non-parametric paired Wilcoxon signed-rank test to compare individual time points (T0 vs T1, T1 vs T2 and T0 vs T2) revealed no significant differences.



(a)

(b)

(c)

(d)

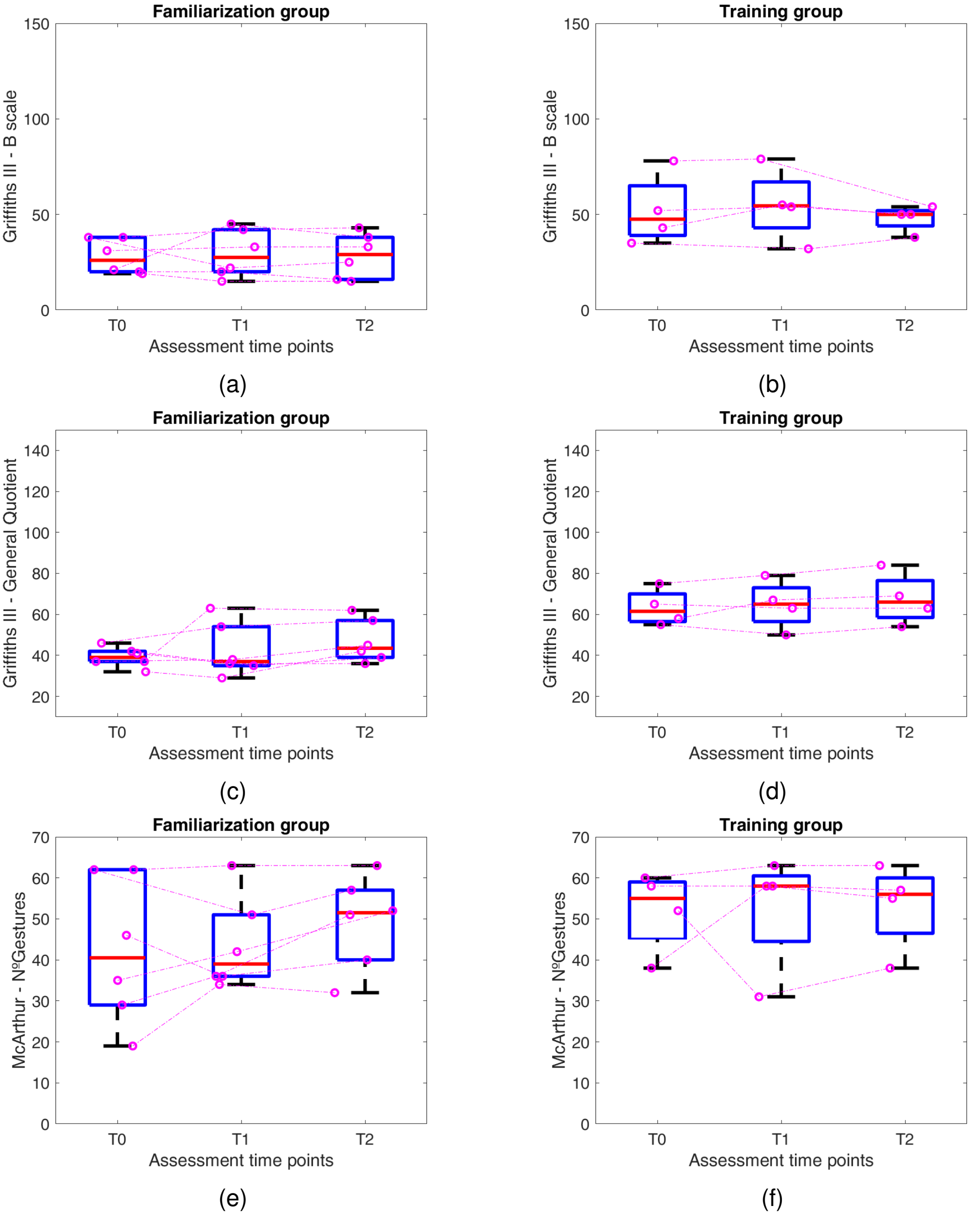
(e)

(f)

(g)

(h)

Figure 7.3: ABAS-II Scores for the familiarization and the training groups at different time points (T0-T1-T2) for the (a,b) Conceptual Adaptive Domain (CAD), (c,d) Social Adaptive Domain (SAD), (e,f) Practical Adaptive Domain (PAD), (g,h) General Adaptive Composite Score (GAC). The red asterisks represent the outliers.



(a) (b) (c) (d) (e) (f)

Figure 7.4: Griffith-III scores for the familiarization and the training groups at different time points (T0-T1-T2) for (a,b) the Communication Scale (B scale), (c,d) the General Quotient and (e,f) the number of Gestures performed in the MacArthur's scale.

Although these results are preliminary and do not allow strong conclusions, this study constituted an important milestone for this thesis since the full robot system was tested by 10 children during 14 weeks. Moreover, we verified the heterogeneity problem described previously by [21, 22]. We think that the differences found between the group that reached the training levels and the group that remained in the familiarization levels were due to the several verbal prompts given by the robot during the protocol. We hypothesize that these prompts are hard to understand and manage by children with severe developmental delay and contribute to a decrease in their engagement during the study. The conclusions of this study were instrumental in shaping the sensorial protocol (Section 3.3), a new protocol developed by Fondazione Don Carlo Gnocchi (FDG). This updated protocol increases the variety of stimuli and incorporates simpler prompts to facilitate easier initial understanding for the children.

7.2 A Pilot Randomized Controlled Trial

In this subsection, we present the results from a Pilot Randomized Controlled Study realized in CADIn association. This study took place from May to June of 2022 with children between 2 and 6 years old, with a prognosis of ASD. Due to their young age, these children presented all ASD characteristics described in DSM-V, but not an official diagnosis. The study lasted nine weeks with one session of 10 min each week. Initially, we recruited 20 children assigning randomly 10 children for the experimental group (robotic group) and 10 children for the control group, with an equal percentage of verbal and non-verbal children. The experimental group has done the hierarchical protocol while the control group has done a revised version of the hierarchical protocol in which both the therapist and the robot were present but all the tasks were done by the therapist. As in the previous study, participants were not restricted from receiving other interventions during their involvement. This study was approved by the Ethical Committee of CADIn, and all parents signed an informed consent.

The primary outcome measure used in this study was the Language Use Inventory (LUI), as it is the main social communication scale utilized by the CADIn team. Other outcome measures were not regularly employed by the CADIn team, and our new scale metric based on the ESCS was still under development. The LUI is a standardized parent-report questionnaire designed to assess pragmatic language in children [173]. In this study, the LUI evaluation was conducted at the beginning (T0) and the end (T1) for children in both the robotic group and the control group.

In total, 9 children from the Robotic group and 7 children from the Control group reached the end of the study. The children who did not arrive at the end of the study were the ones who stopped going to CADIn, for personal reasons. The main characteristics of all children are shown in Table 7.1. In the end, the percentage of verbal children in the group became larger in the robotics group.

Table 7.1: Principal characteristics of the children who arrived to the end of the study.

	Nº Children	Nº Children that delivered the LUI	Age (months) median[1st quartile; 3rd quartile]	% Verbal
Robot group	9	7	50 [44; 61.5]	66%
Control group	7	5	49.5 [33.5; 63.5]	28%
Total	16	12	50 [41.3; 61.8]	50%

In the next paragraphs, we explain how we analysed this dataset to explore the potentialities of our attention classification system further. Moreover, we show how our automatic quantitative measures can be related to the chosen clinical scale.

7.2.1 Generalizability of the model

First, we decided to test the generalizability of the attention model (Section 5.1) in a subsample of seven children. A main difference in relation to the initial study where we tested our attention model in the Associação Portuguesa para o Autismo e as Perturbações do Desenvolvimento (APPDA) (Section 5.1.5), was that in this study the computer was replaced by a tablet to diminish the child's attention towards a distractor object. Thus, there were two targets: the robot and the other person. For each child, we analysed three sessions that by chronological order were called A, B and C.

In this study, the number of skeletons not detected by the Kinect was considerably high, due to the younger age of the participants and the fact that the acquisition room had more light. Therefore we used the pose reconstructor described in Section 6.1 and then processed the skeletons as the ones of APPDA acquired with the Kinect.

Then we first used the parameters (the AOI widths of the robot and of the other person) found in the APPDA sessions and calculated the model's accuracy for the three sessions. The ground truth was established by two annotators who labelled selected frames. 68% of the selected frames were kept for the evaluation. Second, we applied the learning approach to choose the best parameters. Sessions A, B, and C were considered training, validation and test sets, respectively. In Table 7.2, we verify that using the same parameters or the learning approach leads to similar results although better with this last technique. In addition, even though there is a decrease of 10% in the performance in relation to the APPDA performances, the accuracies are still above 74%, which is similar to other state-of-the-art methods, considering our scenario with younger children and more challenging positions.

Table 7.2: Accuracy of our attention classification system in the CADIn study using the widths of the APPDA study and the widths obtained from the learning approach. Bold indicates the highest accuracy for each session.

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

7.2.2 Comparison between clinical and automatic measures of attention

From the children who arrived at the end of the study, we calculated the Total Fixation Duration (TFD) for the several sessions in which they participated. The number of sessions was different for the different children. Then, we compared this variable with the LUI scale, maintaining all the children whose parents delivered a completed LUI at T0 and T1 (Table 7.1). The main goals of this analysis were to verify if the attention versus the therapist increased and if the attention versus Elsewhere decreased. In addition, we wanted to compare the longitudinal trend in the LUI scores with the proposed attention outcome measures.

We started by doing a non-parametric unpaired Mann-Whitney U Test to compare the TFD measures versus each target in the robot and control at T0 and then at T1. The same was done for the LUI. The objective of these tests is always to verify the null hypothesis that the distributions of both populations are identical (e.g: the LUI at T0 for the robot group has an identical distribution to the LUI at T0 for the control group).

Secondly, we did a paired test with the Wilcoxon signed-rank test, comparing the evolution of LUI at T0 and T1 and the same for the evolution of TFD. In this case, the null hypothesis is that for example the distribution of (LUI at T0, LUI at T1) is equal to the distribution of (LUI at T0, LUI at T1). In the end, through a Spearman correlation, using the measurements of both the control and robotic group, we compared the evolution of the LUI [LUI final (T1)- LUI initial (T0)] first, with the evolution of the TFD and then with the median value of the TFD across sessions.

Regarding the Total Fixation Duration (Table 7.3), we noticed that although the two groups were not different at T0, they became significantly different at T1 for all targets. Regarding the LUI (Table 7.4), there were no significant differences neither at T0 nor at T1 between the two groups.

Concerning the paired analysis, we had one significant result in the Robotic Group for the evolution of LUI (Figure 7.5). We verified that in the robotic group, the LUI significantly increases ($p = 0.047$), showing the potential of our therapy to augment communication capabilities. In addition, we noticed that the decrease of TFD towards elsewhere in the robotic group was very close to a significant value ($p = 0.054$). The decrease of the TFD towards elsewhere shows that the child is less distracted, demonstrating that the therapy can augment the engagement of the children in the tasks (Figure 7.6).

In the end, a moderate correlation ($\rho = -0.62$) between the median TFD towards Elsewhere and

Table 7.3: Median[1st quartile; 3rd quartile] of the Total Fixation Duration towards the three areas of interest (NAO, therapist and elsewhere) for the control and experimental group and respective p-values. Bold indicates the significant p-values.

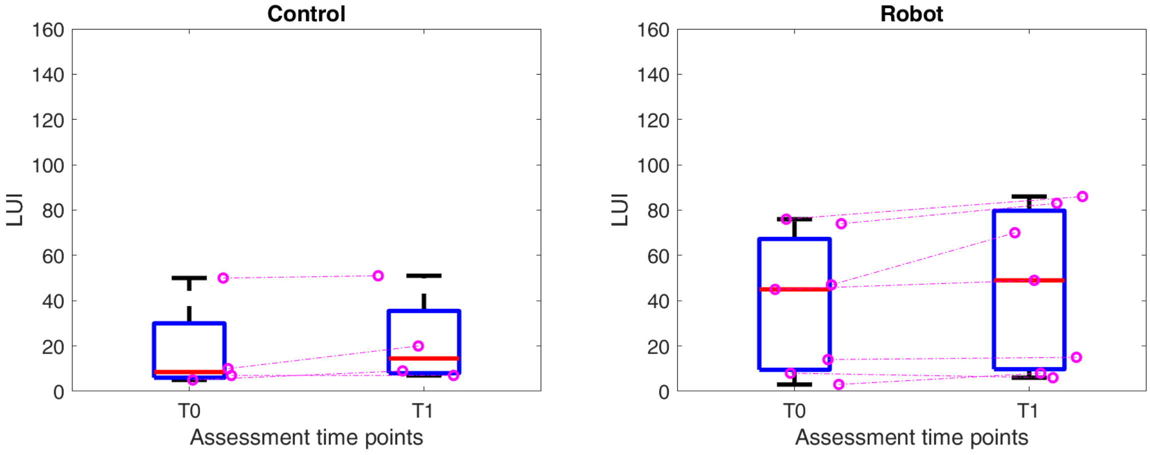
T0	Control	Robot	p-value
NAO	26.0 [22.9; 56.7]	52.5 [38.2; 62.9]	0.25
Therapist	35.2 [17.9; 48.4]	19.2 [14.4; 34.5]	0.41
Elsewhere	26.0 [19.4; 33.3]	18.9 [7.1; 38.1]	0.61

T1	Control	Robot	p-value
NAO	15.2 [11.1; 33.0]	55.6 [34.7; 74.4]	**0.01**
Therapist	44.0 [42.6; 52.5]	30.6 [20.6; 34.0]	**0.01**
Elsewhere	30.9 [24.1; 40.0]	11.2 [7.0; 21.7]	**0.05**

Table 7.4: Median[1st quartile; 3rd quartile] of the LUI scale for the control and experimental group and respective p-values.

T0	Control	Robot	p-value
LUI	8.5 [6; 30]	45.0 [9.5; 67.3]	0.53

T1	Control	Robot	p-value
LUI	14.5 [8; 35.5]	49.0 [9.8; 79.8]	0.41



(a)

(b)

Figure 7.5: Boxplot of LUI scale at T0 and T1 for (a) the control and (b) the robotic group.

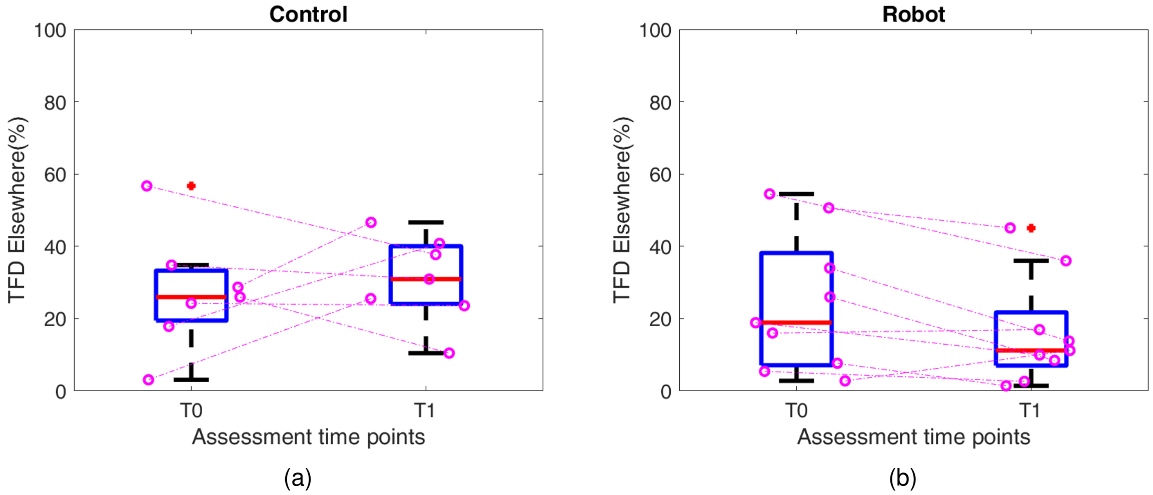


Figure 7.6: Boxplot of Total Fixation Duration (TFD) towards Elsewhere at T0 and T1 for (a) the control and (b) the robotic group.

the evolution of the LUI (LUI final-LUI initial) was found and it was significant ($p = 0.04$). This indicates that a decrease in the median TFD towards Elsewhere—reflecting reduced distraction in the child—is associated with greater improvements in the LUI scale and, consequently, in communication. In this way, there is a relation between our constructed metric and the expected evolution of a clinical scale.

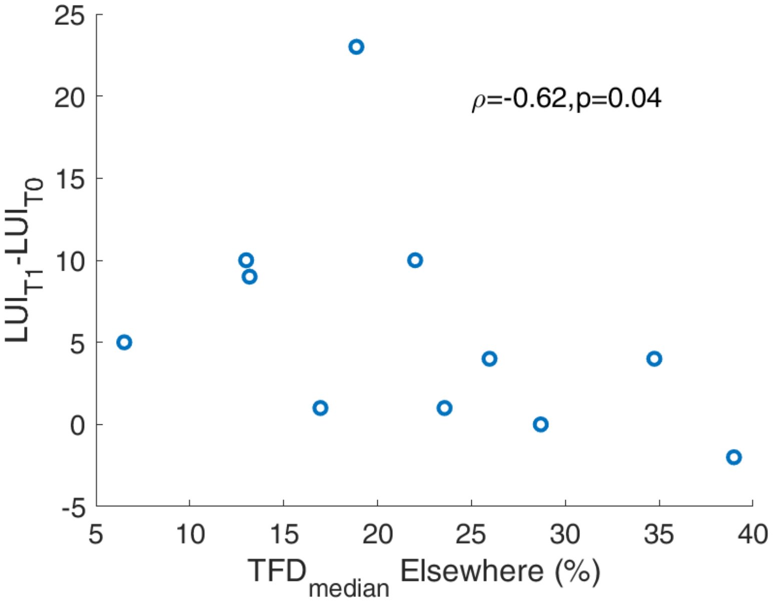


Figure 7.7: Evolution of the LUI considering the median TFD when looking at elsewhere for both the control and robotic groups.

7.3 Discussion

These two studies allowed us to test our robotic-assisted protocol with a considerable number of children for several weeks. In the first study, we could assess the impact on the results of the heterogeneity characteristic of children with ASD. We also noticed that maybe the metrics chosen were not sensitive enough for the capabilities that we were testing, since the overall sample patterns were not observed.

That was one of the main reasons we decided to have a future outcome measure based on the attention of the children, the Early Social Communication Scale (ESCS) that we have adapted in the scale protocol in Section 3.4.

Given that this new outcome measure was not ready at the time of acquisitions, for our second study we used the LUI scale, which principally focuses on communication. We verified a significant improvement in this scale in the robotic group from T0 to T1. This improvement was not present in the Control Group. This measure is based on a questionnaire done by the parents and they were not blinded to the intervention, thus, the result could be biased.

Nevertheless, when joining the fixation values of both the control group and robotic group, we found a relation between the evolution of the LUI and the median Total Fixation Duration towards Elsewhere, showing that our designed metric can be connected with a clinical metric related to communication. The increase in the child's engagement during the session is therefore connected with a better improvement in the communication scale.

Both studies represent the achievement of the principal goals of this thesis:

(i) We evaluated a protocol relevant to clinical practice across two distinct clinical settings.

(ii) We investigated an automated metric derived from a computer vision algorithm—total fixation duration—which enabled the assessment of children during sessions as well as before and after treatment.

(iii) We conducted a pilot study and a randomized controlled trial, confirming that participants who underwent the training phase showed improvements in the primary outcome measures.

Chapter 8

Conclusions

The main goal of this thesis was to develop protocols and quantitative measures that could be used in robotic-assisted therapies for children with Autism, aiming to introduce new technological tools into clinical settings to assist therapists and positively impact children with ASD.

The state-of-the-art highlights the benefits of robotic interventions, with several randomized controlled trials demonstrating positive outcomes compared to control interventions (Chapter 2). One of the main obstacles to the breadth of these therapies is the lack of quantitative measures and their relation with the medical scales already used by clinicians. Quantitative measures should be obtained. However, due to the characteristics of children with ASD, these sensors are mainly cameras. The challenges associated with the computer vision field (occlusions, tracking, etc) are exacerbated by the highly unconstrained nature of autism therapies. Moreover, the heterogeneity of this disorder prevents the creation of homogeneous datasets. These datasets are also generally small, as most robotic therapies involve short sessions over a limited timeframe. The lack of data makes implementing machine learning algorithms challenging.

To deal with the unconstrained scenario, most robotic therapies rely on a Wizard of Oz setup, where an external operator controls the robot according to the child's behavior. This is not sustainable in clinical practice, leading to a growing trend toward more automated and adaptive protocols. Our protocols were always conceptualized with a view towards greater adaptability to the children, aiming to enlarge the number of children engaged while working important capabilities for their development (Chapter 3). In this phase, synergistically collaborating with the clinical staff of the three different institutions was crucial. The development of these protocols integrated their clinical knowledge with our system's engineering capabilities. Notably, the transition from the basic mirroring protocol to the hierarchical protocol, associated to an increase of hierarchy and diversity of the exercises, was followed by a single level protocol both in the sensorial and the bingo protocol. This simplification granted a larger flexibility to the therapist to adapt to the characteristics of the children.

We focused on three main fields due to the characteristics of our protocols: action analysis, attention analysis and 3D pose estimation. Regarding action analysis (Chapter 4), the mirroring measure emerged as a potential metric for assessing protocol difficulty and evaluating a fundamental aspect of our protocols: mirroring. It evidenced particular differences between the ASD child and the other neurotypical children, suggesting its potential as a metric for tracking improvements in mirroring abilities of children with ASD.

The various metrics to assess different features of actions were all dependent on the determination of the beginning and end of the movements, established manually in that case. The Kinetic Parameter Method solved this issue and allowed a considerable improvement of our recognition system compared to the sliding window method traditionally used in literature [144]. Our online recognition system achieved similar accuracies to the ones reported in literature, for adults. During child's therapy sessions it served as an engagement tool for children, triggering feedback signals.

In terms of attention analysis (Chapter 5), our attention classification system had an accuracy similar to the state of the art [112] in a more unconstrained environment. Additionally, this system appeared to align with the therapist's observations and was easy for her to interpret. This suggests that the tool could be expanded beyond children with ASD to support continuous evaluations after therapy for children with neurodevelopmental disorders more broadly.

The transition to an online system for integration in the robot controller led to a substantial reduction of the computational times. However, the performance results were unsatisfactory. The creation of a new evaluation protocol based on the ESCS scale and the association with a neural network provided considerably good results in an adult sample. Further tests with children are required to confirm these findings.

A significant limitation of our setup, which notably affected our attention classification system, was the high number of skeletons lost during data acquisition. Our 3D personalised pose estimator (Chapter 6) recovered a considerable number of skeletons lost by our depth camera during the therapies and outperforming state of the art algorithms [3] when compared to the gold standard system. Its linear regression model facilitated the development of an online model reliable and consistent, suitable for clinical environments.

Across multiple studies, we demonstrated how our system functioned effectively with diverse children with ASD, particularly in the studies with the largest sample sizes (Chapter 7): the pilot study developed in FDG and the pilot randomized controlled trial that took place in CADIn.The first study revealed how the heterogeneity of children influenced their progress within the intervention protocol. Children who reached the training levels showed improvements on all evaluated metrics, whereas those who remained at the familiarization levels did not. The second study was fundamental not just to prove that our attention system could generalize to younger children than initially tested but, principally, in establishing a relationship between our constructed attention evaluation metric and a clinical scale (LUI) documented in the literature [173]. This study revealed that larger reduction of the median fixation time towards ewas associated to a larger evolution of the LUI from the beginning to the end of the study. This finding underscores the importance of the designed metric for the quantitative evaluation of the children's development. In addition, we noticed that in the robotic group, there was a significant improvement in the LUI metric that was not present in the control group, which suggest the potential of our robotic therapy in modifying ASD symptoms.

During the development of these technologies and the associated metrics, we not only confirmed findings from other robotic therapies but also observed several aspects reported in the fundamental autism literature. Attention maps clearly highlighted a greater focus on non-social stimuli compared to social stimuli. Similarly, mirroring measures revealed that children with ASD experienced greater difficulties in this area compared to neurotypical children. In the end, in the pilot studies developed, the heterogeneity present in ASD was revealed in all our results.

8.1 Future Work

The first short-term direction of the work is increasing the robustness of the results through further testing. In fact, at the end of this thesis, a randomized controlled trial of the sensorial protocol and a pilot study with the bingo protocol are underway. This randomized controlled trial also includes testing the scale protocol with the simultaneous collection of the ESCS scale. A multi-centric analysis can now be conducted combining data from multiple centres across two countries. This will help investigate, for example, how children's attention evolved across different protocols and over several weeks.

Looking toward the long-term scenario, we propose three key future directions that we believe are the most important:

- **Development of a personalized robotic system**

  This thesis has demonstrated the importance of having a flexible robotic system since all our protocols have evolved in that direction based on test acquisitions in clinical scenarios. Additionally, the quantitative measures developed could be leveraged for this purpose. Personalization could follow two distinct approaches:

  (i) a system that suggests to the therapist which exercises the child could do based on their performance and attention metrics recorded in the previous sessions.

  (ii) a system that dynamically adapts the type of feedback and exercises based on the child's attention and performance during sessions.

  For both systems it is important to establish the order in which the exercises and feedback should be provided. Even though at this moment we have a considerable dataset of acquisitions and could use reinforcement learning algorithms to determine this order, substantial work with the therapists is essential. Their extensive experience and expert knowledge—far beyond the number of recorded sessions—are invaluable for refining this process.

  Additionally, for the second type of system, all our metrics must function in real-time. First, the kinetic parameter method should be associated with other kinematic measures, as the velocity or acceleration, to determine whether only the arms were moving or other body parts were involved. This would improve the segmentation of the gesture and, consequently, the recognition method. This method could also be used to transform our mirroring measure into an online metric. Regarding the attention system, exploring alternative face detection methods could enhance performance metrics.

- **Generalization and modularization of the analysis system**

  This thesis has introduced different quantitative measures and their analysis was based on the acquired data and deep learning methods. Nowadays deep learning methods are constantly evolving and new cameras, pose estimators, gaze detectors or recognition methods emerge each year. There should be an effort for the standardization and integration of all the modules so that each module can be easily replaced without disrupting the overall analysis flow.

  Specifically, for recognition methods, since the our set of gestures evolve, an important future direction will include transfer learning. In this way, we expect that the knowledge of our trained networks can pass to the new ones. Given the limited availability of children with ASD, most testing is conducted on larger groups of adult subjects. Therefore, it is important to explore how networks trained on adults can be fine-tuned for a small group of available children. A potential approach is using weakly supervised data from the initial therapy sessions. During these sessions, the therapist would establish if the gesture was recognized correctly or not by the algorithm and this feedback/label could be used to adjust the neural network's weights.

- **Increasing parents' involvement**

  Throughout this PhD project, therapists were extensively involved in the design of several protocols, which was important for their implementation and for facilitating testing with children with this specific disorder. However, we observed that most dropouts occurred due to familiar organizational challenges. To address this, we aim to increase parental involvement by providing them with more information at the outset and offering opportunities to interact with the system during initial sessions. This would raise awareness about the system's potential benefits.

Moreover, a next step would be extending this protocol to a home setting. In the future, humanoid robots are expected to become more affordable, meanwhile, a solution could involve using Choregraphe, the NAO robot simulator. Although this approach would lack the benefits of physical embodiment, it would allow for increased rehabilitation hours and greater parental participation in therapy. Furthermore, this could help assess the generalizability of the learned behaviour to a scenario different from the clinical room.

8.2 Final considerations

This thesis has shown the various challenges and potential benefits of robotics therapies for autism. Attention and movement/gestures were the main targets of our quantitative metrics. Their strong performance in terms of accuracy suggest that they could be used not just in the autism field but also in other neurodevelopmental rehabilitation scenarios. In fact, one of our attention metrics was related to the evolution of a clinical scale, demonstrating its potential usefulness in the clinical practice.

The developed protocols were tested with more than 30 children over several weeks. The long-term impact of these therapies on these children can only be assessed in several years. In the meantime, in our memory, there remain many individual-subject, single episodes: the moment a child kissed the robot; the first time a nonverbal child spontaneously pointed, showing the robot to therapist; the mirroring dances between therapist and child mediated by the robot; the various ”hi” and ”goodbye” gestures presented to the robot, at the beginning and end of the sessions.

While these results require more observations to draw statistically relevant analyses, they were found to be very meaningful and relevant for the therapists. In a disorder characterized by heterogeneity and the presence of repetitive and restricted behaviors, even a single change in a child is invaluable. We thrust that these developments demonstrate the role of technology to enhance the therapies of subjects with ASD and pave the way for future endeavours.

Bibliography

[1] A. S. Ivani, A. Giubergia, L. Santos, A. Geminiani, S. Annunziata, A. Caglio, I. Olivieri, and A. Pedrocchi, “A gesture recognition algorithm in a robot therapy for ASD children,” *Biomedical Signal Processing and Control*, vol. 74, p. 103512, 2022. doi: 10.1016/j.bspc.2022.103512

[2] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, “Coherent reconstruction of multiple humans from a single image,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5578–5587, 2020. doi: 10.1109/CVPR42600.2020.00562

[3] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, “Putting people in their place: Monocular regression of 3d people in depth,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 243–13 252, 2022. doi: 10.1109/CVPR52688.2022.01289

[4] 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. doi: 10.1109/ICCV.2019.00701

[5] L. Santos, A. Geminiani, I. Olivieri, J. Santos-Victor, and A. Pedrocchi, “Copyrobot: Interactive mirroring robotics game for asd children,” in *XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019*, J. Henriques, N. Neves, and P. de Carvalho, Eds., pp. 2014–2027. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-030-31635-8_239

[6] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding*, vol. 170, pp. 51–66, 2018. doi: 10.1016/j.cviu.2018.03.003

[7] A. A. Morgan, J. Abdi, M. A. Q. Syed, G. E. Kohen, P. Barlow, and M. P. Vizcaychipi, “Robots in healthcare: a scoping review,” *Current Robotics Reports*, vol. 3, no. 4, pp. 271–280, 2022. doi: 10.1007/s43154-022-00095-4

[8] 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. doi: 10.1146/annurev-bioeng-071811-150036

[9] R. Sacco, N. Camilleri, J. Eberhardt, K. Umla-Runge, and D. Newbury-Birch, “The prevalence of autism spectrum disorder in europe,” in *Autism Spectrum Disorders - Recent Advances and New Perspectives*, A. P. M. Carotenuto, Ed. Rijeka: IntechOpen, 2022, ch. 13. doi: 10.5772/intechopen.108123

[10] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-V*. American Psychiatric Association, 2013. doi: 10.1176/appi.books

[11] J. Baio, L. Wiggins, D. L. Christensen, M. J. Maenner, J. Daniels, Z. Warren, M. Kurzius-Spencer, W. Zahorodny, C. Robinson Rosenberg, T. White, M. S. Durkin, P. Imm, L. Nikolaou, M. Yeargin-Allsopp, L.-C. Lee, R. Harrington, M. Lopez, R. T. Fitzgerald, A. Hewitt, S. Pettygrove, J. N. Constantino, A. Vehorn, J. Shenouda, J. Hall-Lande, K. Van Naarden Braun, and N. F. Dowling, "Prevalence of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 sites, united states, 2014," *Morbidity and mortality weekly report. Surveillance summaries*, vol. 67, no. 6, pp. 1–23, Apr 2018. doi: 10.15585/mmwr.ss6706a1

[12] T. Hirota and B. H. King, "Autism spectrum disorder: A review," *JAMA*, vol. 329, no. 2, pp. 157–168, 2023. doi: 10.1001/jama.2022.23661

[13] H. Stieglitz Ham, A. Bartolo, M. Corley, G. Rajendran, A. Szabo, and S. Swanson, "Exploring the relationship between gestural recognition and imitation: Evidence of dyspraxia in autism spectrum disorders," *Journal of Autism and Developmental Disorders*, vol. 41, no. 1, p. 1–12, 2011. doi: 10.1007/s10803-010-1011-1

[14] P. A. G. Forbes, X. Pan, and A. F. de C. Hamilton, "Reduced mimicry to virtual reality avatars in autism spectrum disorder," *Journal of Autism and Developmental Disorders*, vol. 46, no. 12, pp. 3788–3797, 2016. doi: 10.1007/s10803-016-2930-2

[15] H. Dadgar, J. Alaghband Rad, Z. Soleymani, A. Khorammi, J. McCleery, and S. Maroufizadeh, "The relationship between motor, imitation, and early social communication skills in children with autism," *Iranian Journal of Psychiatry*, vol. 12, no. 4, pp. 236–240, Oct 2017.

[16] M. Jouaiti and P. Hénaff, "Robot-based motor rehabilitation in autism: A systematic review," *International Journal of Social Robotics*, vol. 11, no. 5, pp. 753–764, 2019. doi: 10.1007/s12369-019-00598-9

[17] 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. doi: 10.1186/s13229-018-0230-8

[18] W.-C. So, C.-H. Cheng, W.-Y. Lam, Y. Huang, K.-C. Ng, H.-C. Tung, and W. Wong, "A robot-based play-drama intervention may improve the joint attention and functional play behaviors of chinese-speaking preschoolers with autism spectrum disorder: A pilot study," *Journal of Autism and Developmental Disorders*, vol. 50, no. 2, pp. 467–481, 2020. doi: 10.1007/s10803-019-04270-z

[19] 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. doi: 10.1109/ACCESS.2020.3045042

[20] E. C. Zabor, A. M. Kaizer, and B. P. Hobbs, "Randomized controlled trials," *Chest*, vol. 158, no. 1S, pp. S79–S87, 2020. doi: 10.1016/j.chest.2020.03.013

[21] W.-C. So, C.-H. Cheng, W.-Y. Lam, T. Wong, W.-W. Law, Y. Huang, K.-C. Ng, H.-C. Tung, and W. Wong, "Robot-based play-drama intervention may improve the narrative abilities of chinese-speaking preschoolers with autism spectrum disorder," *Research in Developmental Disabilities*, vol. 95, p. 103515, 2019. doi: 10.1016/j.ridd.2019.103515

[22] Z. Zheng, G. Nie, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, “A randomized controlled trial of an intelligent robotic response to joint attention intervention system,” *Journal of Autism and Developmental Disorders*, vol. 50, no. 8, pp. 2819–2831, 2020. doi: 10.1007/s10803-020-04388-5

[23] L. Santos, S. Annunziata, A. Geminiani, A. Ivani, A. Giubergia, A. Caglio, E. Brazzoli, R. Lipari, M. C. Carrozza, I. Olivieri, and A. Pedrocchi, “Applications of robotics for Autism Spectrum Disorder: a scoping review,” *Review Journal of Autism and Developmental Disorders*, 2023. doi: 10.1007/s40489-023-00402-5

[24] C. Lord, T. S. Brugha, T. Charman, J. Cusack, G. Dumas, T. Frazier, E. J. H. Jones, R. M. Jones, A. Pickles, M. W. State, J. L. Taylor, and J. Veenstra-VanderWeele, “Autism spectrum disorder,” *Nature Reviews Disease Primers*, vol. 6, no. 1, p. 5, 2020. doi: 10.1038/s41572-019-0138-4

[25] I. Kamp-Becker, J. Tauscher, N. Wolff, C. Küpper, L. Poustka, S. Roepke, V. Roessner, D. Heider, and S. Stroth, “Is the combination of ADOS and ADI-R necessary to classify ASD? rethinking the “gold standard” in diagnosing ASD,” *Frontiers in Psychiatry*, vol. 12, p. 727308, 2021. doi: 10.3389/fpsyt.2021.727308

[26] K. Strimbu and J. A. Tavel, “What are biomarkers?” *Current Opinion on HIV and AIDS*, vol. 5, no. 6, pp. 463–466, 2010. doi: 10.1097/COH.0b013e32833ed177

[27] H. R. Park, J. M. Lee, H. E. Moon, D. S. Lee, B.-N. Kim, J. Kim, D. G. Kim, and S. H. Paek, “A short review on the current understanding of autism spectrum disorders,” *Experimental Neurobiology*, vol. 25, no. 1, pp. 1–13, 2016. doi: 10.5607/en.2016.25.1.1

[28] C. Ames and S. Fletcher-Watson, “A review of methods in the study of attention in autism,” *Developmental Review*, vol. 30, no. 1, pp. 52–73, 2010. doi: 10.1016/j.dr.2009.12.003

[29] M. Chita-Tegmark, “Attention allocation in asd: a review and meta-analysis of eye-tracking studies,” *Review Journal of Autism and Developmental Disorders*, vol. 3, no. 3, pp. 209–223, 2016. doi: 10.1007/s40489-016-0077-x

[30] X. Ma, H. Gu, and J. Zhao, “Atypical gaze patterns to facial feature areas in autism spectrum disorders reveal age and culture effects: A meta-analysis of eye-tracking studies,” *Autism Research*, vol. 14, no. 12, pp. 2625–2639, 2021. doi: 10.1002/aur.2607

[31] P. Mundy, “A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder,” *European Journal of Neuroscience*, vol. 47, no. 6, pp. 497–514, 2018. doi: 10.1111/ejn.13720

[32] P. Mundy, C. Delgado, J. Block, M. Venezia, A. Hogan, and J. Seibert, “A manual for the early social communication scales (escs),” *Davis: MIND Institute, University of California at Davis*, 2003.

[33] M. G. Logrieco, E. Annechini, L. Casula, S. Guerrera, M. Fasolo, S. Vicari, and G. Valeri, “Nonverbal skills evolution in children with autism spectrum disorder one year post-diagnosis,” *Children*, vol. 11, no. 12, 2024. doi: 10.3390/children11121520

[34] Q. Ye, L. Liu, S. Lv, S. Cheng, H. Zhu, Y. Xu, X. Zou, and H. Deng, “The gestures in 2–4-year-old children with autism spectrum disorder,” *Frontiers in Psychology*, vol. 12, 2021. doi: 10.3389/fpsyg.2021.604542

[35] L. A. Edwards, “A meta-analysis of imitation abilities in individuals with autism spectrum disorders,” *Autism Research*, vol. 7, no. 3, pp. 363–380, 2014. doi: 10.1002/aur.1379

[36] Y. Huang, M. K.-Y. Wong, W.-Y. Lam, C.-H. Cheng, and W.-C. So, “What affects gestural learning in children with and without autism? the role of prior knowledge and imitation,” *Research in Developmental Disabilities*, vol. 129, p. 104305, 2022. doi: 10.1016/j.ridd.2022.104305

[37] J. A. Colebourn, A. C. Golub-Victor, and A. Paez, “Developing overhand throwing skills for a child with autism with a collaborative approach in school-based therapy,” *Pediatric Physical Therapy*, vol. 29, no. 3, 2017. doi: 10.1097/PEP.0000000000000405

[38] K. Boshoff, H. Bowen, H. Paton, S. Cameron-Smith, S. Graetz, A. Young, and K. Lane, “Child development outcomes of dir/floortime tm-based programs: A systematic review,” *Canadian Journal of Occupational Therapy*, vol. 87, no. 2, pp. 153–164, 2020. doi: 10.1177/0008417419899224

[39] 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. doi: 10.1007/s10803-019-03953-x

[40] S. Ali, F. Mehmood, D. Dancey, Y. Ayaz, M. J. Khan, N. Naseer, R. D. C. Amadeu, H. Sadia, and R. Nawaz, “An adaptive multi-robot therapy for improving joint attention and imitation of asd children,” *IEEE Access*, vol. 7, pp. 81 808–81 825, 2019. doi: 10.1109/ACCESS.2019.2923678

[41] 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. doi: 10.3389/fpsyg.2019.01503

[42] F. Mehmood, Y. Ayaz, S. Ali, R. De Cassia Amadeu, and H. Sadia, “Dominance in visual space of asd children using multi-robot joint attention integrated distributed imitation system,” *IEEE Access*, vol. 7, pp. 168 815–168 827, 2019. doi: 10.1109/ACCESS.2019.2951366

[43] Y. Feng, Q. Jia, M. Chu, and W. Wei, “Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation,” *IEEE Access*, vol. 5, pp. 19 494–19 504, 2017. doi: 10.1109/ACCESS.2017.2754291

[44] D. O. David, C. A. Costescu, S. Matu, A. Szentagotai, and A. Dobrean, “Effects of a robot-enhanced intervention for children with asd on teaching turn-taking skills,” *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 29–62, 2020. doi: 10.1177/0735633119830344

[45] 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. doi: 10.3389/fpsyt.2019.00370

[46] L. Boccanfuso, S. Scarborough, R. K. Abramson, A. V. Hall, H. H. Wright, and J. M. O’Kane, “A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned,” *Autonomous Robots*, vol. 41, no. 3, pp. 637–655, 2017. doi: 10.1007/s10514-016-9554-4

[47] 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. doi: 10.1109/TNSRE.2015.2475724

[48] R. S. Moorthy and S. Pugazhenthi, “Imitation based training to enhance psychomotor skills in autistic children using a snatcher robot,” in *2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA)*, pp. 1–6, 2016. doi: 10.1109/RAHA.2016.7931903

[49] Z. Telisheva, A. Turarova, A. Zhanatkyzy, G. Abylkasymova, and A. Sandygulova, “Robot-assisted therapy for the severe form of autism: Challenges and recommendations,” in *Social Robotics*, M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, Á. Castro-González, and H. He, Eds., pp. 474–483. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-35888-4_44

[50] P. Chevalier, J.-C. Martin, B. Isableu, C. Bazile, and A. Tapus, “Impact of sensory preferences of individuals with autism on the recognition of emotions expressed by two robots, an avatar, and a human,” *Autonomous Robots*, vol. 41, no. 3, pp. 613–635, 2017. doi: 10.1007/s10514-016-9575-z

[51] F. Petric, D. Miklić, M. Cepanec, P. Cvitanović, and Z. Kovačić, “Functional imitation task in the context of robot-assisted autism spectrum disorder diagnostics: Preliminary investigations,” in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 1471–1478, 2017. doi: 10.1109/ROMAN.2017.8172498

[52] F. Petric and Z. Kovačić, “Hierarchical pomdp framework for a robot-assisted asd diagnostic protocol,” in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 286–293, 2019. doi: 10.1109/HRI.2019.8673295

[53] M. Del Coco, M. Leo, P. Carcagnì, F. Famà, L. Spadaro, L. Ruta, G. Pioggia, and C. Distante, “Study of mechanisms of social interaction stimulation in autism spectrum disorder by assisted humanoid robot,” *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 993–1004, 2018. doi: 10.1109/TCDS.2017.2783684

[54] A. A. Ramírez-Duque, T. Bastos, M. Munera, C. A. Cifuentes, and A. Frizera-Neto, “Robot-assisted intervention for children with special needs: A comparative assessment for autism screening,” *Robotics and Autonomous Systems*, vol. 127, p. 103484, 2020. doi: 10.1016/j.robot.2020.103484

[55] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, Y. Yoshimura, T. Ikeda, C. Hasegawa, D. N. Saito, J. Shimaya, H. Ishiguro, M. Mimura, and M. Kikuchi, “Brief report: A novel system to evaluate autism spectrum disorders using two humanoid robots,” *Journal of Autism and Developmental Disorders*, vol. 49, no. 4, pp. 1709–1716, 2019. doi: 10.1007/s10803-018-3848-7

[56] F. Alnajjar, M. Cappuccio, A. 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, no. 1, pp. 67–82, 2021. doi: 10.1007/s12369-020-00639-8

[57] K. Baraka, F. S. Melo, M. Couto, and M. Veloso, “Optimal action sequence generation for assistive agents in fixed horizon tasks,” *Autonomous Agents and Multi-Agent Systems*, vol. 34, no. 2, p. 33, 2020. doi: 10.1007/s10458-020-09458-7

[58] A. Amanatiadis, V. G. Kaburlasos, C. Dardani, S. A. Chatzichristofis, and A. Mitropoulos, “Social robots in special education: Creating dynamic interactions for optimal experience,” *IEEE Consumer Electronics Magazine*, vol. 9, no. 3, pp. 39–45, 2020. doi: 10.1109/MCE.2019.2956218

[59] F. Petric and Z. Kovacic, “Design and validation of momdp models for child–robot interaction within tasks of robot-assisted asd diagnostic protocol,” *International Journal of Social Robotics*, vol. 12, no. 2, pp. 371–388, 2020. doi: 10.1007/s12369-019-00577-0

[60] U. Qidwai, S. B. A. Kashem, and O. Conor, “Humanoid robot as a teacher’s assistant: Helping children with autism to learn social and academic skills,” *Journal of Intelligent & Robotic Systems*, vol. 98, no. 3, pp. 759–770, Jun 2020. doi: 10.1007/s10846-019-01075-1

[61] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban, H. Liu, V. Nair, S. Matu, A. Mazel, M. Selescu, E. Senft, S. Thill, B. Vanderborght, D. Vernon, and T. Ziemke, “The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy,” *PLOS ONE*, vol. 15, no. 8, pp. 1–15, 2020. doi: 10.1371/journal.pone.0236939

[62] E. Y.-h. Chung, “Robot-mediated social skill intervention programme for children with autism spectrum disorder: An aba time-series study,” *International Journal of Social Robotics*, 2020. doi: 10.1007/s12369-020-00699-w

[63] M. W. D. Korte, I. van den Berk-Smeekens, M. van Dongen-Boomsma, I. J. Oosterling, J. C. D. Boer, E. I. Barakova, T. Lourens, J. K. Buitelaar, J. C. Glennon, and W. G. Staal, “Self-initiations in young children with autism during pivotal response treatment with and without robot assistance,” *Autism*, vol. 24, no. 8, pp. 2117–2128, 2020. doi: 10.1177/1362361320935006

[64] J. A. Barnes, C. H. Park, A. Howard, and M. Jeon, “Child-robot interaction in a musical dance game: An exploratory comparison study between typically developing children and children with autism,” *International Journal of Human–Computer Interaction*, vol. 37, no. 3, pp. 249–266, 2021. doi: 10.1080/10447318.2020.1819667

[65] B. R. Schadenberg, D. Reidsma, D. K. J. Heylen, and V. Evers, “Differences in spontaneous interactions of autistic children in an interaction with an adult and humanoid robot,” *Frontiers in Robotics and AI*, vol. 7, p. 28, 2020. doi: 10.3389/frobt.2020.00028

[66] M. Salvador, A. S. Marsh, A. Gutierrez, and M. H. Mahoor, “Development of an aba autism intervention delivered by a humanoid robot,” in *Social Robotics*, A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, Eds., pp. 551–560. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-47437-3_54

[67] I. B. Wijayasinghe, I. Ranatunga, N. Balakrishnan, N. Bugnariu, and D. O. Popa, “Human–robot gesture analysis for objective assessment of autism spectrum disorder,” *International Journal of Social Robotics*, vol. 8, no. 5, pp. 695–707, 2016. doi: 10.1007/s12369-016-0379-2

[68] E. Marinoiu, M. Zanfir, V. Olaru, and C. Sminchisescu, “3d human sensing, action and emotion recognition in robot assisted therapy of children with autism,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2158–2167, 2018. doi: 10.1109/CVPR.2018.00230

[69] G. Palestra, G. Varni, M. Chetouani, and F. Esposito, “A multimodal and multilevel system for robotics treatment of autism in children,” in *DAA ’16: Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, no. 3, pp. 1–6. Association for Computing Machinery, 2016. doi: 10.1145/3005338.3005341

[70] H. Kumazaki, T. Muramatsu, Y. Yoshikawa, B. A. Corbett, Y. Matsumoto, H. Higashida, T. Yuhi, H. Ishiguro, M. Mimura, and M. Kikuchi, “Job interview training targeting nonverbal communication using an android robot for individuals with autism spectrum disorder,” *Autism*, vol. 23, no. 6, pp. 1586–1595, 2019. doi: 10.1177/1362361319827134

[71] H. Kumazaki, Z. Warren, A. Swanson, Y. Yoshikawa, Y. Matsumoto, Y. Yoshimura, J. Shimaya, H. Ishiguro, N. Sarkar, J. Wade, M. Mimura, Y. Minabe, and M. Kikuchi, “Brief report: Evaluating the utility of varied technological agents to elicit social attention from children with autism spectrum disorders,” *Journal of Autism and Developmental Disorders*, vol. 49, no. 4, pp. 1700–1708, 2019. doi: 10.1007/s10803-018-3841-1

[72] H. Kumazaki, Z. Warren, A. Swanson, Y. Yoshikawa, Y. Matsumoto, H. Takahashi, N. Sarkar, H. Ishiguro, M. Mimura, Y. Minabe, and M. Kikuchi, “Can robotic systems promote self-disclosure in adolescents with autism spectrum disorder? a pilot study,” *Frontiers in Psychiatry*, vol. 9, p. 36, 2018. doi: 10.3389/fpsyt.2018.00036

[73] H. Kumazaki, Z. Warren, T. Muramatsu, Y. Yoshikawa, Y. Matsumoto, M. Miyao, M. Nakano, S. Mizushima, Y. Wakita, H. Ishiguro, M. Mimura, Y. Minabe, and M. Kikuchi, “A pilot study for robot appearance preferences among high-functioning individuals with autism spectrum disorder: Implications for therapeutic use,” *PLOS ONE*, vol. 12, no. 10, pp. 1–13, 2017. doi: 10.1371/journal.pone.0186581

[74] A. Zaraki, M. Khamassi, L. J. Wood, G. Lakatos, C. Tzafestas, F. Amirabdollahian, B. Robins, and K. Dautenhahn, “A novel reinforcement-based paradigm for children to teach the humanoid kaspar robot,” *International Journal of Social Robotics*, vol. 12, no. 3, pp. 709–720, 2020. doi: 10.1007/s12369-019-00607-x

[75] B. Robins, K. Dautenhahn, L. Wood, and A. Zaraki, “Developing interaction scenarios with a humanoid robot to encourage visual perspective taking skills in children with autism – preliminary proof of concept tests,” in *Social Robotics*, A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He, Eds., pp. 147–155. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-70022-9_15

[76] A. Zaraki, L. Wood, B. Robins, and K. Dautenhahn, “Development of a semi-autonomous robotic system to assist children with autism in developing visual perspective taking skills,” in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 969–976, 2018. doi: 10.1109/ROMAN.2018.8525681

[77] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, and R. W. Picard, “Measuring engagement in robot-assisted autism therapy: A cross-cultural study,” *Frontiers in Robotics and AI*, vol. 4, p. 36, 2017. doi: 10.3389/frobt.2017.00036

[78] 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. doi: 10.1016/j.patrec.2018.03.007

[79] S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, and M. J. Matarić, “Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders,” *Science Robotics*, vol. 5, no. 39, 2020. doi: 10.1126/scirobotics.aaz3791

[80] S.-S. Yun, H. Kim, J. Choi, and S.-K. Park, “A robot-assisted behavioral intervention system for children with autism spectrum disorders,” *Robotics and Autonomous Systems*, vol. 76, pp. 58 – 67, 2016. doi: 10.1016/j.robot.2015.11.004

[81] K. Dautenhahn, “Human-centred social robotics: Autonomy, trust and interaction challenges,” 2020, international Conference on

Robotics and Automation. [Online]. Available: https://ieeetv.ieee.org/icra-2020-keynote-human-centred-social-robotics-autonomy-trust-and-interaction-challenges

[82] 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. doi: 10.1109/ROMAN.2018.8525747

[83] J. Bharatharaj, L. Huang, A. Al-Jumaily, R. E. Mohan, and C. Krägeloh, “Sociopsychological and physiological effects of a robot-assisted therapy for children with autism,” *International Journal of Advanced Robotic Systems*, vol. 14, no. 5, p. 1729881417736895, 2017. doi: 10.1177/1729881417736895

[84] L. Desideri, M. Negrini, M. C. Cutrone, A. Rouame, M. Malavasi, E.-J. Hoogerwerf, P. Bonifacci, and R. Sarro, “Exploring the use of a humanoid robot to engage children with autism spectrum disorder (asd),” *Studies in health technology and informatics*, vol. 242, pp. 501–509, 2017. doi: 10.3233/978-1-61499-798-6-501

[85] J. Bharatharaj, L. Huang, A. Al-Jumaily, M. R. Elara, and C. Krägeloh, “Investigating the effects of robot-assisted therapy among children with autism spectrum disorder using bio-markers,” *IOP Conference Series: Materials Science and Engineering*, vol. 234, p. 012017, sep 2017. doi: 10.1088/1757-899X/234/1/012017

[86] L. Desideri, M. Negrini, M. Malavasi, D. Tanzini, A. Rouame, M. C. Cutrone, P. Bonifacci, and E.-J. Hoogerwerf, “Using a humanoid robot as a complement to interventions for children with autism spectrum disorder: a pilot study,” *Advances in Neurodevelopmental Disorders*, vol. 2, no. 3, pp. 273–285, 2018. doi: 10.1007/s41252-018-0066-4

[87] H. 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. doi: 10.1109/MRA.2019.2904121

[88] F. S. Melo, A. Sardinha, D. Belo, M. Couto, M. Faria, A. Farias, H. Gambôa, C. Jesus, M. Kinarullathil, P. Lima, L. Luz, A. Mateus, I. Melo, P. Moreno, D. Osório, A. Paiva, J. Pimentel, J. Rodrigues, P. Sequeira, R. Solera-Ureña, M. Vasco, M. Veloso, and R. Ventura, “Project inside: towards autonomous semi-unstructured human–robot social interaction in autism therapy,” *Artificial Intelligence in Medicine*, vol. 96, pp. 198 – 216, 2019. doi: 10.1016/j.artmed.2018.12.003

[89] H. Cai, Y. Fang, Z. Ju, C. Costescu, D. David, E. Billing, T. Ziemke, S. Thill, T. Belpaeme, B. Vanderborght, D. Vernon, K. Richardson, and H. Liu, “Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study,” *IEEE Sensors Journal*, vol. 19, no. 4, pp. 1508–1518, 2019. doi: 10.1109/JSEN.2018.2877662

[90] I. Giannopulu, K. Terada, and T. Watanabe, “Communication using robots: a perception-action scenario in moderate asd,” *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, no. 5, pp. 603–613, 2018. doi: 10.1080/0952813X.2018.1430865

[91] K. Silva, M. Lima, A. Santos-Magalhães, C. Fafiães, and L. de Sousa, “Living and robotic dogs as elicitors of social communication behavior and regulated emotional responding in individuals with

autism and severe language delay: A preliminary comparative study," *Anthrozoös*, vol. 32, no. 1, pp. 23–33, 2019. doi: 10.1080/08927936.2019.1550278

[92] P. Ponce, A. Molina, D. Grammatikou, and O. Mata, "Fuzzy logic type 1 and 2 for social robots and apps for children with autism," in *2017 Sixteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pp. 1–8, 2017. doi: 10.1109/MICAI-2017.2017.00009

[93] S. Matsuda, E. Nunez, M. Hirokawa, J. Yamamoto, and K. Suzuki, "Facilitating social play for children with pdds: Effects of paired robotic devices," *Frontiers in Psychology*, vol. 8, 2017. doi: 10.3389/fpsyg.2017.01029

[94] E. Y.-h. Chung, "Robotic intervention program for enhancement of social engagement among children with autism spectrum disorder," *Journal of Developmental and Physical Disabilities*, vol. 31, no. 4, pp. 419–434, 2019. doi: 10.1007/s10882-018-9651-8

[95] A. Taheri, A. Meghdari, M. Alemi, and H. Pouretemad, "Human–robot interaction in autism treatment: A case study on three pairs of autistic children as twins, siblings, and classmates," *International Journal of Social Robotics*, vol. 10, no. 1, pp. 93–113, 2018. doi: 10.1007/s12369-017-0433-8

[96] B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, and F. Shic, "Improving social skills in children with asd using a long-term, in-home social robot," *Science Robotics*, vol. 3, no. 21, 2018. doi: 10.1126/scirobotics.aat7544

[97] A. D. Nuovo, D. Conti, G. Trubia, S. Buono, and S. D. Nuovo, "Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability," *Robotics*, vol. 7, no. 2, p. 25, 2018. doi: 10.3390/robotics7020025

[98] J. Bharatharaj, L. Huang, R. Mohan, A. Al-Jumaily, and C. Krägeloh, "Robot-assisted therapy for learning and social interaction of children with autism spectrum disorder," *Robotics*, vol. 6, no. 1, p. 4, 2017. doi: 10.3390/robotics6010004

[99] L. Boccanfuso, E. Barney, C. Foster, Y. A. Ahn, K. Chawarska, B. Scassellati, and F. Shic, "Emotional robot to examine different play patterns and affective responses of children with and without asd," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 19–26, 2016. doi: 10.1109/HRI.2016.7451729

[100] S. Attawibulkul, N. Asawalertsak, P. Suwawong, P. Wattanapongsakul, W. Jutharee, and B. Kaewkamnerdpong, "Using a daily routine game on the bliss robot for supporting personal-social development in children with autism and other special needs," in *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, pp. 695–700, 2019. doi: 10.23919/SICE.2019.8859853

[101] A. Taheri, A. Meghdari, M. Alemi, and H. Pouretemad, "Teaching music to children with autism: A social robotics challenge," *Scientia Iranica*, vol. 26, no. Special Issue on: Socio-Cognitive Engineering, pp. 40–58, 2019. doi: 10.24200/sci.2017.4608

[102] Y. Nakadoi, "Usefulness of animal type robot assisted therapy for autism spectrum disorder in the child and adolescent psychiatric ward," in *New Frontiers in Artificial Intelligence*, M. Otake, S. Kurahashi, Y. Ota, K. Satoh, and D. Bekki, Eds., pp. 478–482. Cham: Springer International Publishing, 2017. doi: 10.1007/978-3-319-50953-2_35

[103] R. Suzuki and J. Lee, “Robot-play therapy for improving prosocial behaviours in children with autism spectrum disorders,” in *2016 International Symposium on Micro-NanoMechatronics and Human Science (MHS)*, pp. 1–5, 2016. doi: 10.1109/MHS.2016.7824238

[104] B. Han, D. Yim, Y. T. Kim, S. jung Lee, and K. H. Hong, “The effect of a story intervention on the syntactic skills of children with autism spectrum disorders by using an educational humanoid robot,” *Communication Sciences and Disorders*, vol. 21, pp. 244–261, 2016. doi: 10.12963/csd.16316

[105] K. Arent, J. Kruk-Lasocka, T. Niemiec, and R. Szczepanowski, “Social robot in diagnosis of autism among preschool children,” in *2019 24th International Conference on Methods and Models in Automation and Robotics (MMAR)*, pp. 652–656, 2019. doi: 10.3390/app12178399

[106] 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. doi: 10.3389/frobt.2020.00043

[107] A. Aryania, H. S. Aghdasi, E. A. Beccaluva, and A. Bonarini, “Social engagement of children with autism spectrum disorder (asd) in imitating a humanoid robot: a case study,” *SN Applied Sciences*, vol. 2, no. 6, p. 1085, 2020. doi: 10.1007/s42452-020-2802-4

[108] G.-b. Wan, F.-h. Deng, Z.-j. Jiang, S.-z. Lin, C.-l. Zhao, B.-x. Liu, G. Chen, S.-h. Chen, X.-h. Cai, H.-b. Wang, L.-p. Li, T. Yan, and J.-m. 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, no. 3, pp. 374–387, 2019. doi: 10.1631/FITEE.1800555

[109] E. Short, E. Deng, D. Feil-Seifer, and M. Mataric, “Understanding agency in interactions between children with autism and socially assistive robots,” *Journal of Human-Robot Interaction*, vol. 6, p. 21, 2017. doi: 10.5898/JHRI.6.3.Short

[110] E. J. Choi, Y. T. Kim, S. J. Yeon, J. Kim, and K.-H. Hong, “Effects of robot and computer-based intervention on learning action word symbols of aac for children with autism spectrum disorder,” *Journal of College Student Development*, vol. 21, pp. 744–759, 2016. doi: 10.12963/csd.16344

[111] X. Liu, X. Zhou, C. Liu, J. Wang, X. Zhou, N. Xu, and A. Jiang, “An interactive training system of motor learning by imitation and speech instructions for children with autism,” in *2016 9th International Conference on Human System Interactions (HSI)*, pp. 56–61, 2016. doi: 10.1109/HSI.2016.7529609

[112] 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. doi: 10.1109/ROMAN.2018.8525634

[113] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, “Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with asd,” *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 2, pp. 125–135, 2018. doi: 10.1109/THMS.2017.2776865

[114] F. Askari, H. Feng, T. D. Sweeny, and M. H. Mahoor, “A pilot study on facial expression recognition ability of autistic children using ryan, a rear-projected humanoid robot,” in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 790–795, 2018. doi: 10.1109/ROMAN.2018.8525825

[115] K. Carlson, A. H. Y. Wong, T. A. Dung, A. C. Y. Wong, Y. K. Tan, and A. Wykowska, “Training autistic children on joint attention skills with a robot,” in *Social Robotics*, S. S. Ge, J.-J. Cabibihan, M. A. Salichs, E. Broadbent, H. He, A. R. Wagner, and Á. Castro-González, Eds., pp. 86–92. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-030-05204-1_9

[116] I. Giannopulu, V. Montreynaud, and T. Watanabe, “Minimalistic toy robot to analyze a scenery of speaker–listener condition in autism,” *Cognitive Processing*, vol. 17, no. 2, pp. 195–203, May 2016. doi: 10.1007/s10339-016-0752-y

[117] S. M. Srinivasan, I.-M. Eigsti, L. Neelly, and A. N. Bhat, “The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (asd): A pilot randomized controlled trial,” *Research in Autism Spectrum Disorders*, vol. 27, pp. 54 – 72, 2016. doi: 10.1016/j.rasd.2016.01.004

[118] W.-C. So, M. K.-Y. Wong, C. K.-Y. Lam, W.-Y. Lam, A. T.-F. Chui, T.-L. Lee, H.-M. Ng, C.-H. Chan, and D. C.-W. Fok, “Using a social robot to teach gestural recognition and production in children with autism spectrum disorders,” *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 6, pp. 527–539, 2018. doi: 10.1080/17483107.2017.1344886

[119] W.-C. So, M. K.-Y. Wong, W.-Y. Lam, C.-H. Cheng, J.-H. Yang, Y. Huang, P. Ng, W.-L. Wong, C.-L. Ho, K.-L. Yeung, and C.-C. Lee, “Robot-based intervention may reduce delay in the production of intransitive gestures in chinese-speaking preschoolers with autism spectrum disorder,” *Molecular Autism*, vol. 9, no. 1, p. 34, 2018. doi: 10.1186/s13229-018-0217-5

[120] C. Pop, B. Vanderborght, and D. David, “Robot-enhanced cbt for dysfunctional emotions in social situations for children with asd,” *Journal of Evidence-Based Psychotherapies*, vol. 17, pp. 119–132, 2017. doi: 10.24193/jebp.2017.2.7

[121] S.-S. Yun, J. Choi, S.-K. Park, G.-Y. Bong, and H. Yoo, “Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system,” *Autism Research*, vol. 10, no. 7, p. 1306–1323, 2017. doi: 10.1002/aur.1778

[122] World Health Organization, “International classification of functioning, disability and health : Icf,” 2001.

[123] World Health Organization Team, *Towards a Common Language for Functioning, Disability and Health: ICF*. Geneva: World Health Organization, 2002.

[124] L. V. Hedges, “Distribution theory for glass’s estimator of effect size and related estimators,” *Journal of Educational Statistics*, vol. 6, no. 2, pp. 107–128, 1981. doi: 10.2307/1164588

[125] M. Borenstein, L. Hedges, J. Higgins, and H. Rothstein, *Introduction to meta-analysis*. wiley, 2009. doi: 10.1002/9780470743386

[126] Cochrane, “Revman.” [Online]. Available: https://revman.cochrane.org/

[127] J. A. C. Sterne, J. Savović, M. J. Page, R. G. Elbers, N. S. Blencowe, I. Boutron, C. J. Cates, H.-Y. Cheng, M. S. Corbett, S. M. Eldridge, J. R. Emberson, M. A. Hernán, S. Hopewell, A. Hróbjartsson,

D. R. Junqueira, P. Jüni, J. J. Kirkham, T. Lasserson, T. Li, A. McAleenan, B. C. Reeves, S. Shepperd, I. Shrier, L. A. Stewart, K. Tilling, I. R. White, P. F. Whiting, and J. P. T. Higgins, "Rob 2: a revised tool for assessing risk of bias in randomised trials," *BMJ*, vol. 366, 2019. doi: 10.1136/bmj.l4898

[128] J. Higgins, J. Thomas, J. Chandler, M. Cumpston, T. Li, M. Page, and V. Welch, *Cochrane handbook for systematic reviews of interventions version 6.5*. Cochrane, 2024. [Online]. Available: www.training.cochrane.org/handbook

[129] C. Ricciardi, N. Pisani, L. Donisi, F. Abate, M. Amboni, P. Barone, M. Picillo, M. Cesarelli, and F. Amato, "Agreement between optoelectronic system and wearable sensors for the evaluation of gait spatiotemporal parameters in progressive supranuclear palsy," *Sensors (Basel)*, vol. 23, no. 24, 2023. doi: 10.3390/s23249859

[130] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, vol. 3, no. 19, 2018. doi: 10.1126/scirobotics.aao6760

[131] H. Javed, R. Burns, M. Jeon, A. M. Howard, and C. H. Park, "A robotic framework to facilitate sensory experiences for children with autism spectrum disorder: A preliminary study," *Journal of Human-Robot Interaction*, vol. 9, no. 1, 2019. doi: 10.1145/3359613

[132] C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, and M. Matarić, "Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders," *Frontiers in Robotics and AI*, vol. 6, p. 110, 2019. doi: 10.3389/frobt.2019.00110

[133] M. Meshry, M. E. Hussein, and M. Torki, "Linear-time online action detection from 3D skeletal data using bags of gesturelets," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*, pp. 1–9, 2016. doi: 10.1109/WACV.2016.7477587

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

[135] S. R. Ltda, "Eyelink 1000 plus user manual (version 1.0.12)," 2013-2017.

[136] T. Studio, "User manual—tobii studio (version 3.3. 0)," 2015.

[137] M. Tölgyessy, M. Dekan, and L. Chovanec, "Skeleton tracking accuracy and precision evaluation of kinect v1, kinect v2, and the azure kinect," *Applied Sciences*, vol. 11, no. 12, 2021. doi: 10.3390/app11125756

[138] G. Kurillo, E. Hemingway, M.-L. Cheng, and L. Cheng, "Evaluating the accuracy of the azure kinect and kinect v2," *Sensors (Basel)*, vol. 22, no. 7, 2022. doi: 10.3390/s22072469

[139] S. Moon, Y. Park, D. W. Ko, and I. H. Suh, "Multiple kinect sensor fusion for human skeleton tracking using kalman filtering," *International Journal of Advanced Robotic Systems*, vol. 13, no. 2, p. 65, 2016. doi: 10.5772/62415

[140] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding*, vol. 210, p. 103225, 2021. doi: https://doi.org/10.1016/j.cviu.2021.103225

[141] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7122–7131, 2017. doi: 10.1109/CVPR.2018.00744

[142] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “Smpl: a skinned multi-person linear model,” *ACM Transactions on Graphics*, vol. 34, no. 6, 2015. doi: 10.1145/2816795.2818013

[143] N. Hesse, S. Pujades, J. Romero, M. J. Black, C. Bodensteiner, M. Arens, U. G. Hofmann, U. Tacke, M. Hadders-Algra, R. Weinberger, W. Müller-Felber, and A. Sebastian Schroeder, “Learning an infant body model from rgb-d data for accurate full body motion analysis,” in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds., pp. 792–800. Cham: Springer International Publishing, 2018. doi: 10.1007/978-3-030-00928-1_89

[144] G. Luzhnica, J. Simon, E. Lex, and V. Pammer-Schindler, “A sliding window approach to natural hand gesture recognition using a custom data glove,” in *2016 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 81–90, 2016. doi: 10.1109/3DUI.2016.7460035

[145] F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, “Motion analysis: Action detection, recognition and evaluation based on motion capture data,” *Pattern Recognition*, vol. 76, pp. 612–622, 2018. doi: 10.1016/j.patcog.2017.12.007

[146] R. Ibañez, Á. Soria, A. Teyseyre, and M. Campo, “Easy gesture recognition for Kinect,” *Advances in Engineering Software*, vol. 76, pp. 171–180, 2014. doi: 10.1016/j.advengsoft.2014.07.005

[147] I. J. Ding and C. W. Chang, “Feature design scheme for Kinect-based DTW human gesture recognition,” *Multimedia Tools and Applications*, vol. 75, no. 16, pp. 9669–9684, 2016. doi: 10.1007/s11042-015-2782-3

[148] A. Taheri, A. Meghdari, and M. H. Mahoor, “A Close Look at the Imitation Performance of Children with Autism and Typically Developing Children Using a Robotic System,” *International Journal of Social Robotics*, pp. 6–10, 2020. doi: 10.1007/s12369-020-00704-2

[149] Z. Zeng, Q. Gong, and J. Zhang, “CNN Model Design of Gesture Recognition Based on Tensorflow Framework,” in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, pp. 1062–1067, 2019. doi: 10.1109/ITNEC.2019.8729185

[150] Y. Du, Y. Fu, and L. Wang, “Skeleton based action recognition with convolutional neural network,” in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 579–583, 2015. doi: 10.1109/ACPR.2015.7486569

[151] A. Papadakis, E. Mathe, I. Vernikos, A. Maniatis, E. Spyrou, and P. Mylonas, “Recognizing human actions using 3d skeletal information and cnns,” in *Engineering Applications of Neural Networks 2019*. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-20257-6_44

[152] E. Mathe, A. Mitsou, E. Spyrou, and P. Mylonas, “Arm gesture recognition using a convolutional neural network,” in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 37–42, 2018. doi: 10.1109/SMAP.2018.8501886

[153] L. Zhang, G. Zhu, P. Shen, J. Song, S. Afaq Shah, and M. Bennamoun, “Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition,” in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3120–3128, 2017. doi: 10.1109/ICCVW.2017.369

[154] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, “Rgb-d-based human motion recognition with deep learning: A survey,” *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018. doi: https://doi.org/10.1016/j.cviu.2018.04.007

[155] J. qing Liu, R. Fujii, T. Tateyama, Y. Iwamoto, and Y.-W. Chen, “Kinect-based gesture recognition for touchless visualization of medical images,” *International Journal of Computer and Electrical Engineering*, vol. 9, no. 2, pp. 421–429, 2017. doi: 10.17706/IJCEE.2017.9.2.421-429

[156] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damaševičius, R. Maskeliūnas, and K. H. Abdulkareem, “Real-time hand gesture recognition based on deep learning yolov3 model,” *Applied Sciences*, vol. 11, p. 4164, 2021. doi: 10.3390/app11094164

[157] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “OpenFace 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019

[158] Y. Zhou and J. Gregson, “Whenet: Real-time fine-grained estimation for wide range head pose,” in *31st British Machine Vision Conference (BMVC)*, 2020.

[159] 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. 334–352, 2018. doi: 10.1007/978-3-030-01249-6_21

[160] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016. doi: 10.1109/CVPR.2016.91

[161] R. A. Güler, N. Neverova, and I. Kokkinos, “Densepose: Dense human pose estimation in the wild,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7297–7306, 2018. doi: 10.1109/CVPR.2018.00762

[162] G. Alvari, L. Coviello, and C. Furlanello, “EYE-C: Eye-contact robust detection and analysis during unconstrained child-therapist interactions in the clinical setting of autism spectrum disorders,” *Brain Sciences*, vol. 11, no. 12, 2021. doi: 10.3390/brainsci11121555

[163] G. Fassina, L. Santos, A. Geminiani, A. Caglio, S. Annunziata, I. Olivieri, and A. Pedrocchi, “Development of an interactive total body robot enhanced imitation therapy for asd children,” in *2022 International Conference on Rehabilitation Robotics (ICORR)*, pp. 1–6, 2022. doi: 10.1109/ICORR55369.2022.9896536

[164] C. C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, and J. T. Mcconville, “Anthropometric survey of u.s. army personnel: Methods and summary statistics 1988,” ANTHROPOLOGY RESEARCH PROJECT INC YELLOW SPRINGS OH, Tech. Rep., 1989.

[165] M. Reed and K. DeSantis Klinich, *A New Database of Child Anthropometry and Seated Posture for Automotive Safety Applications*. SAE International, 2010.

[166] 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. doi: 10.1016/j.jml.2015.07.004

[167] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochemia medica*, vol. 22, pp. 276 – 282, 2012. doi: 10.11613/BM.2012.031

[168] M. Murgo, “Objective attention classifier for evaluation of new robotic therapy for asd children,” Master’s thesis, Politecnico di Milano, 2022.

[169] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, p. 381–395, 1981. doi: 10.1145/358669.358692

[170] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, “Segment anything,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023. doi: 10.1109/ICCV51070.2023.00371

[171] L. R. Dice, “Measures of the amount of ecologic association between species,” *Ecology*, vol. 26, no. 3, pp. 297–302, 1945. doi: 0.2307/1932409

[172] T. Sørensen, “A method of establishing group of equal amplitude in plant sociobiology based on similarity of species content and its application to analyses of the vegetation on danish commons,” in *Kongelige Danske Videnskabernes Selskab*, vol. 5, no. 4, pp. 1–34, 1948.

[173] C. Guimarães, A. Cruz-Santos, and L. Almeida, “Inventário para o uso da linguagem (lui): Estudo piloto do instrumento de avaliação das competências pragmáticas em português,” in *II Seminário Internacional Contributos da Psicologia em Contextos Educativos*, 2012.