Journal of Intelligent & Robotic Systems (2019) 96:267–281
https://doi.org/10.1007/s10846-018-00975-y

Robot-Assisted Autism Spectrum Disorder Diagnostic Based on Artificial Reasoning

Andrés A. Ramírez-Duque[1] · **Anselmo Frizera-Neto**[1] · **Teodiano Freire Bastos**[1]

Received: 25 April 2018 / Accepted: 20 December 2018 / Published online: 29 March 2019

Abstract

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that affects people from birth, whose symptoms are found in the early developmental period. The ASD diagnosis is usually performed through several sessions of behavioral observation, exhaustive screening, and manual coding behavior. The early detection of ASD signs in naturalistic behavioral observation may be improved through Child-Robot Interaction (CRI) and technological-based tools for automated behavior assessment. Robot-assisted tools using CRI theories have been of interest in intervention for children with Autism Spectrum Disorder (CwASD), elucidating faster and more significant gains from the diagnosis and therapeutic intervention when compared to classical methods. Additionally, using computer vision to analyze child's behaviors and automated video coding to summarize the responses would help clinicians to reduce the delay of ASD diagnosis. In this article, a CRI to enhance the traditional tools for ASD diagnosis is proposed. The system relies on computer vision and an unstructured and scalable network of RGBD sensors built upon Robot Operating System (ROS) and machine learning algorithms for automated face analysis. Also, a proof of concept is presented, with participation of three typically developing (TD) children and three children in risk of suffering from ASD.

Keywords Child-Robot interaction · Autism spectrum disorder · Convolutional neural network · Robot reasoning model · Statistical shape modeling

1 Introduction

Research in Child-Robot Interaction (CRI) aims to provide the necessary conditions for the interaction between a child and a robotic device taking into account some fundamental features, such as child's neurophysical and physical condition, and the child's mental health [1]. That is how Robot-Assisted Therapies (RAT) using CRI theories have been of interest as an intervention for CwASD, elucidating faster and more significant gains from the therapeutic intervention when compared to traditional therapies [2–4].

ASD is a neurodevelopmental disorder that affects people from birth, and its symptoms are found in the early developmental period. Individuals suffering from ASD exhibit persistent deficits in social communication, social interaction and repetitive patterns of behavior, interests, or activities [5]. Some of the ASD signs may be observed before the age of 10 months, although a reliable diagnosis can only be performed at 18 months of age, according to [6], or 24 months according to [7].

The use of computer vision to analyze the child's behaviors, and automated video coding to summarize the interventions, can help the clinicians to reduce the delay of ASD diagnosis, providing the CwASD with access to early therapeutic interventions. In addition, CRI-based intervention can transform traditional diagnosis methods through a robotic device to systematically elicit child's behaviors that exhibit ASD signs [8].

Some of the first systems developed to assist ASD therapists and make diagnosis based on robotic devices have primarily been open loop and remotely operated systems. However, these approaches are unable to perform autonomous feedback to enhance the interaction [9–11].

✉ Andrés A. Ramírez-Duque
aaramirezd@gmail.com

[1] Universidade Federal do Espiríto Santo., Av. Fernando Ferrari, 514 (29075-910), Vitoria, Brazil

Nevertheless, different systems are able to modify the behavior of the robot according to environmental interactions and the child's response, using a closed-loop and artificial cognition approaches [12–16]. These systems have been hypothesized to offer technological mechanisms for supporting more flexible and potentially more naturalistic interaction [17]. In fact, literature reports that automatic robot's social behaviors modulation according to specifics scenarios has a strong effect on child's social behavior [12]. However, despite the increase of positive evidence, this technology has rarely been applied to specific ASD diagnosis.

This work aims to present a robot-assisted framework using an artificial reasoning module to assist clinicians with the ASD diagnostic process. The framework is composed of a responsive robotic platform, a flexible and scalable vision sensor network, and an automated face analysis algorithm based on machine learning models. In this research we take advantage of some neural models available as open sources projects to build a completely new pipeline algorithm for global recognition and tracking of child's face among many faces present in a typical unstructured clinical intervention, in order to estimate the child's visual focus of attention along the time. The proposed system can be used in different behavioral analysis scenarios typical of an ASD diagnostic process. In order to illustrate the feasibility of the proposed system, in this paper an experimental trial to assess joint-attention behavior is presented employing an in-clinic setup (unstructured environment).

The main contributions of this paper are: (i) the development of a new artificial reasoning module upon a flexible and scalable ROS-based vision system using state-of-the-art machine learning neural models; (ii) the proposal and implementation of a supervised CRI (child-robot interaction) based on an open source social robotic platform to enhance the traditional tools for ASD diagnosis using an in-clinic setup protocol. For the best of our knowledge, there are no open source projects available for face analysis based on a multi-camera approach using ROS with the characteristics described in our research.

2 Related Work

Recent researches have shown the acceptance and efficiency of technologies used as auxiliary tools for therapy and teaching of individuals with ASD [18–21]. Such technologies may also be useful for people surrounding ASD individuals (therapists, caregivers, family members). For example, the use of artificial vision systems to measure and analyze the child's behavior can lead to alternative screening and monitoring tools that help the clinicians to get feedback from the effectiveness of the intervention [22]. Additionally, social robots have great potential for aid in the diagnosis and therapy of children with ASD [18, 23]. A higher degree of control, prediction and simplicity may be achieved in interactions with robots, impacting directly on frustration and reducing the anxiety of these individuals [24].

Respect to the use of computer vision techniques, previous studies already analyzed child's behaviors, such as visual attention, eye gaze, eye contact, smile events, and visual exploration using cameras and eye trackers [25, 26] and RGBd cameras [27, 28]. These studies have shown the potential of vision systems in improving the behavioral coding in ASD therapies. However, these studies did not implement techniques of CRI to enhance the intervention.

On the other hand, studies about how CwASD respond to a robot mediator compared to human mediator have been reported, such as intervention scenarios with imitation games [29, 30], telling stories [9] and free play tasks [12, 31]. These works used features, such as proxemics, body gestures, visual contact and eye gaze as behavioral descriptors, whereas the behavior analysis was estimated using manual video coding.

Researchers of Vanderbilt University published a series of research showing an experimental protocol to assess joint attention (JA) tasks defined as the capacity for coordinated orientation of two people toward an object or event [6]. The protocol consisted of directing the attention of the child towards objects located in the room through adaptive prompts [32]. Bekele et al. inferred the participant's eye gaze by the head pose, which was calculated in real-time by an IR camera array [17]. In their last works, Zheng et al. and Warren et al. used a commercial eye tracker to estimated the children's eye gaze around the robot and manual behavioral coding for global evaluation [10, 33]. However, eye tracker devices require pre-calibration and may limit the movement of the individual. The results of these works showed that the robot attracted children's attention and that CwASD reached all JA task. Nevertheless, developing JA tasks is more difficult with a robot than with humans [10]. Anzalone et al. developed a CRI scenario using the NAO robot to perform JA tasks, in which the authors used an RGBD camera to estimate only body and head movements. The results showed that JA performance of children with ASD was similar to the performance of TD children when interacting with the human mediator, however, with a robot mediator, the children with ASD presented a lower performance than the TD children, i.e, the children with ASD needed more social cues to finalize the task [34]. Chevalier et al. analyzed in their study, some features, such as proprioceptive and visual integration in CwASD, using an RGBD sensor to record the interventions sessions and manual behavior coding to analyzed the participants' performance [35]. In none of the previous works, a closed-loop subsystem was

implemented to provide some level of artificial cognition to enable automated robot behavior.

In contrast with the aforementioned researches, other works implemented automated face analysis and artificial cognition through robot-mediator and computer vision, which analyzed child's engagement [36, 37], emotions recognition capability [13, 15, 38] and child's intentions [14, 16]. In these works, two different strategies were implemented, where the most common is based on mono-camera approach using an external RGB or RGBd sensor [15, 36, 37] or using on-board RGB cameras mounted on the robotic-platform [13, 16]. Other strategies are based on a highly structured environment composed of an external camera plus an on-board camera [38] or a network of vision sensors attached to a small table [14]. These strategies based on multi-camera methods improve the system's performance, but remain constrained in relation to desired features, such as flexibility, scalability, and modularity. Thus, despite the potential that these techniques have shown, achieving automated child's behavior analysis in a naturalistic way into unstructured clinical-setups with robots that interact accordingly remains a challenge in CRI.

3 System Architecture Overview

The ROS system used in this work is a flexible and scalable open framework for writing modular robot-centered systems. Similar to a computing operating system, ROS manages the interface between robot hardware and software modules and provides common device drivers, data structures and tool-based packages, such as visualization and debugging tools. In addition, ROS uses an interface definition language (IDL) to describe the messages sent between process or nodes, this feature facilitates the multi-language (C++, Python and Lisp) development [39].

The overall system developed here was built using a node graph architecture, taking advantages of the principal ROS design criteria. As with ROS, our system consists of a number of nodes to local video processing together a robot's behavior estimation, distributed around a number of different hosts and connected at runtime in a peer-to-peer topology. The inter-node connection is implemented as a hand-shaking and occurs in XML-RPC protocol along with a web-socket communication for robot's web-based node (/ONO_node, see Fig. 1). The node structure is flexible, scalable and can be dynamically modified, i.e., each node can be started and left running along an experimental session or resumed and connected to each other at runtime. In addition, from a general perspective, any robotic platform with web-socket communication can be integrated. The developed system is composed of two interconnected modules as shown in Fig. 1: an artificial reasoning module and a CRI-channel module. The module architectures are detailed in the following subsections.

3.1 Architecture of Reasoning Module

In this module, a distributed architecture for local video processing is implemented. The data of each RGBD sensor in the multi-camera system are processed for two nodes, in which the first is a driver level node and the second is a processing node. The driver[1] node transforms the streaming data of the RGBD sensor into the ROS messages format. The driver addresses the data through a specialized transport provided by plugings to publishes images in a compressed representations while the receptor node only sees *sensor_msgs/Image* messages. The data processing node executes the face analysis algorithm. This node uses a *image_transport* subscriber and a ROS packages called CvBridge to turn the data into a image format supported for the typical computer vision algorithms. Later, the same node publishes the head pose and eye gaze direction by means of a ROS navigation message defined as *nav_msgs/Odometry*.

An additional node hosted in the most powerful workstation carries out a data fusion of all navigation messages that were generated in the local processing stage. In addition to the fusion, this node computes the visual focus of attention (VFOA) and publishes this as a *std_msgs/Header*, in which the time stamp and the target name of the VFOA estimation are registered.

3.2 Architecture of CRI-Channel

The system proposed here has two bidirectional communication channels, a robot-device, and a web-based application to interact with both the child and the therapist. The robot device can interact with the CwASD executing different physical actions, such as facial expression, upper limb poses, and verbal communication. Thus, according to the child's performance, the reasoning module can modify the robot's behavior through automatic gaze shifting, changing the facial expression and providing sound rewards. The client-side application was developed to allow the therapist to control and register all step of the intervention protocol. This interface was also used to supervise and control the robot's behavior and to offer feedback to the therapist about the child's performance along the intervention. This App has two channels of communication for interacting with the reasoning module. The first connection uses a web-socket protocol and a RosBridge_suite package to support the interpretation of ROS messages, as well as, JSON-based commands in ROS. The second one uses a ROS module

[1]Tools for using the Kinect One (Kinect V2) in ROS, https://github.com/code-iai/iai_kinect2.

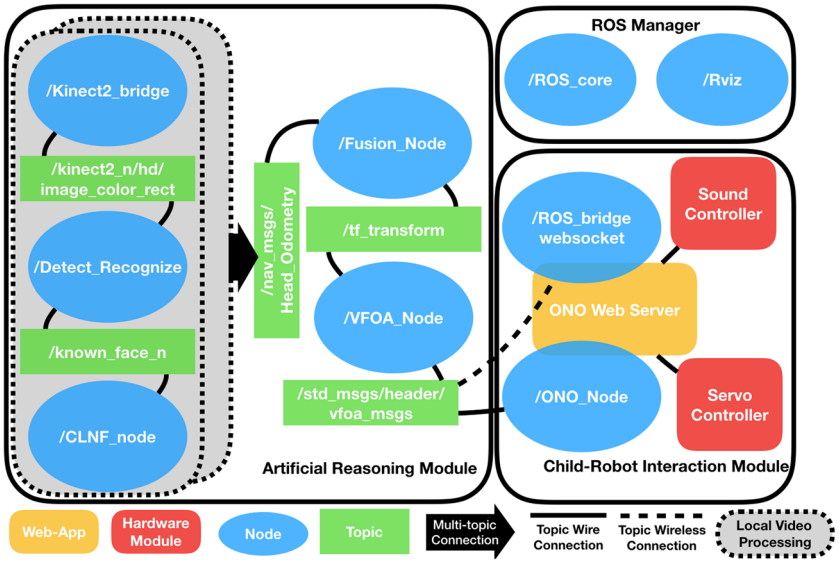

Fig. 1 Node graph architecture of the proposed ROS-based system. The system is composed of two interconnected modules, an artificial reasoning module and a CRI-channel module. The ONO web server has two way of bidirectional communication: a websocket and a standard ROS Subscriber

developed in the server-side application to directly run a ROS node and communicate with standard ROS publishers and subscribers.

4 The Robotic Platform ONO

The CRI is implemented through the open source platform for social robotics (OPSORO),[2] which is a promising and straightforward system developed for face to face communication composed of a low-cost modular robot called ONO (see Fig. 2) and web-based applications [40]. Some of the most important requirements and characteristics that make ONO interesting for this CRI strategy are explained in the following sections.

4.1 Appearance and Identity

The robot is covered in foam and also fabric to have a more inviting and huggable appearance to the children. The robot has an oversized head to make its facial expressions more prominent and to highlight the importance for communication and emotional interaction. As a consequence of its size and pose, children can interact with the robot at eye height when the robot is placed on a table.

The robot ONO has not a predefined identity, as the only element previously conceived is the name. Unlike other robots that have well-defined identities, such as Probo [9] or Kaspar [41], in this work the ONO's identity is built with the participation of the child through a co-creation process. For this reason, a neutral appearance is initially used. In the intervention, the therapist can provide the child with clothes and accessories to define the identity of ONO.

4.2 Mechanics Platform

As the initial design of ONO is composed only of the actuated face, in this work it was needed to provide the ONO with some body language. For this purpose, motorized arms were designed and implemented.

The new design of ONO has a fully face and two arms actuated, giving a total of 17 Degrees of Freedom (DOF). The ONO is able to perform facial expressions and nonverbal cues, such as waving, shake hands and pointing towards objects, moving its arms (2 DOF x 2), eyes (2 DOF x 2), eyelids (2 DOF x 2), eyebrows (1 DOF x 2), and mouth (3 DOF). The robot has also a sound module that allows explicit positive feedback as well as reinforcement learning through playing words, conversations and other sounds.

4.3 Social Expressiveness

In order to improve social interaction with a child, the ONO is able to exhibit different facial expressions. The ONO's expressiveness is based on the Facial Action Coding System (FACS) developed in [42]. Each DOF that composes the ONO's face is linked with a set of Action Units (AU) defined by the FACT, and each facial expression is determined for specific AU values. The facial expressions are represented as a 2D vector $fe = (v, a)$ in the emotion circumplex model defined by valence and arousal [9]. In this context, the basic facial expressions are specified on a unit circle, where the neutral expression corresponds to the origin of the space $fe_0 = (0, 0)$. The relation between the DOF position and the AU values is resolved through a lookup table algorithm using a predefined configuration file [40].

[2] Open Source Platform for Social Robotics (OPSORO) http://www.opsoro.com.

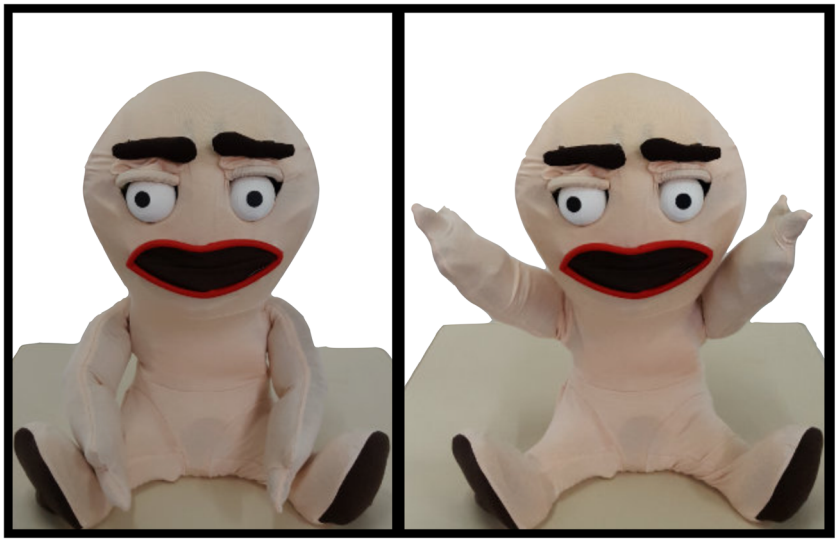

Fig. 2 ONO robot, developed through the open source platform for social robotics (OPSORO)

4.4 Adaptability and Reproducibility

The application of the Do-It-Yourself (DIY) concept is the principal feature of ONO's design, which facilitates its dissemination and use in research areas other than engineering as health care. These characteristics allow ONO building for any person without specialized engineering knowledge. Additionally, it is possible to replicate ONO without the need for high-end components or manufacturing machines [40]. The electronic system is based on a Raspberry Pi single-board computer combined with a custom OPSORO module with circuitry to control up to 32 servos, drive speakers and touch sensors. Any sensor or actuator compatible with the embedded communication protocols (UART, I2C, SPI) implemented on the Raspberry Pi can be used by this platform.

4.5 Control and Autonomy

With the information delivered for the automated reasoning module, it was possible to automate the ONO's behavior and, then, the robot can infer and interpret the children's intentions to react most accurately to the action performed by them, thus enabling a more efficient and dynamic interaction with ONO. In this work, the automated ONO's behavior is partially implemented, i.e., the framework can modify some physical actions of ONO using the feedback information about the child's behavior. The actions suitable to be modified are gaze shift toward the child in specifics events, changing from neutral to positive facial expression when the child looks toward the target, and providing sound rewards. Also, an Aliveness Behavior Module (ABM) is implemented to improve the CRI, which consist of blinking the robot's eyes and changing its arms among some predefined poses. Also, the robot can be manually operated through a remote controller hosted in the client-side application.

5 Reasoning Module: Machine Learning Methods for Child's Face Analysis

The automated child's face analysis consists of monitoring nonverbal cues, such as head and body movements, head pose, eye gaze, visual contact and visual focus of attention. In this work, a pipeline algorithm is implemented using machine learning neural models for face analysis. The chosen methods were developed using state-of-art trained neural models, available by Dlib[3] [43] and OpenFace[4] [44]. Some modification such as, turn the neural model an attribute of the ROS node class and evaluate this in each topic callback, were needed to run the neural models into a common ROS node.

The algorithm proposed for child's face analysis involves face detection, recognition, segmentation and tracking, landmarks detection and tracking, head pose, eye gaze and visual focus of attention (VFOA) estimation. In addition, the architecture proposed here also implement new methods for asynchronous matching and fusion of all local data, visual focus of attention estimation based on Hidden Markov Model (HMM) and direct connection with the CRI-channel to influence the robot's behaviors. A scheme of the pipeline algorithm is shown in Fig. 3.

[3] Dlib C++ Library http://dlib.net/.

[4] A Open Source Facial Behavior Analysis https://github.com/TadasBaltrusaitis/OpenFace.

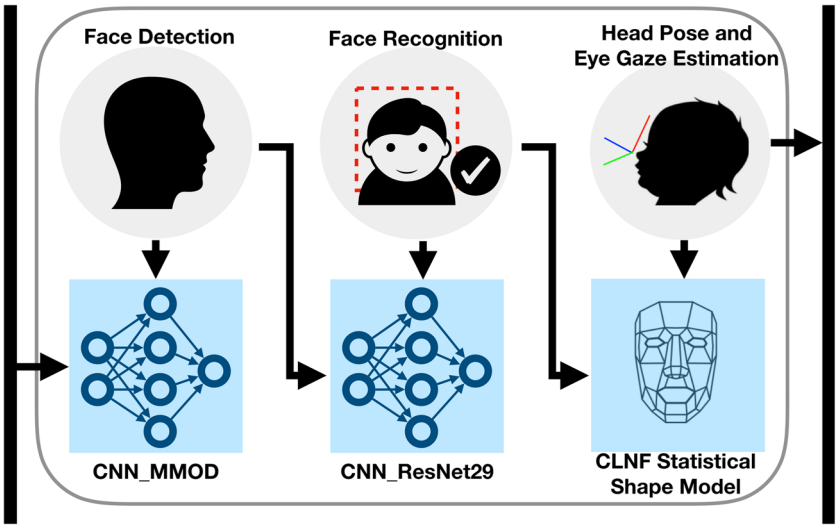

Fig. 3 Pipeline algorithm of the automated child's face analysis

5.1 Child's Face Detection and Recognition

The in-clinic setup requires differentiate the child's face from other faces detected and found in the scene. For this reason, a face recognition process was also implemented in this work. First, the face detection is executed to initialize the face recognition process and, subsequently, initialize the landmarks detection. In this work, both detection and recognition are implemented using deep learning models, which are described in this section.

In the detection process, a Convolutional Neural Network (CNN) based face detector with a Max-Margin Object Detection (MMOD) as loss layer is used [45]. The CNN consist first of a block composed of three downsampling layers, which apply convolution with a 5x5 filter size and 2×2 stride to reduce the size of the image up to eight times its original size and generate a feature map with 16 dimensions. Later, the result are processed for one more block composed of four convolutional layers to get the final output of the network. The three first layers of the last block have 5×5 filter size and 1x1 stride, but, the last layer has only 1 channel and a 9×9 filter size. The values in the last channel are large when the network thinks it has found a face at a particular location. All convolutional block above are implemented with two additional layers among convolutional layers, pointwise linear transformation, and Rectified Linear Units (RELU) to apply the non-saturating activation function $f(x) = \max(0, x)$. The training dataset used to create the model is composed of 6975 faces and is available at Dlib's homepage.[5]

The face recognition algorithm used in this work is inspired on the deep residual model from [46]. The residual network (ResNet) model developed by He *et. al* reformulates the convolutional layers to learn a residual functions $F(x) := H(x) - x$ with reference to the layer inputs x, instead of learning unreferenced functions. In the practical implementation, the previous formulation means inserting shortcut connections, which turn the network into its counterpart residual version [46]. The CNN model then transforms each face detected to a 128D vector space in which images from the same person will be close to each other, but faces from different people will be far apart. Finally, the faces are classified as child's face, caregiver's face and therapist's face.

Both detection and recognition CNN model were implemented and trained from [43] and released in Dlib 19.6.

5.2 Face Analysis, Landmarks, Head Pose and Eye Gaze

This work uses the technique for landmarks detection, head pose and eye gaze estimation developed by Baltrušaitis et al., named Conditional Local Neural Fields (CLNF) [47]. This technique is an extension of the Constrained Local Model (CLM) algorithm using specialized local detectors or patch experts. CNLF model consists of a statistical shape model, which its learned from data examples and is parametrized for m components of linear deformation to control the possible shape variations of the non-rigid objects [48]. Approaches based on CLM [49, 50] and CLNF [47] model the object appearance in a local fashion, i.e, each feature point has its own appearance model to describe the amount of misalignment.

CLNF-based landmark detection consists of three main parts: the shape model, the local detectors or patch experts, and the fitting algorithm, which are detailed below.

[5] http://dlib.net/files/data/dlib_face_detection_dataset-2016-09-30.tar.gz.

5.2.1 Shape Model

The CLNF technique uses a linear model to describe non-rigid deformations called Point Distribution Model (PDM). The PDM is used to estimate the likelihood of the shapes being in a specific class, given a set of feature points [48]. This is important for model fitting and shape recognition.

The shape of a face that has n landmark points can be described as:

$$X = [X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_n, Z_1, Z_2, \ldots, Z_n], \tag{1}$$

and the class that describes a valid instance of a face using PDM can be represented as:

$$X = \bar{X} + \Phi q, \tag{2}$$

where $\bar{X}$ is the mean shape of the face, Φ described the principal deformation modes of the shape, and q represent the non-rigid deformation parameters. Both $\bar{X}$ and Φ are learned automatically from labeled data using Principal Component Analysis (PCA). The probability density distribution of the instances into the shape class is expressed as a zero mean Gaussian with Covariance matrix $\Lambda = ([\lambda_1; \ldots; \lambda_m])$ evaluated at q:

$$p(\mathbf{q}) = \mathcal{N}(q; 0; \Lambda) = \frac{1}{\sqrt{(2\pi)^m |\Lambda|}} exp\left\{-\frac{1}{2}(q^T \Lambda^{-1} q)\right\} \tag{3}$$

Once the model is defined, it is necessary to place the 3D PDM in an image space. The following equation is used to transform between 3D space to image space using weak perspective projection [49]:

$$x_i = s \cdot R_{2D} \cdot (\bar{X}_i + \Phi_i q) + t, \tag{4}$$

where $\bar{X}_i = [\bar{x}_i, \bar{y}_i, \bar{z}_i]^T$ is the mean value of the i^{th} landmark. The instance of the face in an image is, therefore, controlled using the parameter vector $\mathbf{p} = [s, w, t, q]$, where q represents the local non-rigid deformation, s is a scaling term, w is the rotation term that controls the 2×3 matrix R_{2D}, and t is the translation term.

The global parameters are used to estimate the head pose in reference to the camera space using orthographic camera projection and solving the Perspective-n-Point (PnP) problem respect to the detected landmarks. The PDM used in [44] was trained on two public datasets [51, 52]. This result in a model with 34 non-rigid (Principal modes) and 6 rigid shape parameters.

5.2.2 Patch Experts

The patch experts scheme is the main novelty implemented in the CLNF model. The new Local Neural Field (LNF) patch expert takes advantage of the non linear relationship between pixel values and the patch response maps. The LNF captures two kinds of spatial characteristics between pixels, such as similarity and sparsity [47].

LNF patch expert can be interpreted as a three layer perceptron with a sigmoid activation function followed by a weighted sum of the hidden layers. It is also similar to the first layer of a Convolutional Neural Network [44]. The new LNF patch expert is able to learn from multiple illuminations and retain accuracy. This becomes important when creating landmark detectors and trackers that are expected to work in unseen environments and on unseen people.

The learning and inference process is developed using a gradient-based optimization method to help in finding locally optimal model parameters faster and more accurately.

In the CLNF model implemented in [44], 28 set in total of LNF patch experts were trained for seven views and four scales. The framework uses patch experts specifically trained to recognize the eyelids, iris and the pupil, in order to estimate the eye gaze [44].

5.2.3 Fitting Algorithm

For each new image or video frame, the fitting algorithm of CLNF-based landmark detection process attempts to find the value of the local and global deformable model parameters $\mathbf{p}$ that minimizes the following function [49]:

$$\mathcal{E}(\mathbf{p}) = \mathcal{R}(\mathbf{p}) \sum_{i=1}^{n} \mathcal{D}_i(x_i; \mathcal{I}), \tag{5}$$

where $\mathcal{R}$ is a weight to penalize unlikely shapes, which depends on the shape model, and $\mathcal{D}$ represents the misalignment of the i^{th} landmark in the image $\mathcal{I}$, which is function of both the parameters $\mathbf{p}$ and the patch experts. Under the probabilistic point of view, the solution of (5) is equivalent to maximize the *a posteriori* probability (MAP) of the deformable model parameters p:

$$p\left(\mathbf{p} \mid \{l_i = 1\}_{i=1}^{n}, \mathcal{I}\right) \propto p((p)) \prod_{i=1}^{n} p\left(l_i = 1 \mid x_i, \mathcal{I}\right), \tag{6}$$

where, $l_i \in \{1, -1\}$ is a discrete random variable indicating whether the i^{th} landmark is aligned or misaligned, $p(\mathbf{p})$ is the prior probability of the deformable parameters $\mathbf{p}$, and $p\left(l_i = 1 \mid x_i, \mathcal{I}\right)$ is the probability of a landmark being aligned at a particular pixel location x_i, which is quantified from the response maps created by patch. Therefore, the last term in (6) represents the joint probability of the patch expert response maps.

The MAP problem is solved using a optimization strategy designed specifically for CLNF fitting called non-uniform

regularized landmark mean shift (NU-RLMS) [47], which uses two step process. The first step evaluates each of the patch experts around the current landmark using a Gaussian Kernel Density Estimator (KDE). The second step iteratively updates the model parameters to maximize (6).

The NU-RLMS uses expectation maximization algorithm, where the E-step involves evaluating the posterior probability over the candidates, and the M-step finds the parameter updated through the mean shift vector $\mathbf{v}$. The mean shift vector points in the direction where the feature point should go, but the motion is restricted by the statistical shape model and the $\mathcal{R}(\mathbf{p})$. This interpretation leads to the new update function:

$$\underset{\Delta\mathbf{p}}{\operatorname{argmin}} \left\{ \|J\Delta\mathbf{p} - \mathbf{v}\|_W^2 + r\,\|\mathbf{p} + \Delta\mathbf{p}\|_{\tilde{\Lambda}^{-1}}^2 \right\}, \tag{7}$$

where r is a regularization term, J is the Jacobian, which describe how the landmarks location are changing based on the infinitesimal changes of the parameters $\mathbf{p}$, $\tilde{\Lambda}^{-1} = diag([0; 0; 0; 0; 0; 0; \lambda_1^{-1}; \ddots; \lambda_m^{-1}])$, and W allows for weighting of mean-shift vectors. Non-linear least squares leads to the following update rule:

$$\Delta\mathbf{p} = -\left(J^T W J + r\Lambda^{-1}\right)\left(r\Lambda^{-1}\mathbf{p} - J^T W\mathbf{v}\right). \tag{8}$$

To construct W, the performance of patch experts on training data is used.

5.3 Data Fusion

The fusion of the local results for the head pose estimation is done applying a consensus over the rotation algorithm [53]. This algorithm consists of calculating the weighted average pose between each camera estimation and its immediate sensors' estimation neighbors using the axis-angle representation. The local pose is penalized by two weights: the alignment confidence of landmarks detection procedure and the Mahalanobis distances between the head pose and a neutral pose.

5.4 Field of View (FoV) and Visual focus of Attention (VFOA)

The VFOA estimation model is implemented as a dynamic Bayesian network through a Hidden Markov Model (HMM). The model assumes a specific set of child's attention attractors or targets $\mathbb{F}$. The estimation process decodes the sequence of child's head poses $H_t = (H_t^{yaw}, H_t^{pitch}) \in \mathbb{R}^2$ in terms of VFOA states $F_t \in \mathbb{F}$ at time t [54]. The probability distribution of the head poses in reference to a given VFOA target is represented by a Gaussian distribution, whereas the transitions among these targets are represented by the transition matrix A. The HMM equations can then be written as follows:

$$P(H_t \mid F_t = f, \mu_t^h) = \mathcal{N}(H_t \mid \mu_t^h(f), \Sigma_H(f)) \tag{9}$$

$$p(F_t = f \mid F_{t-1} = \hat{f}) = A_{f\hat{f}}. \tag{10}$$

The Gaussian covariances is defined manually to reflect target sizes and head pose estimation variability. Moreover, the Gaussian means corresponding to each specific target μ_t^h is calculated through a gaze model that sets this parameter as a fixed linear combination of the target direction and the head reference direction [55]:

$$\mu_t^h(f) = \alpha \star \mu_t(f) + (1_2 - \alpha) \star R_t, \tag{11}$$

where $\star$ denotes the component wise product $1_2 = (1, 1)$, $\alpha = (\alpha^{yaw}, \alpha^{pitch}) = (0.7, 0.5)$ are adjustable constants that describe the fraction of the gaze shift that corresponds to the child's head rotation, $\mu_t \in (\mathbb{R}^2)^K$ is the directions of the given K targets, and $R_t \in \mathbb{R}^2$ represents the reference direction, which is the average head pose over a time window W^R. The above assumption describes the body orientation behavior of any child who tends to orient himself/herself towards the set of gaze targets to make more comfortable to rotate his/her head towards different targets [55].

$$R_t = \frac{1}{W^R} \sum_{i=t-W^R}^{t} H_i. \tag{12}$$

Finally, for the estimation of the VFOA sequence a classic Viterbi algorithm of HMM is implemented [54].

6 Case Study

For the case study, the vision system is composed of three Kinect V2 sensors. Each sensor is connected to a workstation equipped with a processor of Intel Core i5 family and a GeForce GTX GPU board (two workstation with GTX960 board, and one workstation with GTX580 board). All workstation are connected through a local area network synchronized using the NTP protocol.[6] The sensors were intrinsically and extrinsically calibrated through a conventional calibration process using a standard black-white chessboard.[7]

6.1 In-clinic Setup

A multidisciplinary team of psychologists, doctors and engineers developed a case study using a psychology room equipped with a unidirectional mirror to perform behavioral

[6] Network Time Protocol Homepage, http://www.ntp.org.

[7] Tools for using the Kinect One (Kinect V2) in ROS, https://github.com/code-iai/iai_kinect2.

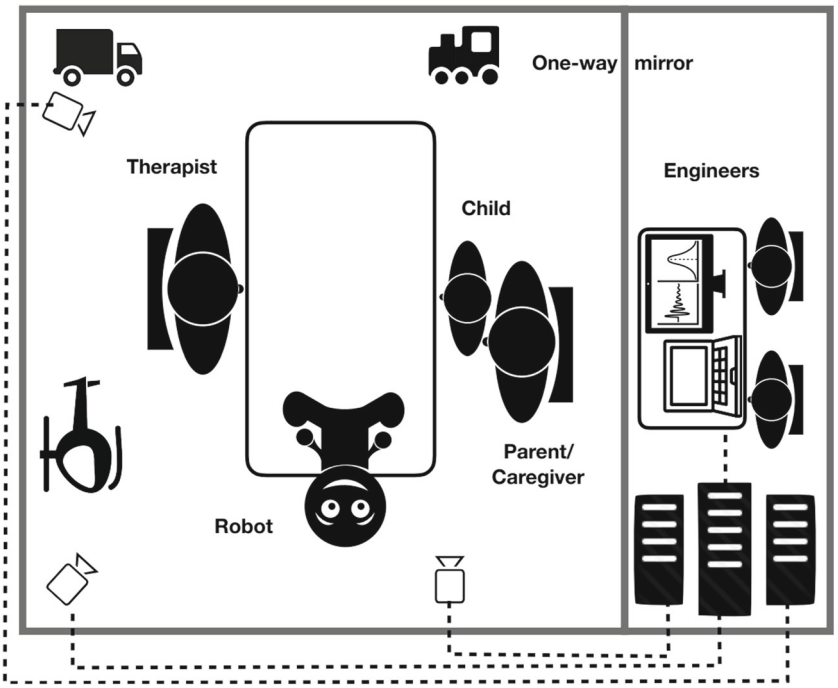

Fig. 4 Representation of the interventions room of in-clinic setup

observation appropriately. The room was prepared with a table and three chairs: one for the child, another for the caregiver and a third one for the therapist. The robot was placed on the table, and the following toys, a helicopter, a truck and a train, were attached to room's walls. The RGBD sensors were located close to the walls, and no additional camera was placed on the robot or the table, so as not to attract the child's attention. A representation of the interventions room of in-clinic setup is shown in Fig. 4.

6.2 Intervention Protocol

In this work, a technology-based system was used as a tool in various stages of the ASD diagnostic process. The framework can be implemented to extract different behavioral features to be assessed, e.g., eye contact, stereotyped movements of the head, concentration and excessive interest in objects or events. However, for the scope of this research, a specific clinical setup intervention to assess Joint Attention (JA) behaviors is presented. The intervention aims to evaluate the capacity of JA; which can be divided into three classes: initiation of joint attention (IJA), responding to joint attention bids (RJA), and initiation of request behavior (IRB) [6]. The therapist guides the intervention all the time and leverages the robot device as an alternative channel of communication with the child, for the above, both the specialist and the robot remained in the room during the intervention. The children were accompanied throughout the session by a caregiver who was oriented not to help the child in the execution of the

Fig. 5 The child's nonverbal cues elicited by the CRI, to look towards the therapist, towards the robot, point and self occlusion

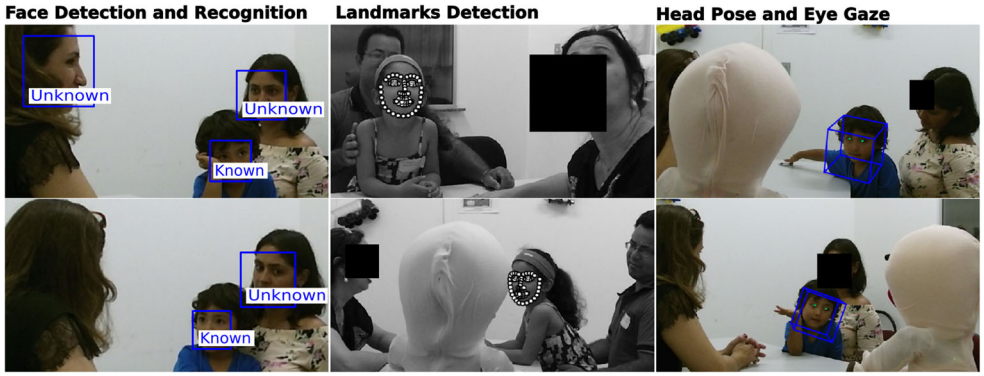

Fig. 6 Performance of the child's face analysis pipeline for the case study. Face detection and recognition, landmarks detection, head pose and eye gaze estimation were executed

tasks. The exercise developed aimed to direct the attention of the child towards objects located in the room through stimuli, such as, look at, point and speak. The stimuli were generated first only by the therapist and later just by the robot.

6.3 Subjects

Three children without confirmed ASD diagnosis, but with evidence of risk factors, and three typically developing (TD) children as the control group participated in the experiments. All volunteers participated with their parent's consent, which were five boys (3 ASD, 2 TD) and one TD girl, between 36 months to 48 months. Each volunteer participated in one single session. The goal was to analyze the based-line of the child's behavior and establish differences in the behavioral reaction between TD and ASD children for stimuli generated through CRI and leverage the novelty effect raised by the robot mediator.

7 Results and Discussion

The child's nonverbal cues elicited by the CRI can be observed in Fig. 5. Some examples of children's behavior tagged to perform the behavioral coding are shown in the six pictures. The tagged behaviors were: to look towards an object, towards the robot, and towards the therapist, to point and, to respond to a prompt of both mediators and self occlusion. Typical occlusion problem, as occlusion by hair, hands and the robot were detected.

The performance of video processing in the proof of concept session is reported in Fig. 6. In the case study sessions, the child's face detection and recognition, the

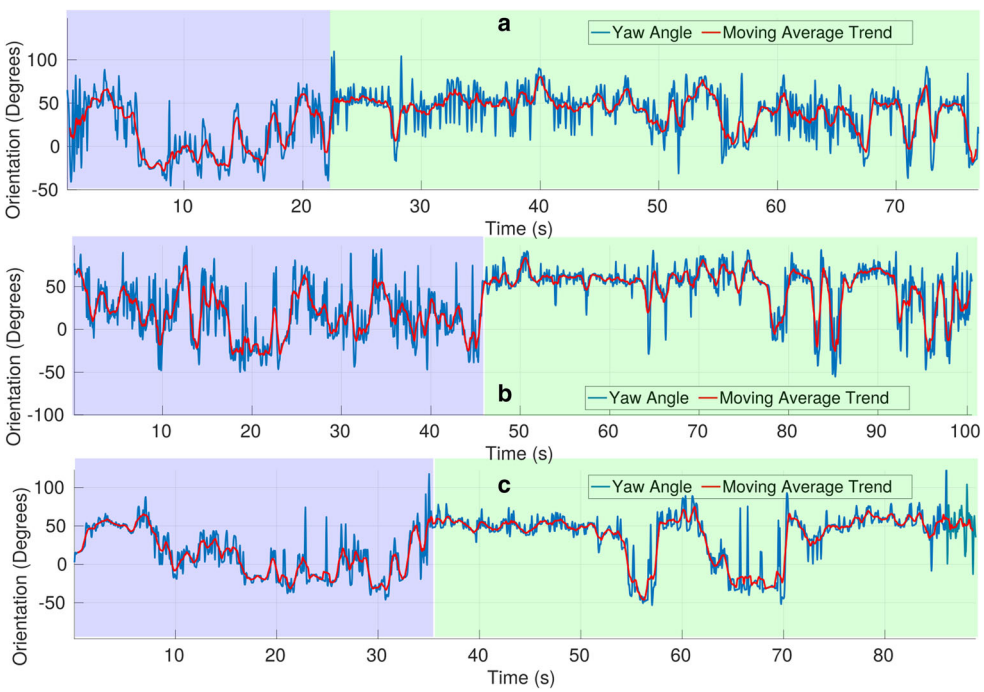

Fig. 7 Evolution over time of the child's head/neck rotation (yaw rotation) for a TD group

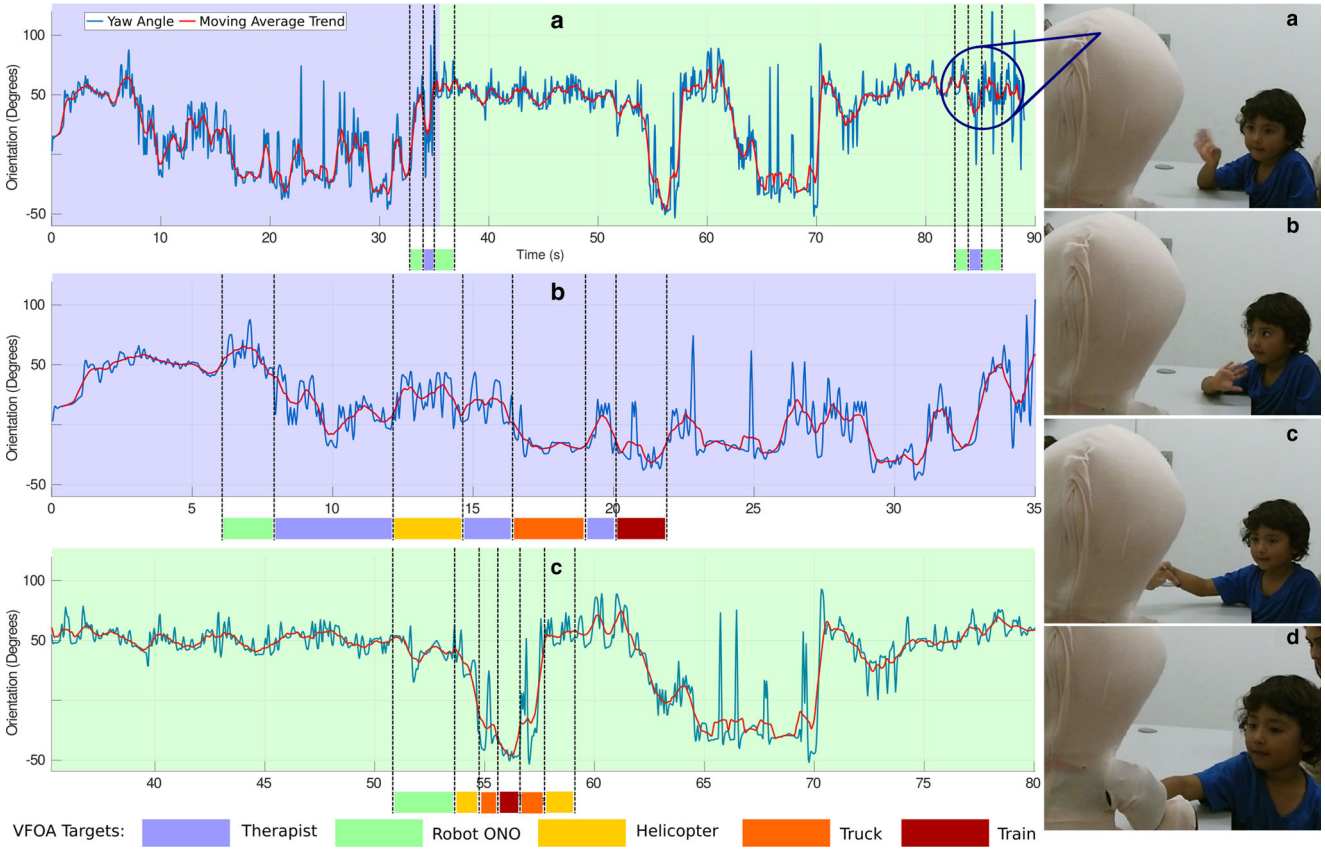

Fig. 8 Evolution over time of the child's head/neck rotation (yaw rotation) for a TD volunteer and VFOA estimation results

landmarks detection, head pose and eye gaze estimation for different viewpoints are shown in Fig. 3. The recognition process was able to detect all faces in the session successfully in most cases.

The child's head pose was captured throughout the session and analyzed automatically to estimate the evolution over time of child's head and the VFOA. Along the session, the child's neck right/left rotation movement was predominant (Yaw axis), while the neck flexion/extension (Pitch axis) and neck R/L lateral flexion movements (Roll axis) remained approximately constant. The Yaw rotation of the TD children group is reported in Fig. 7. The vertical light blue stripe indicates the intervention period with therapist-mediator, and the vertical light green stripe represents the period with robot-mediator. The continuous blue line represents the raw data recorded, and the continuous red line describes the average data trend. From the observation of the three plot, the TD children started the intervention looking towards the robot, evidently, the robot was a naturalistic attention attractor. Subsequently, when the therapist begins the protocol explaining the tasks, the children attention shifts towards the therapist. The children remained this behavior until that the therapist introduced the robot-mediator. In this transition, the children's behaviors, such as, RJA and IJA toward the therapist were observed. Once the therapist changed the mediation with the robot, the children turned his/her attention to the robot and the objects in the room.

A more detailed analysis of one of the TD volunteers is shown in Fig. 8. The plot (A) shows the overall intervention session; the plot (B) and plot (C) are a zoom of the period with therapist and robot mediator, respectively. The colors convention in the three plots of Fig. 8 describes the results generated by the automated estimation of VFOA. From these scenarios, some essential aspects already emerge. In the therapist-mediator interval, the child responded to JA task using only one repetition for all prompt level. The child's behavior of RJA was according to the protocol, i.e., the child looked towards the therapist to wait for instructions, rapidly the child searched in the target, and next looked again toward the therapist (Color sequence: light blue - yellow - light blue - orange - light blue - red). This behavior was the same for all prompts. In contrast, with the robot-mediator, the child did not look toward the robot among indications at consecutive targets (Color sequence: light green - yellow - orange - red - orange - yellow). The above happened because, in the protocol, both mediators executed the instructions in the same order, and

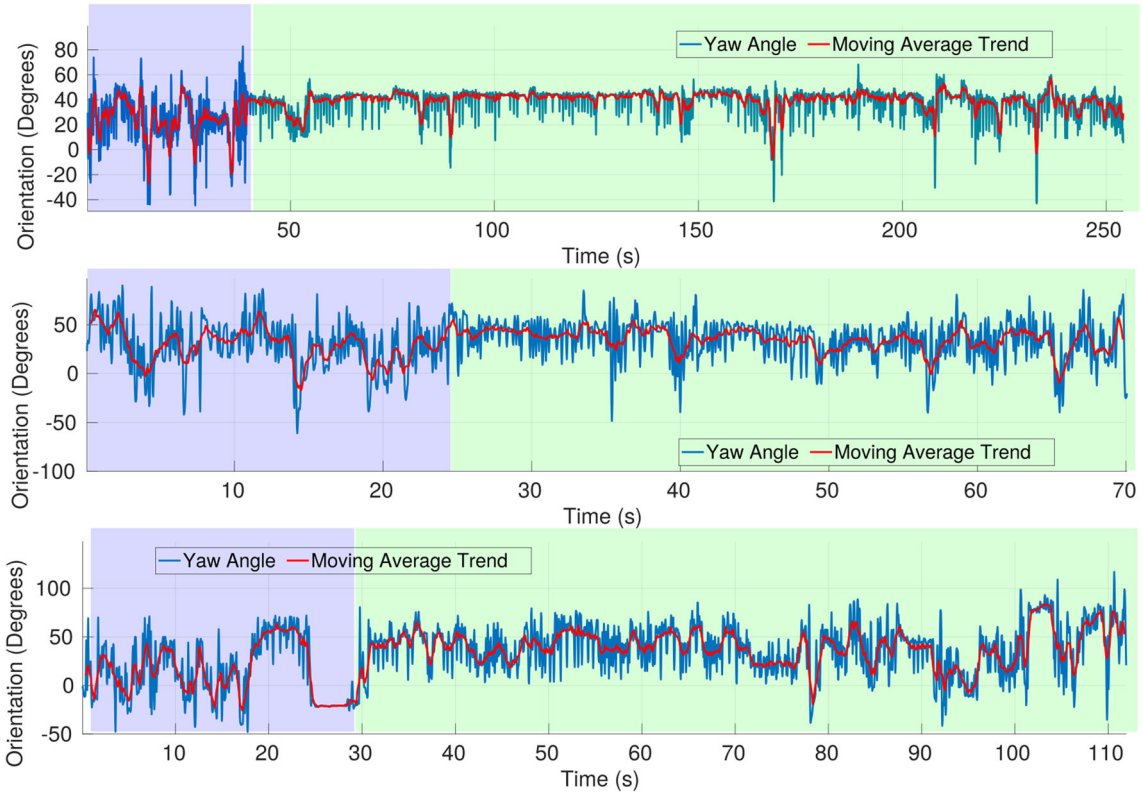

Fig. 9 Evolution over time of the child's head/neck rotation (yaw rotation) for a ASD group

the child memorized the commands and the object's position until the robot mediator interval. This fact did not affect the intervention's aim, as the robot mediator succeeded to elicit the child's behaviors of RJA and IJA. In addition, as highlighted in the plot (A) in Fig. 8, when the session finalized and the robot mediator said goodbye, again, RJA and IJA behaviors were perceived. The pictures (a-d) show these events: first the child said goodbye towards the robot, then, he looked the therapist to confirm that the session ended and looked again towards the robot, finally the child took the robot's hand.

From the analysis of the three TD volunteers, the same reported behaviors were perceived. However, the analysis of the children in the ASD group showed different behavior patterns concerning comfort, visual contact and novelty stimulus effect during the sessions. The evolution over time of the child's head/neck rotation (yaw rotation) for an ASD group is shown in Fig. 9. On the one hand, the three children in the ASD group maintained more visual contact with the robot compared to the therapist and exhibited more interest in the robot platform compared to the TD children. However, the performance of the children in the activities of JA did not improve significantly when the robot executed the prompt. On the other hand, the clinicians manifested that in all cases the first visual contact toward them occurred in the instant that the robot entered the scene and started interacting, i.e., the ONO mediation elicited behaviors of IJA towards the therapist. In addition, the CwASD exhibited less discomfort regarding the session, from the first moment when the robot initiated mediation in the room and, in some cases, when showed appearance of verbal and non-verbal pro-social behaviors. These facts did not arise with the TD children, because the first visual contact with the therapist occurred when they entered the room. Additionally, TD children showed the ability to divide the attention between the robot and the therapist from the beginning to the end of the intervention, exhibiting comfort in every moment. The behavior modulation of CwASD is observed in Fig. 9. Before the period with robot-mediator the children exhibited discomfort (unstable movements of their head), and after of this period, the head movement tended to be more stable.

The novelty of a robot-mediator at diagnostic session can be analyzed as an additional stimulus of the CRI. Accordingly, in this case study the children of the ASD group showed more behavior modification (attention and comfort) produced by the robot interaction at the beginning of the CRI, remaining until the end of the session. On the other hand, the children of the TD group responded to the novelty effect of the robot mediator from the time the child entered the room and saw the robot, until the beginning of the therapist presentation. For the above, despite the novelty

of the stimuli effect, these did not seem to affect the social interaction between the TD children and the therapist, and in contrast, these stimuli seemed to enhance the CwASD social interaction with the therapist along the intervention.

These results are impressive, since they show the potential of CRI intervention to systematically elicit differences between the pattern of behavior on TD and ASD children. We identified RJA and IJA toward the therapist at the beginning of the intervention, at the transition between therapist to robot mediator, and at the end for all TD children. In contrast, we only identified IJA towards the therapist in the transition between mediators, for ASD children. This fact shows a clear difference of behavior pattern between CwASD and TD children, which can be analyzed using a JA task protocol. In fact, these pattern differences can be used as evidence to improve the ASD diagnosis.

8 Conclusions

This work presented a Robot-Assisted tool to assist and enhance the traditional practice of ASD diagnosis. The designed framework combines a vision system with the automated analysis of nonverbal cues in addition to a robotic platform; both developed upon open source projects. This research contributes to the state-of-the-art with an innovative flexible and scalable architecture capable to automatically register events of joint attention and patterns of visual contact before and after of a robot-based mediation as well as the pattern of behavior related to comfort or discomfort along the ASD intervention.

In addition, an artificial vision pipeline based on a multi-camera approach was proposed. The vision system performs face detection, recognition and tracking, landmark detection and tracking, head pose, gaze and estimation of visual focus of attention was proposed, with its performance considered suitable for use into conventional ASD intervention. At least one camera captured the child's face in each sample frame. Furthermore, the feedback information about the child's performance was successfully used to modulate the supervised behavior of ONO, improving the performance of the CRI and the visual attention of the children. Regarding the VFOA estimation, the algorithm was able to estimate the target into the FoV in different situations recurrently. Also, the robot was able to react according to the estimation. However, the algorithm only failed when occlusion by the child's hands is generated. On the other hand, the occlusion by the therapist and the robot was compensated using the multi-camera approach. The child's face recognition system showed to be imperative to analyze the child's behavior in the clinical setup implemented in this work, which required the caregiver's attention in the room.

Despite the limited number of children of this study, preliminary results of this case study showed the feasibility of identifying and quantify differences in the patterns of behavior of TD children and CwASD elicited by the CRI intervention. Through the proof of concept, it is evidenced here the system ability to improve the traditional tools used in ASD diagnosis. As future works, it is recommended a study to replicate the protocol proposed in this paper with ten CwASD and ten TD children. Another suggestion is to quantify other kinds of behaviors in addition to that assessed in this paper, such as verbal utterance patterns, physical and emotional engagement, object or event preferences and gather more evidence to improve the assistance to therapists in ASD diagnosis processes.

Acknowledgements This work was supported by the Google Latin America Research Awards (LARA) program. The first author scholarship was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Disclosure statement No potential conflict of interest was reported by the authors.

References

1. Belpaeme, T., Baxter, P.E., de Greeff, J., Kennedy, J., Read, R., Looije, R., Neerincx, M., Baroni, I., Zelati, M.C.: Child-Robot interaction: perspectives and challenges. In: 5th International Conference, ICSR 2013, pp. 452–459. Springer International Publishing, Bristol (2013)
2. 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. **6**(1), 249–262 (2012)
3. Scassellati, B., Admoni, H., Maja, M.: Robots for use in autism research. Annu. Rev. Biomed. Eng. **14**(1), 275–294 (2012)
4. Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., Pioggia, G.: Autism and social robotics: A systematic review (2016)
5. American Psychiatric Association: DSM-5 diagnostic classification. In: Diagnostic and Statistical Manual of Mental Disorders. American Psychiatric Association, 5 (2013)
6. Eggebrecht, A.T., Elison, J.T., Feczko, E., Todorov, A., Wolff, J.J., Kandala, S., Adams, C.M., Snyder, A.Z., Lewis, J.D., Estes, A.M., Zwaigenbaum, L., Botteron, K.N., McKinstry, R.C., Constantino, J.N., Evans, A., Hazlett, H.C., Dager, S., Paterson, S.J., Schultz, R.T., Styner, M.A., Gerig, G., Das, S., Kostopoulos, P., Schlaggar, B.L., Petersen, S.E., Piven, J., Pruett, J.R.: Joint attention and brain functional connectivity in infants and toddlers. Cerebral Cortex **27**(3), 1709–1720 (2017)
7. Steiner, A.M., Goldsmith, T.R., Snow, A.V., Chawarska, K.: Disorders in infants and toddlers. J. Autism Dev. Disord. **42**(6), 1183–1196 (2012)
8. Belpaeme, T., Baxter, P.E., Read, R., Wood, R., Cuayáhuitl, H., Kiefer, B., Racioppa, S., Kruijff-Korbayová, I., Athanasopoulos, G., Enescu, V., Looije, R., Neerincx, M., Demiris, Y., Ros-Espinoza, R., Beck, A., Canamero, L., Hiolle, A., Lewis, M., Baroni, I., Nalin, M., Cosi, P., Paci, G., Tesser, F., Sommavilla, G., Humbert, R.: Multimodal child-robot interaction: building social bonds. Journal of Human-Robot Interaction **1**(2), 33–53 (2012)

9. Vanderborght, B., Simut, R., Saldien, J., Pop, C., Rusu, A.S., Pintea, S., Lefeber, D., David, D.O.: Using the social robot probo as a social story telling agent for children with ASD. Interact. Stud. **13**(3), 348–372 (2012)
10. Warren, Z.E., Zheng, Z., Swanson, A.R., Bekele, E., Zhang, L., Crittendon, J.A., Weitlauf, A.F., Sarkar, N.: Can robotic interaction improve joint attention skills? J. Autism Dev. Disord. **45**(11), 3726–3734 (2015)
11. Wood, L.J., Dautenhahn, K., Lehmann, H., Robins, B., Rainer, A., Syrdal, D.S.: Robot-mediated interviews: Do robots possess advantages over human interviewers when talking to children with special needs? Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **8239 LNAI**, 54–63 (2013)
12. Feil-Seifer, D., Mataric, M.J.: b3IA A control architecture for autonomous robot-assisted behavior intervention for children with Autism Spectrum Disorders. In: ROMAN 2008 The 17th IEEE International Symposium on Robot and Human Interactive Communication, pp. 328–333 (2008)
13. Leo, M., Del Coco, M., Carcagní, P., Distante, C., Bernava, M., Pioggia, G., Palestra, G.: Automatic emotion recognition in Robot-Children interaction for ASD treatment. In: Proceedings of the IEEE International Conference on Computer Vision, 2015-Febru(c), pp. 537–545 (2015)
14. Esteban, P.G., Baxter, P.E., Belpaeme, T., Billing, E., Cai, H., Cao, H.-L., Coeckelbergh, M., Costescu, C., David, D., De Beir, A., Fang, Y., Ju, Z., Kennedy, J., Liu, H., Mazel, A., Pandey, A., Richardson, K., Senft, E., Thill, S., Van De Perre, G., Vanderborght, B., Vernon, D., Hui, Y., Ziemke, T.: How to build a supervised autonomous system for Robot-Enhanced therapy for children with autism spectrum disorder. Paladyn Journal of Behavioral Robotics **8**(1), 18–38 (2017)
15. Pour, A.G., Taheri, A., Alemi, M., Ali, M.: Human–Robot facial expression reciprocal interaction platform: case studies on children with autism. Int. J. Soc. Robot. **10**(2), 179–198 (2018)
16. Feng, Y., Jia, Q., Wei, W.: A control architecture of Robot-Assisted intervention for children with autism spectrum disorders. J. Robot. **2018**, 12 (2018)
17. Bekele, E., Crittendon, J.A., Swanson, A., Sarkar, N., Warren, Z.E.: Pilot clinical application of an adaptive robotic system for young children with autism. Autism: The International Journal of Research and Practice **18**(5), 598–608 (2014)
18. Huijnen, C.A.G.J., Lexis, M.A.S., Jansens, R., de Witte, L.P.: Mapping robots to therapy and educational objectives for children with autism spectrum disorder. J. Autism Dev. Disord. **46**(6), 2100–2114 (2016)
19. Aresti-Bartolome, N., Begonya, G.-Z.: Technologies as support tools for persons with autistic spectrum disorder: s systematic review. Int. J. Environ. Res. Public Health **11**(8), 7767–7802 (2014)
20. Boucenna, S., Narzisi, A., Tilmont, E., Muratori, F., Pioggia, G., Cohen, D., Mohamed, C.: Interactive technologies for autistic children: a review. Cogn. Comput. **6**(4), 722–740 (2014)
21. Grynszpan, O., Patrice, L., Weiss, T., Perez-Diaz, F., Gal, E.: Innovative technology-based interventions for autism spectrum disorders: a meta-analysis. Autism **18**(4), 346–361 (2014)
22. Rehg, J.M., Rozga, A., Abowd, G.D., Goodwin, M.S.: Behavioral imaging and autism. IEEE Pervasive Comput. **13**(2), 84–87, 4 (2014)
23. Cabibihan, J.J., Javed, H., Ang, M., Aljunied, S.M.: Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. Int. J. Soc. Robot. **5**(4), 593–618 (2013)
24. Sartorato, F., Przybylowski, L., Sarko, D.K.: Improving therapeutic outcomes in autism spectrum disorders: enhancing social communication and sensory processing through the use of interactive robots. J. Psychiatr. Res. **90**, 1–11 (2017)
25. Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R.M., Rozga, A., Rehg, J.M.: Detecting gaze towards eyes in natural social interactions and its use in child assessment. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **1**(3), 43:1–43:20 (2017)
26. Ness, S.L., Manyakov, N.V., Bangerter, A., Lewin, D., Jagannatha, S., Boice, M., Skalkin, A., Dawson, G., Janvier, Y.M., Goodwin, M.S., Hendren, R., Leventhal, B., Shic, F., Cioccia, W., Gahan, P.: JAKE® Multimodal data capture system: Insights from an observational study of autism spectrum disorder. Frontiers in Neuroscience 11(SEP) (2017)
27. Rehg, J.M., Abowd, G.D., Rozga, A., Romero, M., Clements, M.A., Sclaroff, S., Essa, I., Ousley, O.Y., Li, Y., Kim, C., Rao, H., Kim, J.C., Lo Presti, L., Zhang, J., Lantsman, D., Bidwell, J., Ye, Z.: Decoding children's social behavior. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3414–3421 (2013)
28. Adamo, F., Palestra, G., Crifaci, G., Pennisi, P., Pioggia, G., Ruta, L., Leo, M., Distante, C., Cazzato, D.: Non-intrusive and calibration free visual exploration analysis in children with autism spectrum disorder. In: Computational Vision and Medical Image Processing V - Proceedings of 5th Eccomas Thematic Conference on Computational Vision and Medical Image Processing, VipIMAGE 2015, pp .201–208 (2016)
29. Michaud, F., Salter, T., Duquette, A., Mercier, H., Lauria, M., Larouche, H., Larose, F.: Assistive technologies and Child-Robot interaction. American Association for Artificial Intelligence **ii**(3), 8–9 (2007)
30. Duquette, A., Michaud, F., Mercier, H.: Exploring the use of a mobile robot as an imitation agent with children with low-functioning autism. Auton. Robot. **24**(2), 147–157 (2008)
31. Simut, R.E., Vanderfaeillie, J., Peca, A., Van de Perre, G., Bram, V.: Children with autism spectrum disorders make a fruit salad with probo, the social robot: an interaction study. J. Autism Dev. Disord. **46**(1), 113–126 (2016)
32. Bekele, E., Lahiri, U., Swanson, A.R., Crittendon, J.A., Warren, Z.E., Nilanjan, S.: A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism. IEEE Trans. Neural Syst. Rehabil. Eng. **21**(2), 289–299 (2013)
33. Zheng, Z., Zhang, L., Bekele, E., Swanson, A., Crittendon, J.A., Warren, Z.E., Sarkar, N.: Impact of robot-mediated interaction system on joint attention skills for children with autism. In: IEEE International Conference on Rehabilitation Robotics (2013)
34. Anzalone, S.M., Tilmont, E., Boucenna, S., Xavier, J., Jouen, A.L., Bodeau, N., Maharatna, K., Chetouani, M., Cohen, D.: How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D + time) environment during a joint attention induction task with a robot. Res. Autism Spectr. Disord. **8**(7), 814–826 (2014)
35. Chevalier, P., Martin, J.C., Isableu, B., Bazile, C., Iacob, D.O., Adriana, T.: Joint attention using human-robot interaction: impact of sensory preferences of children with autism. In: 25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, pp. 849–854 (2016)
36. Lemaignan, S., Garcia, F., Jacq, A., Dillenbourg, P.: From real-time attention assessment to "with-me-ness" in human-robot interaction. In: ACM/IEEE International Conference on Human-Robot Interaction, 2016-April, pp. 157–164 (2016)
37. Del Coco, M., Leo, M., Carcagni, P., Fama, F., Spadaro, L., Ruta, L., Pioggia, G., Distante, C.: Study of mechanisms of

social interaction stimulation in autism spectrum disorder by assisted humanoid robot. IEEE Transactions on Cognitive and Developmental Systems **8920**(c), 1–1 (2017)
38. Palestra, G., Varni, G., Chetouani, M., Esposito, F.: A multimodal and multilevel system for robotics treatment of autism in children. In: Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents - DAA '16, pp. 1–6. ACM Press, New York (2016)
39. Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J., Berger, E., Wheeler, R., Ng, A.: ROS : an open-source robot operating system. In: ICRA workshop on open source software, number 3.2, pp. 5 (2009)
40. Vandevelde, C., Saldien, J., Ciocci, C., Vanderborght, B.: The use of social robot ono in robot assisted therapy. In: International Conference on Social Robotics, Proceedings, m (2013)
41. Dautenhahn, K.: A paradigm shift in artificial intelligence: why social intelligence matters in the design and development of robots with human-like intelligence. 50 Years of Artificial Intelligence, pp. 288–302 (2007)
42. Ekman, P., Friesen, W.: Facial Action Coding System. Consulting Psychologists Press (1978)
43. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
44. Baltrušaitis, T., Robinson, P., Morency, L.-P.: OpenFace: an open source facial behavior analysis toolkit. IEEE Winter Conference on Applications of Computer Vision (2016)
45. King, D.E.: Max-Margin Object Detection. 1 (2015)
46. He, K., Zhang, X., Ren, S., Jian, S.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, 6
47. Baltrušaitis, T., Robinson, P., Morency, L.P.: Constrained local neural fields for robust facial landmark detection in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 354–361 (2013)
48. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: Proceedings of the British Machine Vision Conference 2006, pp. 1–95 (2006)
49. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. Int. J. Comput. Vis. **91**(2), 200–215 (2011)
50. Baltrušaitis, T., Robinson, P., Morency, L.P.: 3D constrained local model for rigid and non-rigid facial tracking. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2610–2617 (2012)
51. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Neeraj, K.: Localizing parts of faces using a consensus of exemplars. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2930–2940 (2013)
52. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive Facial Feature Localization, pp. 679–692. Springer, Berlin (2012)
53. Jorstad, A., Dementhon, D., Jeng Wang, I., Burlina, P.: Distributed consensus on camera pose. IEEE Trans. Image Process. **19**(9), 2396–2407 (2010)
54. Ba, S.O., Odobez, J.-M.: Multi-Person visual focus of attention from head pose and meeting contextual cues. IEEE Trans. Pattern Anal. Mach. Intell. **33**(August), 1–16 (2008)
55. Sheikhi, S., Jean-Marc, O.: Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. Pattern Recogn. Lett. **66**, 81–90 (2015)

Andrés A. Ramírez-Duque received his bachelor's degree in Mechatronics Engineering from the Universidad Nacional de Colombia, Bogotá, Colombia, in 2009, and his Industrial Automation Master degree from the Universidad Nacional de Colombia, Bogotá, Colombia, in 2011. He is currently working toward a Ph.D. degree in the Assistive Technology Center, Federal University of Espírito Santo, Vitória, Brazil. He won a Google Latin America Research Award 2017. His current research interests include Child-Robot interaction, cloud parallel computing, high performance computing, smart environments and serious games applied to Children with development impairments.

Anselmo Frizera-Neto received his bachelor's degree in Electrical Engineering (2006) from the Federal University of Espírito Santo (UFES) in Brazil and his doctorate in Electronics (2010) at the University of Alcalá, Spain. From 2006 to 2010 he was a researcher of the Bioengineering Group of the Consejo Superior de Investigaciones Científicas (Spain) where he carried out research related to his doctoral thesis. He is currently a permanent professor and adjunct coordinator of the Graduate Program in Electrical Engineering at UFES. He has authored or co-authored more than 250 papers in scientific journals, books and conferences in the fields of electrical and biomedical engineering. He has conducted or co-directed master's and doctoral theses in research institutions from Brazil, Argentina, Italy and Portugal. His research is aimed at rehabilitation robotics, the development of advanced strategies of human-robot interaction and the conception of sensors and measurement technologies with applications in different fields of electrical and biomedical engineering. Along with Andrés Ramírez-Duque, he won a Google Latin America Research Award 2017.

Teodiano Freire Bastos received his B.Sc. degree in Electrical Engineering from Universidade Federal do Espírito Santo (Vitória, Brazil) in 1987, his Specialist degree in Automation degree from Instituto de Automática Industrial (Madrid, Spain) in 1989, and his Ph.D. degree in Physical Science (Electricity and Electronics) from Universidad Complutense de Madrid (Spain) in 1994. He made two postdocs, one at the University of Alcalá (Spain, 2005) and another at RMIT University (Australia, 2012). He is currently a full professor at Universidade Federal do Espírito Santo (Vitória, Brazil), teaching and doing research at the Postgraduate Program of Electrical Enginneering, Postgraduate Program of Biotechnology and RENORBIO Ph.D. Program. His current research interests are signal processing, rehabilitation robotics and assistive technology for people with disabilities

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com