Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability

DI NUOVO, Alessandro <http://orcid.org/0000-0003-2677-2650>, CONTI, Daniela <http://orcid.org/0000-0001-5308-7961>, TRUBIA, Grazia, BUONO, Serafino and DI NUOVO, Santo

robotics

MDPI

Article

Deep Learning Systems for Estimating Visual Attention in Robot-Assisted Therapy of Children with Autism and Intellectual Disability

Alessandro Di Nuovo [1,*], Daniela Conti [1], Grazia Trubia [2], Serafino Buono [2] and Santo Di Nuovo [3]

1 Sheffield Robotics, Sheffield Hallam University, Sheffield S1 1WB, UK; d.conti@shu.ac.uk
2 Psychology Operative Unit, IRCSS Oasi Maria SS, 94018 Troina, Italy; gtrubia@oasi.en.it (G.T.); fbuono@oasi.en.it (S.B.)
3 Department of Educational Sciences, University of Catania, 95124 Catania, Italy; sdinuovo@unict.it
* Correspondence: a.dinuovo@shu.ac.uk; Tel.: +44-114-225-6958

Received: 1 May 2018; Accepted: 30 May 2018; Published: 4 June 2018

Abstract: Recent studies suggest that some children with autism prefer robots as tutors for improving their social interaction and communication abilities which are impaired due to their disorder. Indeed, research has focused on developing a very promising form of intervention named Robot-Assisted Therapy. This area of intervention poses many challenges, including the necessary flexibility and adaptability to real unconstrained therapeutic settings, which are different from the constrained lab settings where most of the technology is typically tested. Among the most common impairments of children with autism and intellectual disability is social attention, which includes difficulties in establishing the correct visual focus of attention. This article presents an investigation on the use of novel deep learning neural network architectures for automatically estimating if the child is focusing their visual attention on the robot during a therapy session, which is an indicator of their engagement. To study the application, the authors gathered data from a clinical experiment in an unconstrained setting, which provided low-resolution videos recorded by the robot camera during the child–robot interaction. Two deep learning approaches are implemented in several variants and compared with a standard algorithm for face detection to verify the feasibility of estimating the status of the child directly from the robot sensors without relying on bulky external settings, which can distress the child with autism. One of the proposed approaches demonstrated a very high accuracy and it can be used for off-line continuous assessment during the therapy or for autonomously adapting the intervention in future robots with better computational capabilities.

Keywords: deep learning neural networks; face detection and alignment; robot-assisted therapy; intellectual disability; autism spectrum disorder; human–robot interaction

1. Introduction

Recent technology solutions in intelligent systems and robotics have made possible innovative intervention and treatment for individuals affected by Autism Spectrum Disorder (ASD), which is a pervasive neurodevelopmental disorder in which deficits in social interaction and communication can make ordinary life challenging from childhood through adulthood [1]. Children with ASD often experience anxiety when interacting with other people due to the complexity and unpredictability of human behaviors. The controllable autonomy of robots has been exploited to provide acceptable social partners for these children [2] in the treatment of this disorder. Indeed, several studies have shown that some individuals with ASD prefer robots to humans and that robots generate a high degree of

motivation and engagement, including children who are unlikely or unwilling to interact socially with human therapists (see [3] for a review).

Epidemiological data show that ASD can often comorbid with some level of Intellectual Disability (ID) [4], in fact, it has been reported that 54% of children with ASD have an IQ below 85, which makes therapeutic interventions more difficult due to the limited capabilities of the subjects, who often need hospitalization. The treatment of these children is more likely to benefit from the introduction of technological aids. Regarding therapeutic training of imitation presented in this article for instance, the standard approach is to employ two clinical personnel: one must be close to constantly support the child while another performs the tasks to imitate. Intelligent semi-autonomous systems like robots could provide assistance by performing the imitation tasks in this case, and, therefore, require only one therapist to support the child while controlling the system.

The work presented in this paper is part of the EU H2020 MSCA-IF CARER-AID project, which aims to improve robot-assisted therapy interventions for children with ASD and ID via an automatic personalization of the robot behavior that should meet the patient's condition. The aim is to fully integrate the robot within a standard treatment, the TEACCH [5] (Treatment and Education of Autistic and related Communication Handicapped Children) approach, which is commonly used for the treatment of ASD.

One of the most common characteristics of ASD is the impairment in using eye gaze to establish attention in social interaction and it is one of the most important of the traits that are assessed for diagnosis and one of the main areas of intervention for therapy [6].

Small robotic platforms that are being used for robot-assisted therapy have usually limited sensors on-board and the use of external devices is very common in Human–Robot Interaction (HRI) for computing the human behavior in this context. Therefore, the challenge is to estimate the child's visual attention directly from the robot cameras, possibly without the need of external devices, such as high-resolution cameras and/or Microsoft Kinects (or equivalent), which can definitely increase the performance, but at the same time limit the portability of the system and make more difficult its actual integration within the standard therapeutic environment.

Moreover, especially in the case of ID, children are unlikely to adhere to any imposed constraint like those typically required to maximize algorithm performance. Therefore, flexibility and adaptability are essential pre-requisites for the inclusion of any technology in the actual therapy [7]. Conversely, in the case of robot-assisted therapy, preliminary experiments suggest that deriving attentional focus from a single camera would not be accurate enough [8]. Quite the reverse, more recent experiments found acceptable levels of agreement between the robot observations and manual annotations [9].

To investigate this issue, the authors considered training attention classifiers based on state-of-the-art algorithms of the popular deep-learning neural networks family [10,11]. These artificial neural network architectures achieved exceptional results in computer vision [12] and the authors hypothesize that these technological improvements could empower the robots and allow them to perform reliably for vision tasks like estimating the engagement via the attention focus, and use this information to personalize the interaction.

The results of the evaluation of two architectures to estimate the child's attention to the robot from low-resolution video recordings taken from the robot camera while interacting with the children are presented here. The approach was tested on videos collected during up to 14 sessions of robot-assisted therapy in an unconstrained real setting with six hospitalized children with ASD and ID. The clinical results of the field experiment presented in this article are analyzed and discussed in [13,14].

The rest of the paper is organized as follows: Section 2 reviews and the scientific literature that constitutes the background for this work; Section 3 introduces our approach for classifying the child attention status during the robot-assisted therapy, describes the procedure to collect the video recordings and produce the dataset for the experimentation, and present the algorithms that constitutes our system; Section 4 analyses the numerical results, finally Section 5 gives our discussion and conclusion.

2. Background

2.1. Social Assistive Robotics for Children with Autism

Considering the complexity and the wide amplitude of the autism "spectrum", which encompasses different disabilities and severity levels, it is appropriate to use a multi-modal intervention that can be adapted to the individual's needs to obtain the best benefits from the therapy. Therefore, the controllable autonomy of robots has been exploited to provide acceptable social partners for these children [2].

Socially Assistive Robotics (SAR) is a novel field of application in robotics, which merges assistive and social robotics to design new platforms and services to help users through advanced interaction driven by their needs (e.g., tutoring, physical therapy, daily life assistance, emotional expression) via multimodal interfaces (speech, gestures, and input devices) [15]. Several studies in this area, have shown that some individuals with ASD prefer social robots to humans. The social robot "Probo" could help the social performance of some children with ASD in specific situations like making a fruit salad [16] for example. Robins et al. [17] showed that children with ASD and limited verbal skills prefer interacting with robots than humans. Recent studies have successfully presented robots as mediators between humans and individuals with ASD [18]. Duquette et al. [19] show improvements in affective behavior and attention sharing with co-participating human partners during an imitation task solicited by a simple robotic doll, for instance,. Social robots might be especially beneficial for individuals with ASD who face communication difficulties because practicing communication can be less intimidating with a robot than with another person [20,21].

Humanoid robots, which look like a human being but are much less complex compared to the human, could make learning easier for a child with ASD and then facilitate the transfer of skills learned through models of imitative human–robot interaction to child-human interaction [22]. Indeed, imitation as a means of communication can be related to positive social behavior and it is considered a good predictor of social skills [23]. Children with autism are often characterized by difficulties in imitating the behavior of other people and, therefore, imitation is employed in therapy to promote better body awareness, sense of self, creativity, leadership, and ability to initiate interaction [24].

Recent robotics research has shown numerous benefits of robot assistants in the treatment of children with ASD [17,25].

2.2. Visual and Social Attention in Children with ASD and Robots

People naturally tend to look and focus their attention on objects which are of immediate interest [26]. Visual attention also is normally established in social contexts like a conversation between two people, and the ability to correctly simulate focusing the visual attention on the interaction partner is considered a way for robots to exhibit social intelligence and awareness and facilitate HRI [27].

This natural process can be identified as part of the social attention, which unifies several domains of attention in social contexts [28]. The authors use social attention to identify the domain of social motivation, in this article, as is common in the case of ASD, indeed, they look at the capability of the children to direct the visual focus of attention on the interaction partner, i.e., the robot, as a way for assessing their engagement in the therapeutic activities.

Due to this innate attitude of human beings toward social attention, the deficit in social attention with others is one of the most noticeable features of ASD [29] and it is among the earliest signs of the disorder [30], consequently it plays a crucial role in the detection and assessment of this disorder as well as in therapies.

Thus, it is crucial that a robot assistant has the capability to evaluate if the child is looking at itself while engaged in a therapeutic session. This capability can be used in the assessment of the child's condition for the diagnosis and, then, during the therapy to keep track of progress. Furthermore, the robot can use this information to autonomously adapt the interaction with the child during the

therapeutic session, for instance playing some sound for attracting the child's attention when he/she is distracted.

The authors remark that they use "Attention" to identify when the child is directing the attentional focus on the robot or "Distraction" when he/she is not monitoring or following the robot prompts.

2.3. Video Analysis for Face Detection

Human face localization plays an essential role in a countless number of applications such as human-computer and human–robot interaction systems [27,31], pilot/driver attention levels [32,33], video surveillance [34], face recognition [35], and facial expression analysis [36,37].

Face detection aims at finding whether faces are present in a given image and, if any is found, return their locations through bounding boxes expressed in pixel coordinates, e.g., upper and lower corners of the rectangle that contains the face, in practice.

Generally, the application of a face detection algorithm is a prerequisite for most face recognition and tracking algorithms, which assume that face location is known [38]. Few techniques originally developed for face recognition have also been used to detect faces, on the contrary, but their computational requirements are excessive for the task and demonstrated limited generalization capability as their performance is significantly reduced when applied to "uncultivated" contexts [39].

Recent studies show that deep learning neural network approaches can achieve impressive performances in face detection and face alignment [38]. Face alignment algorithms aim at identifying the geometric structure of human faces in digital images and return an estimate of the face's landmark positions, also known as facial features points. Facial alignment calculation usually begins from a rectangular bounding box returned by a face detector [40]. However, novel deep learning algorithms have been proposed to simultaneously detect face location and estimate basic landmarks. The algorithm the authors used for this work estimates eyes, nose, mouth corners [41], for instance.

2.4. Face Detection and Alignment for Estimating the Visual Focus of Attention in HRI

Face detection and localization of facial features are part of the common processing scheme for determining driver visual attention from video cameras [42]. Facial features can be used to estimate the head pose [43], which is a reliable indicator of the gaze, and, thus, of the visual focus of attention, which can be effectively estimated by head pose only [44,45].

Visual attention in human–robot interaction has been widely studied because its crucial but, it is often evaluated using external devices, such as high-resolution cameras and/or kinetics, e.g., [46,47], due to the limited video recording resources on the most common commercial robotic platforms for social applications. Authors proposed a system for robot-assisted therapy that employs five individual sensors, including three RGB cameras and two Microsoft Kinects [18].

Research on face detection from a single camera has been carried out in the field of ASD, where the majority of the work has focused on tracking the face of the individual with ASD and moving the robot face accordingly [48–50]. Indeed, these studies usually employed standard software like OpenCV and did not evaluate the accuracy of the algorithms. Two studies evaluated the face detection performance from a single camera in the context of robot-assisted therapy, but they came to different conclusions: other authors suggest that deriving attentional focus from a single camera would not be accurate enough [8]; on the other hand, recent experiments found acceptable levels of agreement between the robot observations and manual post-hoc annotations [9]. However, none of them tested the system in the clinical context with children with ASD and ID.

3. Materials and Methods

The methodology used for this work comprises gathering data from a field experiment in which the authors tested the integration of robot-assisted sessions in the standard therapy of hospitalized children with ASD and ID. The video recordings collected during the field experiment are used for

training and evaluating an attention classifier based on the low-resolution video recording for the robot-assistant during the therapy sessions.

Section 3.1 describes the clinical experiment and explains how data has been collected and its format. Section 3.2 presents our architecture, its components and a widely used algorithm for real-time face detection used for comparison.

3.1. Data Gathering from the Clinical Experiment

This section presents the details of the clinical experiment that generated the recordings used in to create the database for the machine learning experiments. The clinical results of this experiment are presented in [13,14]. To summarize, the robot-assisted therapy demonstrated to be effective in imitation training for 4 children with ID levels from mild to severe, who were able to learn all three imitation tasks. However, the two other children with the lowest IQ (profound ID) did not show significant improvement, probably due to their severe limitation in comprehending the instructions. This result suggests that other ways of interaction or technologies should be investigated for these extremely difficult cases.

3.1.1. Participants

Six children were selected among patients diagnosed with ASD and ID, who are currently receiving treatment at the IRCCS Oasi Maria SS of Troina (Italy), a specialized institution for the rehabilitation of intellectual disabilities. Children are inpatients of the institution, where they live for most of the time and follow a clinical daily program of training using the TEACCH approach with psychologists and highly specialized personnel. All children have ASD of grade 3, while the ID levels range from mild to profound.

Participants' ASD and ID levels have been diagnosed before the start of this study with the following standard psycho-diagnostic instruments: Leiter-R, WISC, PEP-3, VABS, ADI-R, and CARS-2. Regarding more details about the diagnosis procedure and the instruments see [4,22].

Ethical approval had been obtained from both the ethical council of IRCSS Oasi Maria SS of Troina and Sheffield Hallam University. All the parents signed consent forms before their children were included in the study. Children were free to leave the experiment at any time and they were always supported by a professional educator other than the researchers.

3.1.2. The Robot Therapist: the Softbank Robotics Nao

The robot used for leading the robot-assisted therapy was the Softbank Robotics Nao v4, which is a small toy-like humanoid robot, very popular for child–robot interaction studies [18,22,51,52].

Unless otherwise specified, this study used the default settings and the standard equipment of the Nao robot v4, which includes two 1.22 Mega pixels cameras that can be used to take pictures and record videos from the robot's perspective. Camera and other sensor positions are shown in Figure 1. According to the specifications, when image resolution is up to 1280×960 pixels and video recording is up to 30 fps, the actual resolution and frame-rate are usually restricted to 320×240 and 10 fps due the limited computing capacity of the main processor and memory resources.

Among the software features, Nao has the face detection and tracking functionality that was used in the clinical experiments to direct the robot toward the child during the interaction.

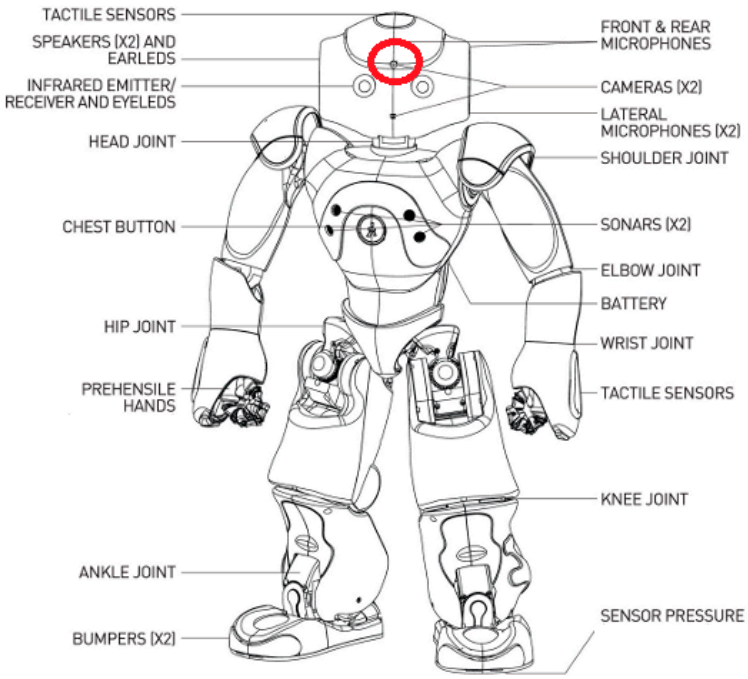

Figure 1. The Nao robot. The camera used for recording the child activity is the one on the top (circled in red).

3.1.3. Protocol for the Clinical Experiments

To assess the children's behavior the study adopted the Verbal Behaviour Milestones Assessment and Placement Program (VB-MAPP) [53]. To match the specific level and training need of the participants, a human therapist conducted a preliminary evaluation of the children's capability to perform the VB-MAPP imitation tasks of levels 1 and 2, and then those tasks that the children were not able to perform were selected and the robot programmed for their training. The authors selected three tasks (T1, T2, and T3) for the experiments, in which the children received a milestone score of 0 or 0.5, defining which of them were not able or did not perform the task properly.

The robot was included in the TEACCH program among the daily activities and identified via a specific "visual schedule". A visual schedule communicates the sequence of upcoming activities or events using objects, photographs, icons, words, or a combination of tangible supports. During each session, the children performed three gross motor imitation tasks managed by the robot only (Figure 2).

To facilitate the interaction, the robot-led therapy sessions were carried out in the same room where children usually did their treatment sessions. During a training encounter, the robot was deployed on a table, to be approximately at the same height of the child, initially at a distance of at least 1 m. The child could move backward or forward to be more comfortable. The training encounters usually comprised three sessions, one for each task. During each session, the children were encouraged to imitate the task performed by the robot. Tasks were proposed in a randomized modality to avoid stereotypical learning. First, the robot verbally presented the behavior to perform in a simple and clear language, then it solicited the child to imitate its movements while doing them (Figure 2).

A professional educator, selected among those involved in the everyday treatment of the children, was always present to represent a "secure base" for the children. The professional educator gave a positive verbal reinforcement ("good" and/or "right") along with, in some cases, a physical reinforcement (a caress). These reinforcements were different for each child and were connected directly to responses, behavior and to the child's difficulties. During all the tasks the robot called the children by name to make the interaction more personalized.

The procedure comprised a preliminary session to decrease the novelty effect. During this preliminary session, the robot was presented to all the children in a non-therapeutic context for a total of approximately 10 min.

The actual experimentation began 7 days after the preliminary encounter. The study comprises a total of 14 encounters over one month, i.e., 3 sessions per week. The total length of each session was approximately 6–8 min per child.

Figure 2. Example of the child–robot interaction during the therapeutic session (the child is imitating the robot moving his arm). A professional educator was always present nearby to support the child.

3.1.4. Video Recording and Labeling

During each therapeutic session, the interaction was recorded using the robot's top camera (Figure 1). The video recording was restricted to 320 × 240 pixels per frame up to 10 fps. The restriction is a default setting due to the limited computing capabilities of the robot's CPU, which was not capable of supporting a higher resolution/framerate video recording while executing the behavior for the therapy. The recordings included the execution of the rehabilitation tasks only. Figure 3 presents one frame as an example of the video recordings.

Figure 3. A frame extracted from one of the videos recorded by the robot (the child is imitating the robot raising his arms up).

The videos were resized to double their resolution to 640 × 480 using bicubic interpolation, because, in many cases, the children's faces were too small to be recognized by the algorithms.

The authors underline that during the sessions, children were free to move and, especially at the beginning, some were positioned more than 1 m from the robot, with the picture of their face contained in less than 20 × 20 pixels.

The robot face detection functionality was active and used with the intention to center the top camera toward the child during the interaction. However, due to the presence of a professional educator, the camera was occasionally centered on the educator's face. Video recordings used for this work were edited to remove the educator's face. To this end, the authors asked the educator to wear a green headband so that their face was easily recognizable with an inexpensive computational algorithm. The educators kindly agreed to comply with this request with no problem.

To build the ground truth for attention estimation experiments, some frames were extracted and manually annotated (attention vs distraction) by two researchers, which separately compiled a record sheet divided into frames. Once completed, the discrepancies were resolved via discussion. Then, the two researchers agreed on the final labels which were used to build the database for this study's machine learning experiments.

The authors remark that the labeling was considered "Attention" when the child was monitoring the robot, i.e., when the visual focus of attention was on the robot and considered "Distraction" otherwise.

3.2. Estimating Attention from Low-Resolution Camera Images

Figure 4 presents the methodology used in the study's machine learning experiments. The authors tested two alternatives with different expected performance and computational requirements for each step of the study approach, to analyze the benefit vs computational cost ratio. The following sections present the alternative algorithms tested in our experiments.

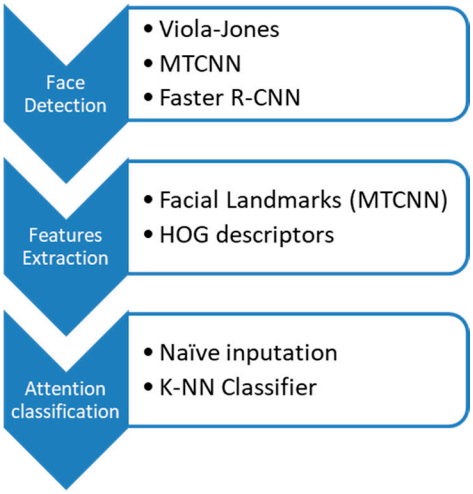

Figure 4. The steps of the study approach. Regarding each step, some alternative approaches were tested.

The Naïve estimation is an inexpensive algorithm that simply assumes that a child established visual attention if the robot detects the child's face.

3.2.1. The Viola–Jones (VJ) Framework for Face Detection

The Viola–Jones is a widely used method for object detection, in particular, face detection which was the first application demonstrated in the seminal paper [54]. The reasons for its success are the real-time computation and its robustness (very high true-positive and very low false-positive rates). It has been used as part of the procedure for the recognition of visual focus of attention and its level in human–robot interaction [27].

The method has 4 key characteristics: (i) Feature Selection using Haar Basis functions; (ii) Creating an "Integral Image" that allows for very fast feature evaluation; (iii) Adaboost Classifier Training; (iv) it combines successively more complex classifiers in a cascade structure to increase the speed of the detector by focusing attention on promising regions of the image.

The rationale for using a "Frontal Face" detector to estimate attention is that the latter is related to the head pose. The basic implementation of the VJ algorithm does not recognize partial faces, i.e., when the child's face is pointing elsewhere and not looking at the robot, in fact. Thus, the VJ approach will recognize a face only when it is orientated toward the robot, which could be interpreted as focusing the visual attention on it.

3.2.2. Faster Convolution Neural Networks with Region Proposal (R–CNN)

State-of-the-art object detection methods include the region proposal-based convolutional neural networks like the popular Faster R–CNN [55], which combine fast prediction with high accuracy by training an auxiliary neural network to make region proposals that can speed-up the execution thanks to a restriction of the search area for a Convolutional Neural Network (CNN). Faster R–CNN is composed of two modules: A Region Proposal Network (RPN) and a CNN detector, which could also be a pre-trained object detection network. A four-step alternating training is employed to better integrate both modules [55].

This study adopted a technique known as "transfer learning" [56], in which pre-trained CNN models are fine-tuned for a specific task using new data. The model that has been tuned in this study's computational experiments is the VGG-16 [57], which is a very deep CNN architecture for image recognition.

3.2.3. Multitask Cascaded Convolutional Networks for Face Detection and Landmark Estimation (MTCNN)

The MTCNN is a deep learning neural network architecture that employs multi-task cascaded CNNs, which are the state-of-the-art for many computer vision applications [58]. The three main characteristics of the MTCNN for performance improvement are: (i) the cascaded CNNs architecture; (ii) an online hard sample mining strategy; and (iii) joint face alignment learning. The last characteristic is of particular interest because facial landmarks could be used to estimate the head pose and, thus, the attention. The MTCNN can outperform state-of-the-art methods across several benchmarks while achieving real-time performance for 640 × 480 VGA images [41]. These are both interesting features for the application being investigated and, therefore, it has been selected for this study's experimental testing.

The coordinates of the landmarks were related to the bounding box position and rescaled according to its size in this study's method:

$$x_l = \frac{x_0 - b_1}{b_2 - b_1} \tag{1}$$

where x_l is the landmark coordinate used for classifying, x_0 is the coordinate in the image and b_1 is the top left corner and b_2 is the bottom right corner.

3.2.4. Histograms of Oriented Gradients (HOGs)

Histograms of oriented gradients (HOGs) are the concatenation of histograms obtained by dividing the image into small connected regions called cells. Thus, the local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. HOG descriptors are particularly suited for human detection as it has been shown that they outperform other feature sets in this task [59].

The authors extracted the HOG descriptors from the face image to test the possible improvement in the attention estimation in this study's approach. Thirty-six HOG features were calculated on the face, only as delimited by the bounding box, and rescaled to 20×20.

3.2.5. K–Nearest Neighbor Classifier (K–NN)

The authors trained a classifier using the K–Nearest Neighbor (K–NN) algorithm to distinguish the two statuses "Attention" (visual focus on the robot) or "Distraction" (not monitoring the robot) from the 5 Facial Landmarks (center of the eyes, nose, and edges of the mouth) or the HOG descriptors.

K–NN which is a classical computational intelligence technique widely used in pattern recognition and data mining applications since it can be applied to both supervised and unsupervised learning problems [60]. The method is based on the simple but powerful idea items can be associated by similarity and, therefore, when applied to classification, any new item can be labeled as the "most similar" in the labeled dataset [61]. This algorithm has been selected after preliminary experiments with different classifier methodologies using a Bayesian optimization process to explore the parameter space and select the best ones [62]. Indeed, in a ten-fold cross-validation, the K–NN classifier achieved the best result in terms of classification accuracy over Support Vector Machines, Decision Trees, and Naïve Bayes Classifier.

3.3. Performance Measures

The aim of the computational experiments was to validate the deep learning approaches and compare them to a standard approach for the evaluation of attention and distraction in human–robot interaction. Given that both conditions are important, along with the classic classification accuracy, the authors considered several measures that focus on the attention and distraction conditions as detailed in Table 1. Where P is the number of positive cases, i.e., frames with attention condition; N identifies the negative cases, i.e., frames with distraction condition; TP is the true positives (correct classification of attention); TN is the true negatives (correct classification of distraction); FP is the false positives (distraction that has been classified as attention); FN is the false negatives (attention that has been classified as distraction).

Table 1. Measures for evaluating classification performance.

Name	Calculation	Explanation and Interpretation
Accuracy	$\frac{TP+TN}{P+N}$	Accuracy is the proximity of the classification results to the true values. It evaluates the overall performance of classification.
Precision	$\frac{TP}{TP+FP}$	Precision is the positive predictive value. This case indicates the reliability of the classification of attention.
Negative Prediction	$\frac{TN}{TN+FN}$	The negative predictive value indicates the reliability of the classification of distraction.
Sensitivity	$\frac{TP}{P}$	Also known as Recall, Hit or True Positive Rate, Sensitivity focuses only on how good the performance in classifying attention is.
Specificity	$\frac{TN}{N}$	Opposite to Sensitivity, Specificity or true negative rate is evaluating only the performance in classifying distraction.

A higher value indicates a better performance for all the measures considered in comparisons.

The authors also have calculated the Receiver Operating Characteristic (ROC) curve, which is a graph that exemplifies the performance of a binary classifier by plotting Sensitivity (true positive rate) against the false positive rate (FP/P) at various threshold settings. The points of the ROC curve are calculated from the K–NN classification scores (group similarities) and varying the discrimination thresholds. A measure for model comparison is the Area Under the Curve (AUC), which, in the case

of ROC, can be interpreted as the probability that a higher score is assigned to a randomly selected positive example than to a random negative example [63].

4. Computational Experiments and Results

4.1. Database Creation

The authors built the training set for the classifier by extracting 4794 frames from the videos recorded during the preliminary encounters. These examples of child–robot interaction were then used both to build ad-hoc detectors using the R-CNN architecture and to train the attention classifier (K–NN) to discriminate between attention and distraction from landmarks or HOGs features.

The authors extracted 31726 frames (i.e., 2 frames per second) from 204 videos recorded by the robot during the therapy sessions for the test set. Video length varied according to the number of tasks to be performed, which was dynamically adapted to the child's condition for each session. The duration ranges from 52 to 126 s, with an average of 77.75 s.

Following the methodology described in Section 3.1.4, two researchers separately labeled the frames with the inter-coder agreement score which was 0.94, producing a reliability (measured by Cohen's kappa) of 0.85. The final labels were agreed upon by both researchers after discussion.

Table 2 reports the total average time each child was demonstrating focusing the visual attention on the robot during the therapy sessions.

Table 2. The rate of attention for each child, the average of all the robot-assisted therapy sessions.

The Rate of Child Attention on the Robot Activities	
C1	5.70%
C2	63.4%
C3	48.3%
C4	67.3%
C5	78.0%
C6	33.3%

The robot-assisted therapy had a different impact on the children, who also have different types of ASD and level of ID. Indeed, the children cover a wide range, with C1 who was the least interested in the robot with as little as 5.7% of attention, while C5 showed high attention consistently during all the sessions. Details of the performance of the automatic attention estimation for each child are in Appendix A.

4.2. Classification Results

The following tables present the performance metrics for the approaches considered; metrics were calculated for each of the 204 recorded videos. The descriptive statistics reported are the Average (avg), Median (median), Maximum (max), Minimum (min) and the standard deviation (*SD*). The detailed results for each child are reported in Appendix A.

The first analysis in Table 3 focus on the Viola–Jones algorithm, which has been tested with both a Naïve classification, i.e., "attention" if a face is detected, "distraction" if no face is detected. The performance metrics of this classic approach show a good result despite its simplicity.

Despite the HOGs extraction and K–NN classification, Table 4 shows very little improvement in the performance of the VJ approach. Indeed, there is a small increase in accuracy and a decrease in sensitivity, i.e., a reduction in true attention classifications. This means that the better performance has been caused by the K-NN classification of the HOGs that underestimates attention producing false detections of distraction.

Table 3. Results of the Viola–Jones with Naïve Classification.

	Avg	Median	Max	Min	*SD*
Accuracy	0.734	0.763	0.957	0.212	0.115
Precision	0.706	0.852	1.000	0.082	0.126
Sensitivity	0.678	0.764	1.000	0.009	0.226
Specificity	0.776	0.797	1.000	0.172	0.170
Negative Prediction	0.696	0.695	1.000	0.078	0.139

Table 4. Results of the Viola–Jones with histograms of oriented gradients (HOGs) and K–Nearest Neighbor (K–NN) classification.

	Avg	Median	Max	Min	*SD*
Accuracy	0.754	0.780	0.969	0.212	0.106
Precision	0.729	0.859	1.000	0.119	0.131
Sensitivity	0.651	0.728	1.000	0.010	0.213
Specificity	0.812	0.836	1.000	0.206	0.146
Negative Prediction	0.706	0.696	1.000	0.157	0.132

Quite the opposite, the Naïve classification of R–CNN detections shows a better sensitivity, i.e., true classification of attention, but achieves the worst overall performance as it overestimates attention by detecting partial faces. Results for the Naïve R–CNN are in Table 5.

Table 5. Results of the Faster Convolution Neural Networks with Region Proposal (R–CNN) Naïve classification.

	Avg	Median	Max	Min	*SD*
Accuracy	0.696	0.725	0.956	0.206	0.123
Precision	0.663	0.767	1.000	0.087	0.138
Sensitivity	0.759	0.819	1.000	0.083	0.201
Specificity	0.642	0.681	1.000	0.096	0.231
Negative Prediction	0.583	0.563	1.000	0.036	0.189

Table 6 reports a significant improvement by applying the K–NN classification to the HOGs extracted from the R–CNN detections. This can be explained by the fact that the R–CNN detects more faces (true positives) than the VJ algorithm thanks to the ad-hoc training. Meanwhile, the K–NN classification stage reduces the false positive detections and therefore improves the overall performance.

Table 6. Results of the R–CNN with HOGs and K–NN classification.

	Avg	Median	Max	Min	*SD*
Accuracy	0.846	0.850	0.939	0.652	0.055
Precision	0.745	0.847	0.995	0.184	0.122
Sensitivity	0.858	0.908	1.000	0.250	0.095
Specificity	0.767	0.754	0.973	0.364	0.102
Negative Prediction	0.782	0.883	0.962	0.261	0.099

The performance behavior of the MTCNN is reported in Tables 7 and 8 and it is similar to the R–CNN. The version with the landmarks has the best sensitivity and reliability in the prediction of the negative statuses, i.e., distraction. However, this is achieved practically by overestimating attention which results in many false positives and the worst specificity (Table 7). This performance is too low, and it cannot be considered reliable enough for a real application.

Table 7. Results of the Multitask Cascaded Convolutional Networks (MTCNN) Landmarks classification via K–NN.

	Avg	Median	Max	Min	*SD*
Accuracy	0.709	0.772	0.966	0.107	0.174
Precision	0.619	0.694	0.982	0.068	0.195
Sensitivity	0.953	0.966	1.000	0.714	0.038
Specificity	0.483	0.490	0.919	0.009	0.215
Negative Prediction	0.902	0.932	1.000	0.333	0.092

Table 8. Results of the MTCNN with HOGs and K–NN classification.

	Avg	Median	Max	Min	*SD*
Accuracy	0.882	0.893	0.972	0.668	0.066
Precision	0.833	0.854	0.949	0.580	0.098
Sensitivity	0.830	0.864	0.978	0.406	0.138
Specificity	0.873	0.888	0.990	0.661	0.079
Negative Prediction	0.818	0.852	0.947	0.452	0.120

Finally, in Table 8 one can see the best overall result in term of all the performance metrics considered.

To summarize the results and show a direct comparison, Figure 5 presents the ROC curves for all the approaches considered.

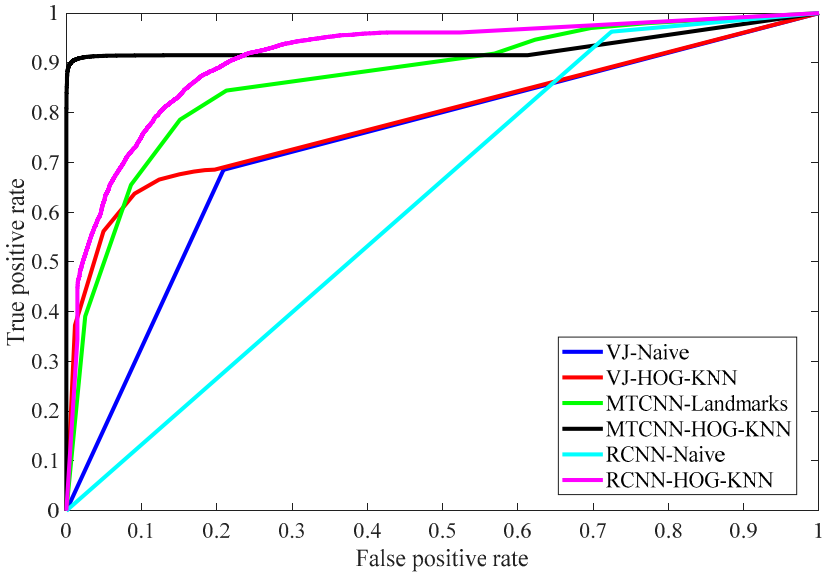

Figure 5. Receiver Operating Characteristic (ROC) curves calculated from the K–NN classification scores with varying discrimination thresholds. Note that Naïve classifiers produced only binary (0 or 1) scores.

The AUCs are: VJ-Naïve 0.7382; VJ-HOG-K–NN 0.7926; MTCNN-Landmarks-K–NN 0.8681; MTCNN-HOG-K–NN 0.9314; RCNN-Naïve 0.6192; RCNN-HOG-K–NN 0.9128. The AUC values are consistent with the other metrics as they confirm that the classification of HOGs with the K–NN is best approach and that the MTCNN is the most efficient algorithm for detecting the faces. However, AUC values give a biased ranking as this measure is based on the true and false positives, i.e., the algorithms that achieve a better result in the sensitivity are in advantage over the others, while the authors remark that in this field of application true negatives (i.e., "Distraction") are also important.

4.3. Computational Execution

The authors also considered the computational execution of the approaches considered in the numerical comparisons for further evaluation. Table 9 presents the performance evaluation of the methods. Experiments were done on a Workstation equipped with Intel Xeon CPU E5-2683 v4 at 2.10 GHz and NVIDIA Tesla K40. If the images can be transferred to the Workstation quickly enough, the Viola–Jones is confirmed to be quickest to process the frames from the robot in a reasonable time, while the R–CNN and MTCNN can achieve decent performance only if accelerated with a GPU.

Table 9. Evaluation of the execution time and maximum frame per seconds (fps) at 640 × 480.

Method	Ratio vs. VJ	Max fps
VJ – Naïve	100%	14.70
VJ + HOG + K–NNs	128%	11.48
MTCNN + K–NN	140%	10.50
MTCNN + HOGs + K–NN	150%	9.80
R–CNN	483%	3.05
R–CNN + HOGs + K–NN	505%	2.92

Note that MTCNN and R-CNN were accelerated using a GPU.

Note that, for the purpose of this application, even the lowest rate of 2 frames per second can be enough to estimate the behavior of the children, which changes in the span of seconds.

5. Discussion and Conclusions

Robot-assisted therapy is a promising field of application for intelligent social robots. However, most of the studies in the literature focus on ASD individuals without ID or neglected to analyze comorbidity. Indeed, very little has been done in this area and it could be considered as one of the current gaps between the scientific research and the clinical application [7,64]. Regarding the clinical context of ASD with ID, the aim is to use social assistive robots to provide assistance to the therapist and, consequently, reduce the workload by allowing the robot to take over some parts of the intervention. This includes monitoring and recording the child's activities, proactively engaging the child when he/she is distracted, and adapting the robot behaviors according to the levels of intervention for every child on an individual basis [18].

To this end, computational intelligence techniques should be utilized to increase the robot capabilities to favor greater adaptability and flexibility that can allow the robot to be integrated into any therapeutic setting according to the specific needs of the therapist and the individual child.

This article describes a step forward in this direction, indeed the authors tackled the problem of estimating the child's visual focus of attention from the robot camera's low-resolution video recordings. To investigate the applicability in a clinical setting, the authors created a database of annotated video recordings from a clinical experiment and compared some computer vision approaches, including popular deep neural network architectures for face detection combined with HOG feature extraction and a K–NN classifier.

The results show that the approaches based on CNN can significantly overcome the benchmark algorithm only if the detection is corrected using HOG features and a classifier to adjust the attention status estimation. Overall, the approach that achieves the best result is the one that makes use of the MTCNN to identify faces. The MTCNN achieved the highest accuracy on the test set, 88.2%, which is very good, and it could be used to adapt the robot behavior to the child's current attention status, even though the computational requirements of CNN demand a proper workstation to be attached and to control the robot. However, thanks to the introduction of ultralow-power processors as accelerators for these architectures [65], the authors hypothesize that these could empower the robots and allow them to perform reliably and in real-time vision tasks like estimating the attention focus and use this information to personalize the interaction. Despite the good result, a deeper analysis of the result

with each child shows some cases in which the performance was poor, and this suggests the need to always perform a careful review by experts if the automated attention estimation is used for diagnosis. It should be noted that the researchers that analyzed the videos did not initially give the same label to some frames, and they had to discuss to come to an agreement for creating the final benchmark. These reasons, along with the low-resolution, might explain the not so remarkably high performance of some computational intelligence approaches.

However, it should be noted that the visual focus of attention is only one of the components of the social attention and the human labels were influenced also by other factors, whether the child's posture, behavior, and actions were coherent with the task, the robot, and the environment for example. Moreover, this study is considering an estimation of the social attention from only one of its components—focusing the visual attention on the interaction partner.

Future work should focus on refining algorithms and, moreover, increase the hardware support for them. Other cues, such as the adherence of the child's behavior to the robot's prompt in the case of social attention evaluation should be evaluated and considered to refine the classification.

Author Contributions: Conceptualization, A.D.N.; Methodology, A.D.N.; Software, A.D.N.; Validation, A.D.N., D.C., G.T., S.B. and S.D.N.; Formal Analysis, A.D.N.; Investigation, D.C. and G.T., Resources, G.T. and S.B.; Data Curation, D.C.; Writing-Original Draft Preparation, A.D.N. and D.C.; Writing-Review & Editing, A.D.N., D.C., G.T., S.B. and S.D.N.; Visualization, A.D.N. and D.C.; Supervision, A.D.N., S.B. and S.D.N.; Project Administration, A.D.N., S.B., S.D.N.; Funding Acquisition, A.D.N.

Funding: This work has been supported by the European Union under the Horizon 2020 Grant n. 703489 (CARER-AID) and the UK EPSRC with the grant EP/P030033/1 (NUMBERS).

Acknowledgments: The authors gratefully thank all children, parents, and educators D. Maccarrone, G. Artimagnella, and S. Nigro. The authors are grateful to the NVIDIA Corporation for the donation of a Tesla K40 that has been used for speeding up the computation of the deep learning architectures.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix Detailed Results per Child

	Viola-Jones						Viola-Jones & HOGs + KNN				
	Accuracy						Accuracy				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.745	0.775	0.957	0.429	0.144	**C1**	0.822	0.841	0.969	0.519	0.100
C2	0.738	0.735	0.888	0.493	0.092	**C2**	0.736	0.737	0.890	0.486	0.094
C3	0.800	0.806	0.934	0.510	0.093	**C3**	0.810	0.816	0.934	0.510	0.093
C4	0.764	0.780	0.913	0.308	0.123	**C4**	0.760	0.784	0.913	0.303	0.122
C5	0.610	0.624	0.890	0.212	0.161	**C5**	0.613	0.634	0.890	0.212	0.161
C6	0.749	0.752	0.901	0.628	0.079	**C6**	0.782	0.776	0.908	0.672	0.066
	Precision						**Precision**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.247	0.229	0.517	0.082	0.145	**C1**	0.307	0.278	0.519	0.119	0.154
C2	0.865	0.880	1.000	0.597	0.106	**C2**	0.856	0.891	1.000	0.375	0.134
C3	0.783	0.835	0.983	0.333	0.170	**C3**	0.809	0.855	0.983	0.400	0.153
C4	0.851	0.868	1.000	0.618	0.088	**C4**	0.856	0.864	1.000	0.606	0.090
C5	0.907	0.930	1.000	0.606	0.085	**C5**	0.913	0.931	1.000	0.630	0.079
C6	0.585	0.567	0.850	0.208	0.165	**C6**	0.631	0.631	0.882	0.211	0.177
	Sensitivity						**Sensitivity**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.648	0.708	1.000	0.313	0.219	**C1**	0.538	0.556	0.667	0.313	0.131
C2	0.686	0.746	0.965	0.223	0.191	**C2**	0.656	0.718	0.953	0.048	0.215
C3	0.705	0.797	1.000	0.029	0.264	**C3**	0.700	0.789	1.000	0.029	0.264
C4	0.732	0.810	0.962	0.115	0.233	**C4**	0.722	0.796	0.962	0.109	0.232
C5	0.550	0.586	0.952	0.009	0.249	**C5**	0.563	0.584	0.952	0.010	0.234
C6	0.746	0.783	1.000	0.188	0.200	**C6**	0.730	0.737	1.000	0.167	0.202

	Specificity						Specificity				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.743	0.794	0.965	0.297	0.181	**C1**	0.836	0.869	0.978	0.412	0.131
C2	0.819	0.842	1.000	0.206	0.173	**C2**	0.825	0.849	1.000	0.206	0.165
C3	0.855	0.861	1.000	0.546	0.119	**C3**	0.883	0.895	1.000	0.600	0.094
C4	0.723	0.783	1.000	0.361	0.197	**C4**	0.731	0.790	1.000	0.367	0.195
C5	0.771	0.800	1.000	0.172	0.197	**C5**	0.793	0.810	1.000	0.241	0.181
C6	0.743	0.779	1.000	0.382	0.155	**C6**	0.802	0.823	0.983	0.515	0.112
	Negative Value						**Negative Value**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.915	0.941	1.000	0.568	0.107	**C1**	0.946	0.963	1.000	0.539	0.084
C2	0.622	0.620	0.903	0.371	0.147	**C2**	0.613	0.603	0.886	0.371	0.141
C3	0.772	0.740	1.000	0.514	0.125	**C3**	0.774	0.759	1.000	0.514	0.123
C4	0.630	0.650	0.920	0.233	0.176	**C4**	0.622	0.633	0.910	0.238	0.178
C5	0.379	0.352	0.813	0.078	0.176	**C5**	0.417	0.388	0.813	0.157	0.168
C6	0.860	0.877	1.000	0.649	0.104	**C6**	0.864	0.882	1.000	0.644	0.099

	MTCNN—Landmarks						**MTCNN & HOGs + KNN**				
	Accuracy						**Accuracy**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.558	0.525	0.816	0.203	0.165	**C1**	0.909	0.923	0.980	0.699	0.056
C2	0.738	0.784	0.941	0.167	0.174	**C2**	0.881	0.876	0.979	0.782	0.042
C3	0.691	0.760	0.961	0.107	0.218	**C3**	0.901	0.917	0.984	0.565	0.081
C4	0.763	0.839	0.957	0.313	0.169	**C4**	0.869	0.866	0.969	0.671	0.059
C5	0.852	0.898	0.966	0.209	0.145	**C5**	0.854	0.887	0.957	0.515	0.108
C6	0.654	0.672	0.924	0.132	0.174	**C6**	0.875	0.887	0.962	0.779	0.052
	Precision						**Precision**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.281	0.226	0.747	0.095	0.200	**C1**	0.589	0.660	0.778	0.238	0.185
C2	0.711	0.729	0.951	0.163	0.178	**C2**	0.901	0.909	0.990	0.730	0.062
C3	0.628	0.658	0.960	0.068	0.240	**C3**	0.902	0.925	0.989	0.700	0.079
C4	0.748	0.837	0.971	0.292	0.190	**C4**	0.896	0.900	0.984	0.672	0.069
C5	0.849	0.898	0.982	0.191	0.152	**C5**	0.952	0.953	1.000	0.867	0.034
C6	0.494	0.500	0.905	0.115	0.209	**C6**	0.759	0.781	0.952	0.276	0.159
	Sensitivity						**Sensitivity**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.899	0.909	1.000	0.750	0.076	**C1**	0.705	0.697	0.917	0.500	0.109
C2	0.980	0.983	1.000	0.917	0.017	**C2**	0.879	0.906	0.986	0.458	0.103
C3	0.965	0.973	1.000	0.882	0.032	**C3**	0.828	0.902	1.000	0.127	0.228
C4	0.960	0.959	1.000	0.868	0.027	**C4**	0.876	0.903	0.991	0.560	0.108
C5	0.982	0.985	1.000	0.931	0.016	**C5**	0.849	0.895	0.990	0.376	0.147
C6	0.935	0.953	1.000	0.714	0.058	**C6**	0.842	0.878	0.986	0.417	0.130
	Specificity						**Specificity**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.530	0.511	0.800	0.182	0.169	**C1**	0.928	0.941	0.986	0.706	0.057
C2	0.388	0.379	0.791	0.009	0.214	**C2**	0.843	0.854	0.986	0.507	0.094
C3	0.490	0.437	0.919	0.037	0.254	**C3**	0.933	0.954	0.988	0.800	0.048
C4	0.446	0.469	0.852	0.049	0.217	**C4**	0.798	0.811	0.988	0.550	0.116
C5	0.518	0.567	0.846	0.027	0.229	**C5**	0.857	0.870	1.000	0.655	0.092
C6	0.530	0.532	0.885	0.025	0.208	**C6**	0.881	0.898	0.989	0.750	0.067

	Negative Value						Negative Value				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.980	0.987	1.000	0.875	0.030	**C1**	0.617	0.649	0.800	0.323	0.141
C2	0.885	0.923	1.000	0.333	0.134	**C2**	0.886	0.894	0.987	0.595	0.074
C3	0.931	0.941	1.000	0.800	0.054	**C3**	0.847	0.915	0.976	0.220	0.194
C4	0.826	0.833	1.000	0.556	0.107	**C4**	0.882	0.905	0.982	0.676	0.080
C5	0.869	0.900	1.000	0.500	0.126	**C5**	0.890	0.925	0.970	0.528	0.104
C6	0.922	0.948	1.000	0.500	0.103	**C6**	0.786	0.822	0.965	0.372	0.128

	RCNN						**RCNN—HOGs + KNN**				
	Accuracy						**Accuracy**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.608	0.590	0.899	0.206	0.180	**C1**	0.842	0.836	0.932	0.692	0.064
C2	0.715	0.734	0.890	0.446	0.110	**C2**	0.842	0.851	0.938	0.713	0.061
C3	0.733	0.725	0.936	0.524	0.094	**C3**	0.874	0.872	0.939	0.730	0.043
C4	0.696	0.725	0.898	0.366	0.128	**C4**	0.844	0.849	0.936	0.736	0.053
C5	0.737	0.768	0.911	0.362	0.116	**C5**	0.868	0.881	0.928	0.725	0.053
C6	0.687	0.674	0.956	0.436	0.108	**C6**	0.805	0.813	0.934	0.652	0.057
	Precision						**Precision**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.229	0.196	0.518	0.087	0.139	**C1**	0.431	0.378	0.707	0.184	0.188
C2	0.804	0.792	1.000	0.438	0.140	**C2**	0.848	0.861	0.966	0.608	0.095
C3	0.720	0.742	0.929	0.416	0.136	**C3**	0.810	0.833	0.959	0.467	0.127
C4	0.822	0.833	1.000	0.520	0.109	**C4**	0.846	0.866	0.981	0.607	0.097
C5	0.850	0.882	1.000	0.438	0.114	**C5**	0.906	0.919	0.995	0.606	0.069
C6	0.550	0.539	0.889	0.191	0.191	**C6**	0.631	0.636	0.837	0.214	0.157
	Sensitivity						**Sensitivity**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.885	0.909	1.000	0.571	0.146	**C1**	0.703	0.743	0.900	0.250	0.160
C2	0.756	0.801	1.000	0.356	0.203	**C2**	0.871	0.891	0.977	0.458	0.102
C3	0.700	0.720	1.000	0.136	0.228	**C3**	0.887	0.926	1.000	0.600	0.084
C4	0.660	0.727	1.000	0.219	0.234	**C4**	0.907	0.918	0.981	0.683	0.058
C5	0.783	0.836	0.990	0.339	0.167	**C5**	0.902	0.925	0.990	0.730	0.067
C6	0.771	0.857	1.000	0.083	0.230	**C6**	0.878	0.897	0.979	0.458	0.103
	Specificity						**Specificity**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.574	0.582	0.904	0.154	0.223	**C1**	0.853	0.849	0.966	0.750	0.068
C2	0.644	0.661	1.000	0.096	0.276	**C2**	0.752	0.744	0.973	0.397	0.124
C3	0.714	0.741	0.991	0.242	0.193	**C3**	0.837	0.837	0.962	0.686	0.065
C4	0.693	0.767	1.000	0.133	0.244	**C4**	0.691	0.660	0.943	0.482	0.118
C5	0.577	0.592	0.929	0.138	0.224	**C5**	0.707	0.735	0.909	0.364	0.146
C6	0.652	0.700	0.987	0.098	0.228	**C6**	0.764	0.763	0.958	0.603	0.088
	Negative Value						**Negative Value**				
	avg	**median**	**max**	**min**	***SD***		**avg**	**median**	**max**	**min**	***SD***
C1	0.864	0.911	1.000	0.417	0.160	**C1**	0.503	0.484	0.708	0.261	0.154
C2	0.563	0.569	1.000	0.118	0.243	**C2**	0.855	0.882	0.962	0.550	0.085
C3	0.600	0.557	0.975	0.319	0.199	**C3**	0.844	0.883	0.946	0.571	0.103
C4	0.408	0.404	0.812	0.124	0.189	**C4**	0.872	0.891	0.959	0.689	0.066
C5	0.325	0.293	0.858	0.036	0.187	**C5**	0.903	0.926	0.961	0.678	0.060
C6	0.740	0.744	0.985	0.387	0.156	**C6**	0.718	0.737	0.883	0.343	0.127

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*; American Psychiatric Association: Washington, DC, USA, 2013.
2. Adams, A.; Robinson, P. An android head for social-emotional intervention for children with autism spectrum conditions. In *Affective Computing and Intelligent Interaction*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 183–190.
3. Rabbitt, S.M.; Kazdin, A.E.; Scassellati, B. Integrating Socially Assistive Robotics into Mental Healthcare Interventions: Applications and Recommendations for Expanded Use. *Clin. Psychol. Rev.* **2014**, *35*, 35–46. [CrossRef] [PubMed]
4. Underwood, L.; McCarthy, J.; Tsakanikos, E. Mental health of adults with autism spectrum disorders and intellectual disability. *Curr. Opin. Psychiatry* **2010**, *23*, 421–426. [CrossRef] [PubMed]
5. Mesibov, G.B.; Shea, V.; Schopler, E. *The TEACCH Approach to Autism Spectrum Disorders*; Springer Science & Business Media: Berlin, Germany, 2004.
6. Pelphrey, K.A.; Morris, J.P.; McCarthy, G. Neural basis of eye gaze processing deficits in autism. *Brain* **2005**, *128*, 1038–1048. [CrossRef] [PubMed]
7. Conti, D.; Di Nuovo, S.; Buono, S.; Di Nuovo, A. Robots in education and care of children with developmental disabilities: A study on acceptance by experienced and future professionals. *Int. J. Soc. Robot.* **2017**, *9*, 51–62. [CrossRef]
8. Kennedy, J.; Baxter, P.; Belpaeme, T. Head pose estimation is an inadequate replacement for eye gaze in child-robot interaction. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts, Portland, OR, USA, 2–5 March 2015; pp. 35–36.
9. Lemaignan, S.; Garcia, F.; Jacq, A.; Dillenbourg, P. From real-time attention assessment to "with-me-ness" in human-robot interaction. In Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; pp. 157–164.
10. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
11. Deng, L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, 1–29. [CrossRef]
12. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
13. Conti, D.; Di Nuovo, A.; Trubia, G.; Buono, S.; Di Nuovo, S. Adapting Robot-Assisted Therapy of Children with Autism and Different Levels of Intellectual Disability: A Preliminary Study. In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, 5–8 March 2018; pp. 91–92.
14. Conti, D.; Trubia, G.; Buono, S.; Di Nuovo, S.; Di Nuovo, A. Evaluation of a Robot-Assisted Therapy for Children with Autism and Intellectual Disability. In Proceedings of the Towards Autonomous Robotic Systems (TAROS) Conference 2018, Bristol, UK, 25–27 July 2018; pp. 1–11.
15. Feil-Seifer, D.; Mataric, M.J. Defining Socially Assistive Robotics. In Proceedings of the 9th International Conference on Rehabilitation Robotics, Chicago, IL, USA, 28 June–1 July 2005; pp. 465–468.
16. Simut, R.E.; Vanderfaeillie, J.; Peca, A.; Van de Perre, G.; Vanderborght, B. Children with Autism Spectrum Disorders Make a Fruit Salad with Probo, the Social Robot: An Interaction Study. *J. Autism Dev. Disord.* **2016**, *46*, 113–126. [CrossRef] [PubMed]
17. Robins, B.; Dautenhahn, K.; Ferrari, E.; Kronreif, G.; Prazak-Aram, B.; Marti, P.; Iacono, I.; Gelderblom, G.J.; Bernd, T.; Caprino, F.; et al. Scenarios of robot-assisted play for children with cognitive and physical disabilities. *Interact. Stud.* **2012**, *13*, 189–234. [CrossRef]
18. Esteban, P.G.; Baxter, P.; Belpaeme, T.; Billing, E.; Cai, H.; Cao, H.L.; Coeckelbergh, M.; Costescu, C.; David, D.; De Beir, A.; et al. How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn J. Behav. Robot.* **2017**, *8*, 18–38. [CrossRef]
19. Duquette, A.; Michaud, F.; Mercier, H. Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. *Auton. Robots* **2008**, *24*, 147–157. [CrossRef]
20. Alemi, M.; Meghdari, A.; Basiri, N.M.; Taheri, A. The effect of applying humanoid robots as teacher assistants to help iranian autistic pupils learn english as a foreign language. In Proceedings of the International Conference on Social Robotics, Paris, France, 26–30 October 2015; pp. 1–10.

21. Kozima, H.; Nakagawa, C.; Yasuda, Y. Interactive robots for communication-care: A case-study in autism therapy. In Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication, Nashville, TN, USA, 13–15 August 2005; pp. 365–370.
22. Conti, D.; Di Nuovo, S.; Trubia, G.; Buono, S.; Di Nuovo, A. Use of Robotics to Stimulate Imitation in Children with Autism Spectrum Disorder: A Pilot Study in a Clinical Setting. In Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication, Kobe, Japan, 31 August–4 September 2015; pp. 1–6.
23. Brooks, D.A.; Howard, A.M. Quantifying upper-arm rehabilitation metrics for children through interaction with a humanoid robot. *Appl. Bionics Biomech.* **2012**, *9*, 157–172. [CrossRef]
24. Williams, J.H.G.; Whiten, A.; Singh, T. A systematic review of action imitation in autistic spectrum disorder. *J. Autism Dev. Disord.* **2004**, *34*, 285–299. [CrossRef] [PubMed]
25. Wainer, J.; Dautenhahn, K.; Robins, B.; Amirabdollahian, F. A Pilot Study with a Novel Setup for Collaborative Play of the Humanoid Robot KASPAR with Children with Autism. *Int. J. Soc. Robot.* **2013**, *6*, 45–65. [CrossRef]
26. Land, M.F. Vision, eye movements, and natural behavior. *Vis. Neurosci.* **2009**, *26*, 51–62. [CrossRef] [PubMed]
27. Das, D.; Rashed, M.G.; Kobayashi, Y.; Kuno, Y. Supporting Human-Robot Interaction Based on the Level of Visual Focus of Attention. IEEE Trans. *Hum. Mach. Syst.* **2015**, *45*, 664–675. [CrossRef]
28. Salley, B.; Colombo, J. Conceptualizing Social Attention in Developmental Research. *Soc. Dev.* **2016**, *25*, 687–703. [CrossRef] [PubMed]
29. Baron-Cohen, S.; Wheelwright, S.; Hill, J.; Raste, Y.; Plumb, I. The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry Allied Discip.* **2001**, *42*, 241–251. [CrossRef]
30. Charman, T. Why is joint attention a pivotal skill in autism? *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **2003**, *358*, 315–324. [CrossRef] [PubMed]
31. Pan, Y.; Ge, S.S.; He, H.; Chen, L. Real-time face detection for human robot interaction. In Proceedings of the RO-MAN 2009—The 18th IEEE International Symposium on Robot and Human Interactive Communication, Toyama, Japan, 27 September–2 October 2009; pp. 1016–1021.
32. Masala, G.L.; Grosso, E. Real time detection of driver attention: Emerging solutions based on robust iconic classifiers and dictionary of poses. *Transp. Res. Part C Emerg. Technol.* **2014**, *49*, 32–42. [CrossRef]
33. Di Nuovo, A.G.; Cannavo, R.B.; Di Nuovo, S. An agent-based infrastructure for monitoring aviation pilot's situation awareness. In Proceedings of the IEEE Symposium on Intelligent Agents (IA), Paris, France, 11–15 April 2011; pp. 1–7.
34. Lan, X.; Xiong, Z.; Zhang, W.; Li, S.; Chang, H.; Zeng, W. A super-fast online face tracking system for video surveillance. In Proceedings of the 2016 IEEE International Symposium on Circuits and Systems (ISCAS), Montreal, QC, Canada, 22–25 May 2016; pp. 1998–2001.
35. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. *A Discriminative Feature Learning Approach for Deep Face Recognition*; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Computer Vision—ECCV 2016; Springer International Publishing: Cham, Switzerland, 2016; pp. 499–515.
36. Pantic, M.; Rothkrantz, L.Ü.M. Automatic analysis of facial expressions: The state of the art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1424–1445. [CrossRef]
37. Ge, S.S.; Samani, H.A.; Ong, Y.H.J.; Hang, C.C. Active affective facial analysis for human-robot interaction. In Proceedings of the RO-MAN 2008—The 17th IEEE International Symposium on Robot and Human Interactive Communication, Munich, Germany, 1–3 August 2008; pp. 83–88.
38. Zafeiriou, S.; Zhang, C.; Zhang, Z. A survey on face detection in the wild: Past, present and future. *Comput. Vis. Image Underst.* **2015**, *138*, 1–24. [CrossRef]
39. Hsu, R.-L.; Abdel-Mottaleb, M.; Jain, A.K. Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 696–706.
40. Wang, N.; Gao, X.; Tao, D.; Yang, H.; Li, X. Facial feature point detection: A comprehensive survey. *Neurocomputing* **2018**, *275*, 50–65. [CrossRef]
41. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
42. Smith, P.; Shah, M.; da Vitoria Lobo, N. Determining driver visual attention with one camera. *IEEE Trans. Intell. Transp. Syst.* **2003**, *4*, 205–218. [CrossRef]

43. Vatahska, T.; Bennewitz, M.; Behnke, S. Feature-based head pose estimation from images. In Proceedings of the 2007 7th IEEE-RAS International Conference on Humanoid Robots, Pittsburgh, PA, USA, 29 November–1 December 2007; pp. 330–335.
44. Stiefelhagen, R. Tracking focus of attention in meetings. In Proceedings of the 4th IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, 14–16 October 2002; p. 273.
45. Senaratne, R.; Jap, B.; Lal, S.; Hsu, A.; Halgamuge, S.; Fischer, P. Comparing two video-based techniques for driver fatigue detection: Classification versus optical flow approach. *Mach. Vis. Appl.* **2011**, *22*, 597–618. [CrossRef]
46. Attamimi, M.; Miyata, M.; Yamada, T.; Omori, T.; Hida, R. Attention Estimation for Child-Robot Interaction. In Proceedings of the Fourth International Conference on Human Agent Interaction, Biopolis, Singapore, 4–7 October 2016; pp. 267–271.
47. Anzalone, S.M.; Boucenna, S.; Ivaldi, S.; Chetouani, M. Evaluating the Engagement with Social Robots. *Int. J. Soc. Robot.* **2015**, *7*, 465–478. [CrossRef]
48. Boccanfuso, L.; O'Kane, J.M. CHARLIE: An Adaptive Robot Design with Hand and Face Tracking for Use in Autism Therapy. *Int. J. Soc. Robot.* **2011**, *3*, 337–347. [CrossRef]
49. Su, H.; Dickstein-Fischer, L.; Harrington, K.; Fu, Q.; Lu, W.; Huang, H.; Cole, G.; Fischer, G.S. Cable-driven elastic parallel humanoid head with face tracking for Autism Spectrum Disorder interventions. In Proceedings of the 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, 31 August–4 September 2010; pp. 467–470.
50. Boccanfuso, L.; Scarborough, S.; Abramson, R.K.; Hall, A.V.; Wright, H.H.; O'Kane, J.M. A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: Field trials and lessons learned. *Auton. Robots* **2017**, *41*, 637–655. [CrossRef]
51. Thill, S.; Pop, C.A.; Belpaeme, T.; Ziemke, T.; Vanderborght, B. Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn J. Behav. Robot.* **2012**, *3*, 209–217. [CrossRef]
52. Conti, D.; Di Nuovo, A.; Cirasa, C.; Di Nuovo, S. A Comparison of Kindergarten Storytelling by Human and Humanoid Robot with Different Social Behavior. In Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction—HRI '17, Vienna, Austria, 6–9 March 2017; pp. 97–98.
53. Sundberg, M.L. *Verbal Behavior Milestones Assessment and Placement Program: The VB-MAPP*; Avb Press: Concord, CA, USA, 2008.
54. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. I-511–I-518.
55. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
56. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 142–158. [CrossRef] [PubMed]
57. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, 2014.
58. Bengio, Y. *Learning Deep Architectures for AI*; Now Publishers Inc.: Breda, The Netherlands, 2009.
59. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–26 June 2005; pp. 886–893.
60. Fix, E.; Hodges, J.L. Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Int. Stat. Rev. Int. Stat.* **1989**, *57*, 238–247. [CrossRef]
61. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
62. Snoek, J.; Larochelle, H.; Adams, R.P. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*; MIT Press Ltd.: Cambridge, MA, USA, 2012; pp. 2951–2959.
63. Ferri, C.; Hernández-Orallo, J.; Flach, P.A. A coherent interpretation of AUC as a measure of aggregated classification performance. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), Bellevue, WA, USA, 28 June–2 July 2011; pp. 657–664.

64. Diehl, J.J.; Schmitt, L.M.; Villano, M.; Crowell, C.R. The Clinical Use of Robots for Individuals with Autism Spectrum Disorders: A Critical Review. *Res. Autism Spectr. Disord.* **2012**, *6*, 249–262. [CrossRef] [PubMed]
65. Ionica, M.H.; Gregg, D. The Movidius Myriad Architecture's Potential for Scientific Computing. *IEEE Micro* **2015**, *35*, 6–14. [CrossRef]