Sensing-enhanced Therapy System for Assessing Children with Autism Spectrum Disorders: A Feasibility Study

Haibin Cai, Yinfeng Fang, Zhaojie Ju, Cristina Costescu, Daniel David, Erik Billing, Tom Ziemke, Serge Thill, Tony Belpaeme, Bram Vanderborght, David Vernon, Kathleen Richardson and Honghai Liu

***Abstract*—It is evident that recently reported robot-assisted therapy systems for assessment of children with autism spectrum disorder (ASD) lack autonomous interaction abilities and require significant human resources. This paper proposes a sensing system that automatically extracts and fuses sensory features such as body motion features, facial expressions, and gaze features, further assessing the children behaviours by mapping them to therapist-specified behavioural classes. Experimental results show that the developed system has a capability of interpreting characteristic data of children with ASD, thus has the potential to increase the autonomy of robots under the supervision of a therapist and enhance the quality of the digital description of children with ASD. The research outcomes pave the way to a feasible machine-assisted system for their behaviour assessment.**

***Index Terms*—Sensing-enhanced, autonomy, autism spectrum disorders, Therapy**

I. Introduction

Autism spectrum disorder (ASD) refers to a group of psychological conditions characterised by highly repetitive behaviour, severely restricted interests and widespread abnormalities in social interactions and communication [1]. It has received much attention in recent years due to its increasing prevalence, and has become an urgent public health concern [2, 3]. One way to alleviate the impact of ASD is the employment of early therapeutic interventions. Recent research showed that early behavioural therapies could result in a significant maintained improvement in IQ, language, social behaviours [4]. With early behavioural therapies, individuals with ASD are expected to gain completely or almost-completely independent lives at a later stage. However, it has been found to be too expensive and time-consuming to provide associated care of individuals of ASD [5, 6].

This work was supported by EU seventh framework programme DREAM (No. 611391). (corresponding author: H. Liu)

H.Cai, Y.Fang, Z.Ju and H.Liu are with the School of Computing, University of Portsmouth, U.K.

C. Costescu and D. David are with Department of Clinical Psychology and Psychotherapy, Babe-Bolyai University, Cluj-Napoca, Romania.

E. Billing and T. Ziemke are with Interaction Lab, School of Informatics, University of Skovde, Skovde, Sweden. (T. Ziemke is also with Department of Computer and Information Science, Linkoping University, Sweden.)

S. Thill and T. Belpaeme are with University of Plymouth, U.K.

B. Vanderborght is with Vrije Universiteit Brussel and Flanders Make, Belgium.

D. Vernon is with Carnegie Mellon University Africa, Rwanda.

K. Richardson is with De Montfort University, UK.

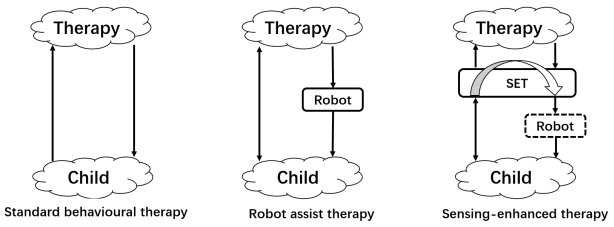


Fig. 1. An illustration of the standard behavioural therapy, robot assisted therapy and sensing-enhanced therapy. The dotted lines in sensing-enhanced therapy mean that the robot can be removed. Thus, the developed system can also be applied in the standard behavioural therapy to reduce the burden for therapists.

Many researchers have suggested ways to reduce the costs and increase the effectiveness of traditional standard behavioural therapies. Robot assisted therapy (RAT) is one of the promising solutions to improve social skills for children with ASD as they exhibit a preference for interacting with non-human agents [7, 8]. Compared to humans, robots are more predictable in repeating specific behaviours and simpler to interact with and thus can be served as an intermediate for human-human interaction [9]. In addition, the robots also have the advantage of being the physical technology which attracts more attention of children with ASD than a human [10]. Consequently, robots have been employed to interact with children with ASD in different ways such as play therapy [11, 12], social communication [13] and joint attention [8].

In a general RAT intervention such as the Wizard of Oz (WoZ) [14], the robot is mostly controlled by an extra operator, hence can not respond autonomously according to children's behaviour. The requirement of the extra operator not only introduces extra costs for the intervention but also increases the complexity to infer the behaviour of children since their information such as facial expression and gaze is mostly non-visible to the operator. Furthermore, extra efforts are needed to analysis children's performance after the intervention. As a result, there is a need to increase the autonomy of the robots to reduce the burden for therapists and get a better consistent therapeutic intervention experience [15].

This paper presents a novel sensing-enhanced therapy (SET) system to improve existing systems of both standard and robot assisted therapy by providing the multi-model data sensing and analysis results. Fig. 1 shows an illustration of the standard behavioural therapy, robot assisted therapy and sensing-enhanced

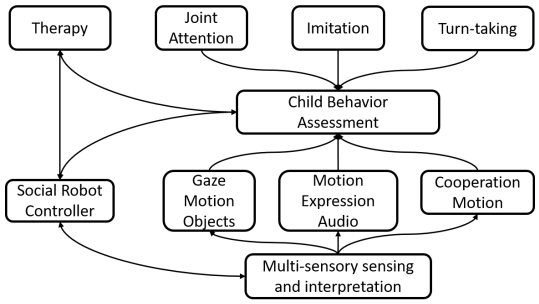


Fig. 2. The framework of the sensing-enhanced therapy system.

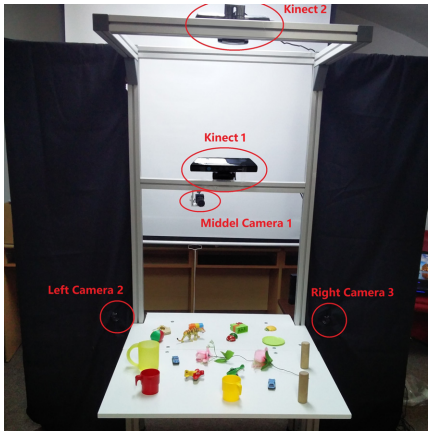


Fig. 3. An illustration of the SET system. The developed sensorized intervention table employs three RGB cameras and two Kinect RGBD sensors. The three cameras are placed in the right, middle and left side of the intervention table. The right and left cameras are hidden by black curtains to avoid distracting the attention of the ASD child. The Kinects are placed on the middle bar and top bar of the intervention table. The width of the middle bar is around 70cm. Its distance to the table is around 60cm to avoid the occlusion of the sensors when a robot is standing on the table.

therapy. In the developed SET system, the child-robot interaction, sensory processing, behaviour assessment, and social robot control can be integrated into a closed loop. Therefore, the robot can operate autonomously according to children's response for a certain time and transfer control to therapists whenever circumstances require. Thus, therapists can not only be released from manually annotating the children's reactions but also benefit from extra data for a better understanding of the children's behavioural state.

The SET system gives priority to three social interventions which cover the principle components of therapeutic interventions and are very common in children with ASD [16]. The three interventions are imitation, joint attention, and turn-taking. It is generally accepted that imitation lays the foundation for personal skill development and communications. Numerous experiments [17–19] have established that repeated imitation training can help children learn new information from the social environment, improve imitation skills and enhance social responsiveness. Joint attention is the progress of following a partner's instructions to interact with objects, either by gazing, pointing or speaking. Studies in [20] showed that the lack of joint attention skills could be seen as an early sign of autism. Turn-taking is the interchanges of behaviours between communicative partners [21]. It is frequently targeted in social skill interventions for children with ASD [7, 22].

Fig. 2 shows the framework of the SET system. Children's behaviors in the three interventions are decomposed into several components (gaze, expression, motion etc.) calculated by the multi-sensory sensing and interpretation module. Based on the outputs of each component, the child behaviour assessment module can provide therapists with useful analysed behaviour information for the diagnosis, care, and treatment of children with ASD, replacing current labour intensive techniques involving papers and pencils, or manual video analysis. Meanwhile, the autonomy of the robot can also be improved by feed this information to the robot social controller module. The therapy module allows users to take back control of the system at any time.

This work fits in a broad project called DREAM to increase the autonomy level of social robots in therapy to move beyond WoZ [23] and fits in a personalized and platform-independent behavior control system for social robots [24]. Unlike our previous work [24] which focused on planning an holistic interaction system and ignored practical challenges of sensory multi-modal data acquiring, fusing and interpreting, this paper proposes a workable sensing system configuration with experimental validation on the recorded data of children with ASD. Our contributions are listed as follows:

1. The development of several components for children's behaviour analysis. These components include gaze estimation, action recognition, facial expression recognition, object tracking, object recognition, sound direction detection and speech recognition.

2. The design of a multi-sensory system configuration which contains a data synchronization strategy and a calibration procedure to address practical challenges and further enable an efficient and effective fusion of the multi-modal data.

3. Experimental results on each component validate the performance of the SET system in analysing the behaviour of children with ASD.

The rest of this paper is organised as follows: an introduction of the SET system and the multi-sensory system configuration is given in Section II. Section III describes the details of each component. Section IV presents the experimental evaluation of each component for assessing children's behaviour. Finally, the paper is concluded with discussions in Section V.

II. SET System Configuration

The SET system aims to infer the psychological disposition of children with ASD and assess their behaviour in joint attention, imitation and turn taking interventions. Fig. 3 shows an example configuration of the system which mainly includes an intervention table, a Nao robot, multiple sensors, fixing accessories and a computer workstation. Three narrow field-of-view cameras with a resolution of 1280∗960 and two RGBD sensors mounted on the intervention table are used for the multi-modal data sensing of the children and therapeutic environment.

The functional structure of the system is shown in Fig. 4. The three cameras faced towards the middle of the intervention

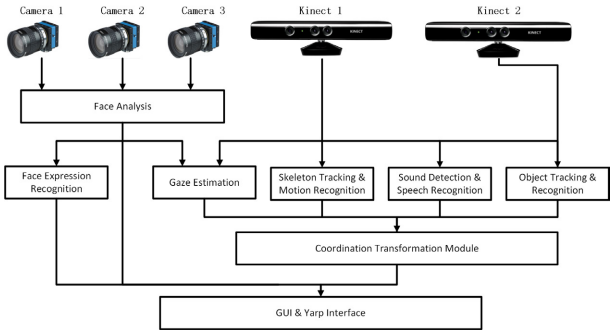


Fig. 4. The overall functional structure of the sensors.

Fig. 5. The Graphical User Interface (GUI) component used by the therapist to control the sensory system. On the right bottom side, the control button panel can be used to control the preview and record function. The "Show3D" button is used to present all sensitised data in a unified 3D world coordinate.

table can capture valid face images even under large head poses. The captured face images can then be used for face analysis purposes such as gaze estimation and facial expression recognition. The front-mounted Kinect 1 aims to capture the child's movement information and audio information. The captured RGBD information is further fused with the images captured by the three cameras for gaze estimation. The top-mounted Kinect 2 used for object tracking and recognition can provide necessary objects' information during the intervention.

The multi-sensor configuration requires the system to address significant challenges of multi-sensory data sensing and fusing. These challenges are addressed by designing a data synchronisation strategy and a calibration procedure.

A. Data Synchronisation

One of the great challenges in multi-modal data sensing is the data synchronisation problem. For example, Funes Mora et al. [25] claimed that one of the main difficulties when constructing a system with a Kinect and a high-resolution camera to capture the eye gaze information is the data synchronisation problem. They proposed to enhance the synchronisation ability by placing several LED lights in the joint view of each sensor. However, this approach requires extra hardware configuration and is not suitable when the views of sensors are not overlapped to a large extent.

To deal with the synchronisation problem, this paper proposes a multi-thread programming framework, in which each sensor owns a separate thread triggered by a central thread. Unlike common multi-thread programming functionalities, each camera thread in this framework has only one loop inside. The target of each camera thread is to capture only one image for gaze estimation and facial expression recognition. Similarly, the Kinect 2 thread aims to capture only one RGB image and one depth image for object tracking and recognition. Compared to the Kinect 2 thread, the Kinect 1 thread needs to capture extra skeleton information for the action recognition component. The central thread is used to repeatedly create these sensing threads and wait for their termination. In this way, we can not only assure a fixed frame numbers for each video but also keep a minimum time difference for the same set of video frames during the recording procedure. Unlike the video processing threads, the audio recording and processing tasks operate separately since they are not measured by video frame numbers. To balance the speech recognition performance and other tasks, a separate thread is constructed for receiving the results of the speech recognition. In the constructed thread, the speech recognition function and a 100 ms sleep operation is repeatedly executed.

Apart from the data sensing and recording, the SET system needs to simultaneously conduct data analysis for the robot system described in [23, 24] to interact with children with ASD. This brings great challenges since the algorithms for gaze estimation, facial expression recognition, object tracking and face alignment require much computational time, which will be presented in detail in Section IV. In the developed multi-thread programming framework, the priority is given to the real-time data sensing and recording by using the aforementioned central thread to repeatedly trigger each sensor's thread. The adopted camera sensors can capture 25 fps and the Kinect sensor has a high speed of 30 fps. This time difference is also utilized by directly integrating the efficient action recognition algorithm into the Kinect 1 sensor's thread. The rest of the algorithms are packaged into an interface function, which will only be executed once required and thus do not affect the performance of data sensing and recording. During the intervention, the interface function is repeatedly called and the results are transferred to the robot system through the Yarp platform [26].

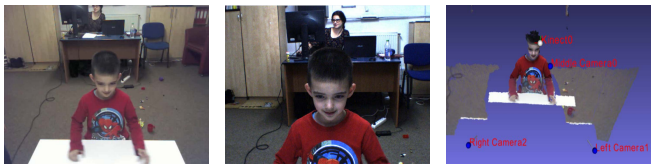
Fig. 5 shows the developed Graphical User Interface (GUI) component. To facilitate the user operation, the SET system provides two modes, namely, a previewing mode and a recording mode. The difference between the two modes is that the recording mode requires an extra saving operation in each thread. The SET system can effectively collect and analyse multi-sensory data in a real-time performance at 25 fps. Some of the analysed results such as detected skeletons, recognised motion, face identification, facial landmarks, facial expression, head pose and gaze are directly shown in the images. The right bottom part of the GUI is a panel to control the system.

B. System Calibration

Due to the multi-sensory configuration, the detected information such as eye centres, object locations, and skeletons are all in the local coordinate of the individual sensors. To

effectively fuse multi-modal data, a coordinate transformation module, which can transfer data from different sensor coordinate systems to a global world coordinate system, is proposed. By doing so, all the collected sensory data can be shown in the global world coordinate system. The centre of the world coordinate system is located at the base of Kinect 1. The vertical axis is defined as the y axis, and the desk plane is defined as the plane of x axis and z axis. The following part of this section shows the methods to calibrate the Kinect 1 with three cameras and the two Kinects respectively.

1) Kinect-Camera calibration: The system places the three cameras at three different locations to capture faces under large head poses. The large angle difference of the three cameras brings challenges for the calibration progress. We have proposed a joint Kinect-Camera calibration framework to simultaneously calibrate relative poses of a Kinect and three cameras. By weighting each camera, a single cost function is constructed to jointly calibrate the Kinect with the three cameras [27]. Fig. 6 shows an illustration of the captured colour images and the reconstructed 3D point clouds. It should be noted that in the reconstructed image, the upper body point clouds are rendered by the high-resolution camera while the rest of the point clouds are rendered by the Kinect. The correct alignment of the point clouds demonstrates the accurate calibration of the sensors.



(a) (b) (c)

Fig. 6. An illustration of the Kinect-Camera calibration result. (a) An image captured by one of the cameras. (b) An image captured by the Kinect 1. (c) The reconstructed 3D point clouds using the camera and the Kinect 1.

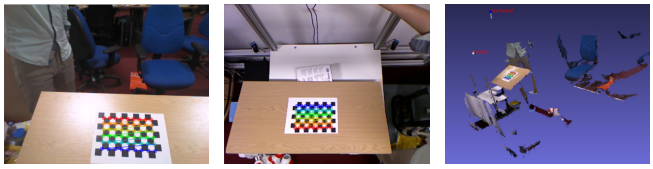
2) Kinect-Kinect calibration: The SET system uses two Kinects to capture both the 3D information of the child and the objects. The relative positions of two Kinects are calibrated by minimising the following equation:

$$min_{R,T} = ||RP_1 + T - P_2|| \tag{1}$$

where P_1 and P_2 represent the coordinates of the image corners detected from the RGB images of the Kinect 2 and the Kinect 1 respectively. R and T are the rotation matrix and the translation matrix for the coordinate transformation between two sensors. Due to the limitation of depth sensors, some of the detected corners might not contain depth information. After removing these points, the iterative closest points (ICP) algorithm is used for the optimisation of R and T. Fig. 7 shows an illustration of the detected corners and the calibration result.

Once the sensors are calibrated, we can transform the coordinates of the objects from one sensor to another sensor via the following equation:

$$\begin{cases} \frac{u_c - u_0}{X} = \frac{v_c - v_0}{Y} = \frac{f}{Z} \\ P_t = RP_0 + T \\ P = (X, Y, Z) \end{cases} \tag{2}$$



(a) (b) (c)

Fig. 7. An illustration of the Kinect-Kinect calibration result. (a) An image captured by the middle Kinect. (b) An image captured by the top Kinect. (c) The calibration result of two Kinects.

where (u_c, v_c) represent the position of detected corners in RGB images and (u_0, v_0) denote the cameras' image centre. P_t, P_o represent the recovered 3D coordinates in the target world coordinate and the local coordinate respectively.

III. Perception Components

The developed perception components in the SET system are gaze, actions, facial expressions, object positions, object identifications, sound direction, and speeches. The algorithms for these components are explained in detail in the following subsections.

A. Gaze Estimation Component

Gaze is an essential part of human's attention system. Although accurate gaze can be obtained using commercial head-mounted eye tracking devices, the uncomfortable wearing experience limits their application for the children with ASD. Apart from the non-wearable requirement, the joint attention tasks also require the SET system to estimate the 3D gaze of children with ASD under large head movements, so that they can freely move their heads while doing motions. These requirements bring extra challenges such as large head movements, occlusions, and various eye appearances. To handle these challenges, we have proposed a gaze estimation method using a single sensor [28] and further developed a multi-sensory configuration to cover large head movements.

As the first step, the boosted cascade face detector [29] is employed to find the rough location of the child's face. Once the face is detected, the supervised descent method proposed by [30] is used to locate the feature points in the facial region. For the detection of eye centre locations, we proposed an accurate convolution based integro-differential method [31] to localise the eye centre even in low-resolution images. The proposed method takes advantage of the drastic intensity changes between the iris and the sclera and localises the eye centre via searching the maximum ratio derivative of the neighbour curve magnitudes in the convolution image.

Once the facial points are located, the Pose from Orthography and Scaling with ITerations (POSIT) proposed by [32] is used to calculate the head pose. To handle the large head movement challenge, we have proposed a real-time gaze estimation method by constructing a multi-sensory fusion system [33]. For those facial points that the camera and Kinect 1 can both capture, it is feasible to find their global 3D coordinates. However, it is sometimes hard for both devices to capture the same facial points in many situations because of the large head

movements. Thus a 2D to 3D coordinate transformation for these located 2D facial points is necessary. The transformation can be performed using the following equation:

$$\begin{cases} P_C = R^{-1} * P_W - R^{-1} * T \\ \frac{X_C^{'}}{u-u_0} = \frac{Y_C^{'}}{v-v_0} = \frac{Z_{PC}^{'}}{f} \\ Z_C^{'} = Z_{PC}^{'} \end{cases} \tag{3}$$

where P_W and P_C represent the head centre positions in the world coordinate system and the local coordinate system respectively. (u_0, v_0) and f denote the cameras' image centre and focal length respectively; $X_C^{'}, Y_C^{'}, Z_C^{'}$ which indicate the 3D coordinate of a point in the local coordinate system correspond to the 2D point (u, v) in the image. The depth value of the head centre point $Z_{PC}^{'}$ is used as a replacement of the missing depth value of those facial landmarks for the calculation of their 3D points in the local coordinate system.

B. Facial Expression Recognition Component

Facial expressions are important aspects of human behaviours. It has also been shown that children with ASD can improve their social skills by participating facial expression related interventions [34]. The SET system aims to recognise five facial expressions which are neutral, angry, fear, happy and sad. We have proposed a face frontalization method [35] to register frontal facial appearances from unconstrained non-frontal facial images. Then we use Local Binary Patterns to represent facial appearance cues and apply a SVM for facial expression classification.

In the proposed approach [35], five different templates are manually designed to match facial expressions. The five templates are constructed by averaging the shape of five manually grouped facial images from the SFEW dataset [36]. For each query image and the detected facial landmarks, the best template will be assigned to it according to the similarity calculated by the geometric distance. In order to reconstruct frontal facial appearances, Active Appearance Model instantiation [37] can be used by minimising:

$$\sum_x \|F - I(W(x; p + \triangle p))\|, s.t. F = \sum_{i=1}^{m} \lambda_i A_i(x) \tag{4}$$

where F is the frontal face which is obtained by a linear combination of a set of pre-defined eigenfaces $A_i(x)$. The input image is warped to the selected template through piecewise affine warp $I(W(x; p + \triangle p))$. The algorithm works iteratively with the update rule $p \leftarrow p + \delta p$.

C. Action Recognition Component

Correctly recognising children's body actions is one of the most important aspects of the imitation tasks. The SET system described in this paper aims to recognise 11 actions as defined in Fig. 8. The actions are *wave hands, hands on eyes, hands over head, open arm, move toy car, drink, knock door and other four complex movements*. These actions are defined according to therapists' view in the imitation tasks for children with ASD.

Fig. 8. Defined actions for children with ASD.

Skeleton information is appealing for human action recognition in that it is invariant to illumination conditions and body appearances. We have proposed a novel skeleton based method [38] for human action recognition. The 3D Moving Trend and Geometry property (3DMTG) from skeleton joints (totally 10 joints from the up-body) are extracted to recognise the behaviour of children with ASD. The moving trend is firstly computed by accumulating all the moving directions in 3D space. Then the geometry property of joints in each frame is modelled by the relative motion information. Finally, the feature descriptor is constructed by integrating the two features for action recognition.

Eq. 5 shows the modelling of the 3D moving trend feature. 3D moving directions are partitioned into m bins $\mathbf{v_j}$. Then the cosine similarity is applied to describe the similarity between $\mathbf{v_t^i}$ and $\mathbf{v_j}$. The quantization of moving directions is implemented using a soft voting strategy.

$$\begin{cases} \mathbf{v_t^i} = \{x_{p_t^i} - x_{p_{t-1}^i}, y_{p_t^i} - y_{p_{t-1}^i}, z_{p_t^i} - z_{p_{t-1}^i}\} \\ cos\theta_j^i(t) = \frac{\mathbf{v_j} \cdot \mathbf{v_t^i}}{\|\mathbf{v_t^i}\| \|\mathbf{v_j}\|}, j \in [1, m] \\ bin_j = \sum_i \|\mathbf{v_t^i}\| \times max\{cos\theta_j^i(t)\}, j \in [1, m] \\ H(i) = bin_1, ..., bin_m \\ \triangle d_t^i = \{x_t^{ri} - x_1^{ri}, y_t^{ri} - y_1^{ri}, z_t^{ri} - z_1^{ri}\} \\ G(t) = \{\triangle d_t^1, ..., \triangle d_t^N\} \end{cases} \tag{5}$$

where $x_{p_t^i}, y_{p_t^i}, z_{p_t^i}$ represent the coordinate of the ith joint. $H(i)$ and $G(t)$ mean the extracted moving trend feature and geometry feature. To eliminate the influence of different initial poses, the displacement between the relative joints in the current frame and the joints in the initial frame is utilised to reflect the geometry property. After extracting the 3DMTG feature descriptor, a SVM classifier is used for action recognition. In practice, a sliding window strategy is used to classify online video streams.

D. Object Tracking and Recognition Component

During the imitation intervention, the child is required to pick up an object on the table and imitate the behaviour related to the object. Thus, there is a need to track and recognise the object. There has been active research in the literature for these tasks. The SET system described in this paper employs the GM-PHD tracker [39] due to its good balance of accuracy and time efficiency for multi-objects tracking. For the object classification, the classic Histogram of Oriented Gradient and SVM are used, and achieve a good performance since the white background of the intervention table reduces the challenge to a large extent.

TABLE I
THE DEMOGRAPHICS OF THE PARTICIPATION.

Participant	No.1	No.2	No.3	No.4	No.5
Gender	M	M	M	F	F
Age	5 years 4 month	5 years 1 month	3 years 9 month	3 years 8 month	4 years 8 month
ADOS	16	12	23	23	25
LF	moderate level	high level	low level	low level	low level
CS	7	4	15	10	11
SIC	9	8	8	12	12
PS	2	1	2	4	4
SBS	4	2	3	6	6

[1] LF = Level of functioning based on the category developed by Gotham et. al [40]; ADOS = Autism Diagnostic Observation Scale [41]; LF = Level of functioning; CS = Communication subscale of ADOS; SIS = Social interaction subscale of ADOS; PS = Play subscale of ADOS; SBS = Stereotype behaviors subscale of ADOS.

E. Sound Direction Detection and Speech Recognition Component

The speech recognition and sound direction localisation can help the SET system better understand the psychological disposition of the children in imitation tasks. Their implementations are based on the Microsoft Kinect SDK. In the defined intervention scene, we mainly focus on eleven pre-defined onomatopoeias, which are recognised by manually scripted similar pronunciation words into an XML file. Although the simulation of these 11 sounds for common kids is simple, it is hard for the children with ASD to accomplish. The recognised speech will also be identified by the system to see if the speech actually comes from the child by using the detected sound direction since the child will always sit in front of the intervention table during the experiment. Based on the analysed information, the system can determine whether to give a positive feedback or encourage the child to do another try.

IV. PERFORMANCE EVALUATION

Participants included in the study are children with a diagnosis of ASD between the ages of 3 to 6 years old. A psychological examination takes place before the intervention to evaluate the presence of autistic symptoms. This diagnosis is given based on scores obtained at Autism Diagnostic Observation Schedule (ADOS), corroborated by the scores form Social Communication Questionnaire (SCQ) and a previous diagnosis. The children with scores that are not in a clinical range are excluded. Table I shows the demographics of five of the recruited participants. Experiments were conducted in a lab where therapists from different organizations and institutions are recruited to provide psychotherapeutic services to children with ASD. Both therapist-based interventions and robot-based interventions were conducted for each child. The sequence order of therapist-based interventions and robot-based interventions is randomized.

Each intervention presented to the child begins with the partner's verbel instructions. For example, the instructions for the imitation, joint attention, and turn taking interventions are "Do it like me!","Please, pay attention to what I am looking!","First is your turn. What's your favorite .../Now is my turn.", respectively. In the imitation intervention, the system will check if the child imitates the partner's actions correctly and recognize the child's positive emotions (happy or neural) or negative emotions (angry, fear or sad) in order to assess their engagement. It should be noted that the partner is not required to perform a specific expression in this intervention. In the joint attention intervention, the partner will indicate one of the two objects on the table by looking at and pointing to the object. The gaze of the Nao robot is represented by its head pose since its eyes always look forward. Then, the system will detect whether the child pays attention to the same object or not by using the gaze estimation result and hand movement detection result. In the turn taking intervention, the system will check if the child makes eye contact with the partner and correctly follows the partner's instructions to respect the turns by detecting his/her gaze and hand movement. Readers are suggested to refer to [42] for more details about the three interventions.

(a) (b)

Fig. 9. Gaze estimation results on an ASD child under small head movements. The white line indicates the gaze direction.

Although many algorithms have been proposed for each component, their performance remains unclear when applying to children with ASD. Besides, as shown in table II, the calculation of the behaviour scores is a direct sum operation of the results from each component. For example, in the imitation intervention, the child will receive a high score if the action of the partner is imitated with positive emotions. Thus, this section firstly conducts function level assessments for each component to give an insight view of the system's performance and then provides a behaviour level evaluation by comparing the behaviour scores obtained using the system with the manually annotated ground truth scores. Since the children don't actively perform the speech tasks and only limited data is collected, the audio-related tasks are not evaluated.

A. Gaze Estimation

In this subsection, we report the experimental evaluation of the proposed multi-sensory based gaze estimation on interaction with children with ASD. Fig. 9 and Fig. 10 show some snapshots of the performance of gaze estimation under small head movements and large head movements respectively. The 3D point clouds are constructed by both the Kinect 1 mounted on the middle bar and the Kinect 2 mounted on the top bar. The white line is the estimated gaze of the child. Thanks to the employment of the multi-sensory configuration, it can even deal with the situation when the attention is entirely away from the intervention table.

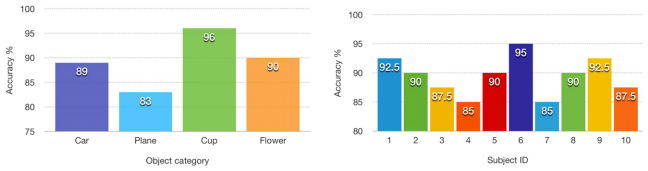
The lack of mutual gaze with social partners is one of the most conspicuous features of ASD. During the interventions,

TABLE II
EVALUATION OF THE CHILDREN'S BEHAVIOUR

Imitation	High Score: The child imitates the movement made by the partner with enthusiasm (positive emotions); Middle Score: The child only finishes part of the behaviour, for example, the movement is imitated without enthusiasm; Low Score: The child does not react or does something else.
Joint Attention	High Score: The child shows something to the partner by using gazing and pointing; Middle Score: The child shows something to the partner by using only part of behaviour; Low Score: The child has no attempts to initiate any joint attention episode.
Turn taking	High Score: The child plays, makes eye contact and respects turns when playing with the partner; Middle Score: The child plays, makes eye contact without considering the partners answers; Low Score: The child does not react or does something else.

(a) (b)

Fig. 10. Gaze estimation results on an ASD child under large head movements. The white line indicates the gaze direction.



(a) (b)

Fig. 11. Performance of multi-sensory based gaze estimation algorithm in joint attention intervention; (a) recognition performance regarding to the 4 different objects; (b) the recognition performance among different children.

TABLE III
MUTUAL GAZE DETECTION PERFORMANCE

Predicted \ Actual	Positive	Negative
Positive	86.53%	13.47%
Negative	5.13%	94.87%

the children are guided by the partner's gaze instructions to look at the robot or human therapist for several times. This system detects mutual gaze by checking if the gaze vector has passed the predefined head area. In the experiment, we randomly extract 1500 images from the recorded videos which consist of 7 children with ASD with individual sample ranging from 98 to 311. Table III shows the mutual gaze recognition performance. The overall accuracy of mutual gaze detection is around 90.7%, which is a good performance considering the large head movement challenge. The performance of mutual gaze detection is affected by the eye center localization algorithm, the gaze estimation algorithm and the head pose estimation algorithm.

The skill to perform joint attention, which is the shared focus of two individuals on an object, plays a critical role in the social development. The impaired development of the joint attention skill is also a prominent feature of children with ASD. Many researchers claim that children with ASD might gain significant maintained improvement in social skills via early intervention therapy of joint attention [43]. During the joint attention section, the partner will look at one of the objects on the table and wait for the child's response. The adopted objects includes a toy car, a toy plane, a cup and a toy flower. Each time, the therapist puts two of the objects on the table ahead and then the partner starts to give indications to the child sitting in front of the table. For the evaluation of the gaze estimation algorithm, we extract 10 successful joint attention samples for each object and each child. The constructed dataset has 10 children and 4 objects, thus the total size of the samples is 400. The algorithm treat it as a successful joint attention if the child's gaze passes through the area of the detected object rectangle.

Fig.11 shows the performance of multi-sensory based gaze estimation algorithm during the joint attention interventions. The proposed algorithm achieves an average recognition rate of 89.5%. The cup object achieves high recognition rate of 96%, which is mainly due to the big size of the object. On the other hand, the small car object achieves a relatively low recognition rate of 83%. In terms of the recognition performance on each child, the proposed algorithm has an maximum accuracy of 95% and a minimum accuracy of 85%.

B. Facial Expression Recognition

The face database is created by manually extracting some frames from recorded videos, which includes 437 images of 5 emotional categories (Angry, Fear, Happy, Neutral and Sad). This dataset contains 7 children with ASD and it is labelled by looking at the video sequence around the specific image to infer the currently status. Fig. 12 shows some snapshots of the performance of the recognition results. It can be intuitively seen that this method has tolerance to small head poses and occlusions.

For facial expression recognition on children's database, the performance is shown in Fig. 13. The overall recognition rate is 63.71%. It is challenging to achieve a clear partition of negative facial expressions as the child tends to perform a combination of emotions (most frequently a combination of fear and angry which is hard to distinguish even by human beings).

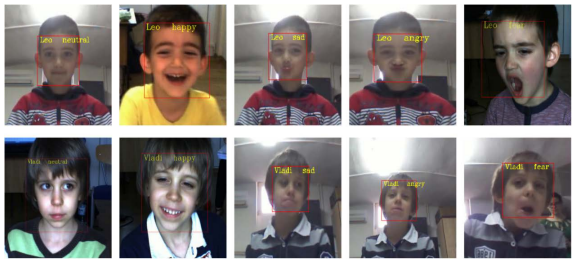


Fig. 12. Snapshots of expression recognition results.

	Neutral	Angry	Fear	Happy	Sad
Neutral	**0.5778**	0.1765	0.0415	0.1107	0.0934
Angry	0.2035	**0.5196**	0.0536	0.1161	0.1071
Fear	0.1509	0.0943	**0.4906**	0.1509	0.1132
Happy	0.0491	0.0552	0.0773	**0.7796**	0.0387
Sad	0.0636	0.0909	0.1515	0.1060	**0.5879**

Fig. 13. Expression recognition results on children ASD.

C. Action Recognition

The action database contains 13 participants with various clothing colours and body sizes. It has 11 actions defined specifically for the imitation interventions, as shown in Fig. 8. Each action is repeated for three times. Thus, the data set contains 429 action segments in total. The starting time and ending time of the actions are manually extracted for the evaluation purpose.

Fig. 14 shows the confusion matrix of our 3DMTG method for the 11 actions. It can be seen that most actions can be correctly recognised by over 80% accuracy. Especially, actions such as *wave hands* and *open arm*, can be 100% recognised by the proposed descriptor because they are simple and have little confusion with the other actions. Actions like *hands on eyes* and *hands over head* are easily confused with each other due to their similar skeleton movement. The recognition accuracies of actions (e.g. *complex 1* and *complex 2*) with large intra-class variations are relatively low.

For the online action recognition scenario where no manually labelled information is available, we used a sliding window strategy with a window size of 26 frames. As a result, the classifier can produce a recognition result for each frame except for the first 25 skeleton frames.

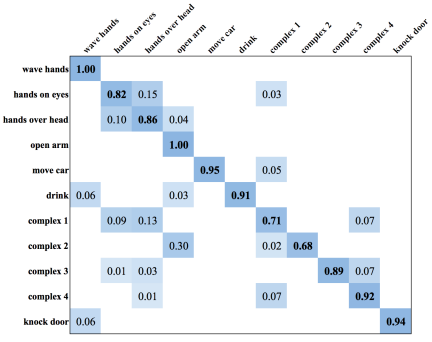


Fig. 14. Actions recognition results on children with ASD.

TABLE IV
TIME PERFORMANCE OF THE SET SYSTEM

Algorithm	Time (ms)
Face alignment	22.13
Gaze estimation	13.85
Facial expression recognition	34.31
Action recognition	9.82
Object tracking and recognition	32.56

D. Time Performance

This subsection provides a speed evaluation of the SET system. The SET system achieves a real-time performance for data sensing and interpretation using the HP Z420 computer (Intel Xeon E5-1650 processor). Due to the heavy load in data transferring, each sensor has to be plug-in to a separate USB controller and the recorded data needs to be saved to a solid-state drive to ensure the real-time performance. By running the system and logging the time cost for each algorithm for 100 times, we obtained the average processing time of each algorithm.

As shown in Table IV, the action recognition has a low computational cost and thus is directly integrated into the Kinect 1 thread. During the experiment, it is found that the build-in speech recognition function consumes less than 0.1 ms. This might due to that a hidden thread, which continuously performs the recognition task, is constructed by the Kinect SDK's speech recognition model. The rest algorithms such as the face alignment algorithm, the gaze estimation algorithm, the facial expression recognition algorithm and the object tracking and recognition algorithm consume more computational time. By packaging them together into the externally invoked interface function, the overall performance of data sensing and recording remains at 25 fps. The drawback is that the analysed information can only be accessed at 10 fps.

E. Assessment of the SET System

To evaluate the performance of the SET system in the imitation, joint attention and turn taking interventions, 70 videos have been extracted with manually labelled scores according to Tabel II. The dataset contains 30 imitation sections, 30 joint attention sections and 20 turn taking sections. The extracted videos have no overlap with the training dataset of each component. The recognition results of different components are directly used to obtain the final behaviour score. Table V shows the confusion matrix of the recognition results. It can be seen from the table that the system achieves 88.24% in identifying low score behaviours. This might due to the fact that it is easy to measure the situations where children do unrelated activities or simply don't respond to the indications. The recognition performance of high score behaviours is lower compared to low score behaviours. The overall recognition performance of the system achieves 82.86% since the low score and middle score behaviours occupy a relatively large proportion (3.7:1) in the testing datasets.

TABLE V
ASSESSMENT OF THE SET SYSTEM

Predicted \ Ground Truth	Low Score	Middle Score	High Score
Low Score	88.24%	11.76%	0%
Middle Score	21.74%	71.74%	6.52%
High Score	5.89%	29.41%	58.82%

[1] The ground truth of the scores are manually labelled and the predicted scores are calculated automatically according to table II.

V. CONCLUSION

This paper made an attempt to improve the existing systems of both standard and robot assisted therapy for children with ASD via a sensing framework with multi-sensory configuration and fusion. The developed SET system has enhanced sensing and interpretation abilities in comparison with the state of the art in the behaviour assessment of children with ASD. Significant contributions comprising acquiring, fusing, and interpreting sensory spatio-temporal multi-modal data have been addressed in the SET system. Experimental evaluations have demonstrated that the SET system is able to effectively perceive the children's behaviour components such as gaze, facial expression and actions, further assessing the behaviour children with ASD.

Future research has been targeted as follows: 1) Experiments for automatic assessment with more children with ASD; 2) Expansion of the activities associated with the intervention table to a multi-task multi-scene smart environment; 3) Evaluation of the effectiveness of SET on improving specific social skills of children with ASD, further gaining more practical social skills.

REFERENCES

[1] F. Edition, *Diagnostic and statistical manual of mental disorders*. American Psychiatric Publishing, Arlington, VA., 2013.

[2] E. Fombonne, "Epidemiology of pervasive developmental disorders," *Pediatric Research*, vol. 65, no. 6, pp. 591–598, 2009.

[3] D. L. Christensen, D. A. Bilder, W. Zahorodny, S. Pettygrove, M. S. Durkin, R. T. Fitzgerald, C. Rice, M. Kurzius-Spencer, J. Baio, and M. Yeargin-Allsopp, "Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network," *Journal of Developmental and Behavioral Pediatrics*, vol. 37, no. 1, pp. 1–8, 2016.

[4] S. Eldevik, R. P. Hastings, J. C. Hughes, E. Jahr, S. Eikeseth, and S. Cross, "Meta-analysis of early intensive behavioral intervention for children with autism," *Journal of Clinical Child & Adolescent Psychology*, vol. 38, no. 3, pp. 439–450, 2009.

[5] R. A. Matthews, S. M. Booth, C. F. Taylor, and T. Martin, "A qualitative examination of the work–family interface: Parents of children with autism spectrum disorder," *Journal of Vocational Behavior*, vol. 79, no. 3, pp. 625–639, 2011.

[6] M. L. Ganz, "The lifetime distribution of the incremental societal costs of autism," *Archives of Pediatrics & Adolescent Medicine*, vol. 161, no. 4, pp. 343–349, 2007.

[7] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, "The clinical use of robots for individuals with autism spectrum disorders: A critical review," *Research in Autism Spectrum Disorders*, vol. 6, no. 1, pp. 249–262, 2012.

[8] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, and Z. E. Warren, "Design, Development, and Evaluation of a Noninvasive Autonomous Robot-Mediated Joint Attention Intervention System for Young Children With ASD," *IEEE Trans. Human-Machine Systems*, pp. 1–11, 2017.

[9] D. François, S. Powell, and K. Dautenhahn, "A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking," *Interaction Studies*, vol. 10, no. 3, pp. 324–373, 2009.

[10] A. Klin, D. J. Lin, P. Gorrindo, G. Ramsay, and W. Jones, "Two-year-olds with autism orient to non-social contingencies rather than biological motion," *Nature*, vol. 459, no. 7244, pp. 257–261, 2009.

[11] L. Boccanfuso, S. Scarborough, R. K. Abramson, A. V. Hall, H. H. Wright, and J. M. O'Kane, "A low-cost socially assistive robot and robot-assisted intervention for children with autism spectrum disorder: field trials and lessons learned," *Autonomous Robots*, vol. 41, no. 3, pp. 637–655, 2017.

[12] R. Simut, C. A. Costescu, J. Vanderfaeillie, B. Van de Perre, Greet Vanderborght, and D. Lefeber, "Can you cure me? children with autism spectrum disorders playing a doctor game with a social robot," *International Journal of School Health*, vol. 3, no. 3, 2016.

[13] F. Sartorato, L. Przybylowski, and D. K. Sarko, "Improving therapeutic outcomes in autism spectrum disorders: Enhancing social communication and sensory processing through the use of interactive robots," *Journal of Psychiatric Research*, vol. 90, pp. 1–11, 2017.

[14] B. Scassellati, H. Admoni, and M. Matarić, "Robots for use in autism research," *Annual Review of Biomedical Engineering*, vol. 14, pp. 275–294, 2012.

[15] S. Thill, C. A. Pop, T. Belpaeme, T. Ziemke, and B. Vanderborght, "Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook," *Paladyn*, vol. 3, no. 4, pp. 209–217, 2012.

[16] C. Wong, S. L. Odom, K. A. Hume, A. W. Cox, A. Fettig, S. Kucharczyk, M. E. Brock, J. B. Plavnick, V. P. Fleury, and T. R. Schultz, "Evidence-Based Practices for Children, Youth, and Young Adults with Autism Spectrum Disorder: A Comprehensive Review," *Journal of Autism and Developmental Disorders*, vol. 45, no. 7, pp. 1951–1966, 2015.

[17] T. Field, T. Field, C. Sanders, and J. Nadel, "Children with autism display more social behaviors after repeated imitation sessions," *Autism*, vol. 5, no. 3, pp. 317–323, 2001.

[18] J.-J. Cabibihan, H. Javed, M. Ang, and S. M. Aljunied, "Why robots? A survey on the roles and benefits of social robots in the therapy of children with autism," *Int. Journal of social robotics*, vol. 5, no. 4, pp. 593–618, 2013.

[19] C. A. Pop, A. C. Petrule, S. Pintea, A. Peca, R. Simut, B. Vanderborght, and D. O. David, "Imitation and Social Behaviors of Children with ASD in Interaction with Robonova. A Series of Single Case experiments," *Transylvanian Journal of Psychology*, vol. 14, no. 1, 2013.

[20] B. A. Taylor and H. Hoch, "Teaching children with autism to respond to and initiate bids for joint attention," *Journal of Applied Behavior Analysis*, vol. 41, no. 3, pp. 377–391, 2008.

[21] T. L. Stanton-Chapman and M. E. Snell, "Promoting turn-taking skills in preschool children with disabilities: The effects of a peer-based social communication intervention," *Early Childhood Research Quarterly*, vol. 26, no. 3, pp. 303–319, 2011.

[22] J. Wainer, K. Dautenhahn, B. Robins, and F. Amirabdollahian, "Collaborating with Kaspar: Using an autonomous humanoid robot to foster cooperative dyadic play among children with autism," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2010, pp. 631–638.

[23] H.-L. Cao, G. Van de Perre, J. Kennedy, E. Senft, P. G. Esteban, A. De Beir, R. Simut, T. Belpaeme, D. Lefeber, and B. Vanderborght, "A personalized and platform-independent behavior control system for social robots in therapy: development and applications," *IEEE Trans. Cognitive and Developmental Systems*, vol. 8920, no. c, p. 1, 2018.

[24] P. G. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H. Cao, M. Coeckelbergh, C. Costescu, D. David, A. D. Beir, Y. Fang, Z. Ju, J. Kennedy, H. Liu, A. Mazel, A. Pandey, K. Richardson, E. Senft, S. Thill, G. V. D. Perre, B. Vanderborght, D. Vernon, and H. Yu, "How to Build a Supervised Autonomous System for Robot-Enhanced Therapy for Children with Autism Spectrum Disorder," *Paladyn, Journal of Behavioral Robotics*, pp. 18–38, 2017.

[25] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "EYEDIAP: A Database for the Development and Evaluation of Gaze Estimation Algorithms from RGB and RGB-D Cameras," in *Proc. ACM Symp. Eye Tracking Research and Applications*, 2014, pp. 255–258.

[26] G. Metta, P. Fitzpatrick, and L. Natale, "YARP – Yet Another Robot Platform," *International Journal of Advanced Robotic Systems*, vol. 3, no. 1, p. 8, 2006.

[27] Y. Liao, Y. Sun, G. Li, J. Kong, G. Jiang, D. Jiang, H. Cai, Z. Ju, H. Yu, and H. Liu, "Simultaneous calibration: a joint optimization approach for multiple kinect and external cameras," *Sensors*, vol. 17, no. 7, p. 1491, 2017.

[28] X. Zhou, H. Cai, Y. Li, and H. Liu, "Two-Eye Model-Based Gaze Estimation from A Kinect Sensor," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2017, pp. 1646–1653.

[29] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[30] X. Xiong and F. De La Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[31] H. Cai, B. Liu, J. Zhang, S. Chen, and H. Liu, "Visual focus of attention estimation using eye center localization," *IEEE Systems Journal*, vol. 11, no. 3, pp. 1320–1325, 2017.

[32] L. S. Dementhon, Daniel F and Davis, "Model-based object pose in 25 lines of code," *Int. Journal of Computer Vision*, 1995.

[33] H. Cai, X. Zhou, H. Yu, and H. Liu, "Gaze Estimation Driven Solution for Interacting Children with ASD," in *Proc. Int. Sym. Micro-Nano Mechatronics and Human Science*, 2015, pp. 1–6.

[34] S. Griffiths, C. Jarrold, I. S. Penton-Voak, A. T. Woods, A. L. Skinner, and M. R. Munafo, "Impaired Recognition of Basic Emotions from Facial Expressions in Young People with Autism Spectrum Disorder: Assessing the Importance of Expression Intensity," *Journal of Autism and Developmental Disorders*, pp. 1–11, 2017.

[35] Y. Wang, H. Yu, J. Dong, B. Stevens, and H. Liu, "Facial expression-aware face frontalization," in *Proc. Asian Conf. Computer Vision*, 2016, pp. 375–388.

[36] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Proc. IEEE Int. Conf. Computer Vision*, 2011, pp. 2106–2112.

[37] I. Matthews and S. Baker, "Active appearance models revisited," *Int. Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[38] B. Liu, H. Yu, X. Zhou, D. Tang, and H. Liu, "Combining 3D Joints Moving Trend and Geometry Property for Human Action Recognition," in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 2016, pp. 332–337.

[39] X. Zhou, H. Yu, H. Liu, and Y. Li, "Tracking multiple video targets with an improved GM-PHD tracker," *Sensors*, vol. 15, no. 12, pp. 30 240–30 260, 2015.

[40] K. Gotham, A. Pickles, and C. Lord, "Standardizing ADOS scores for a measure of severity in autism spectrum disorders," *Journal of autism and developmental disorders*, vol. 39, no. 5, pp. 693–705, 2009.

[41] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter, "The Autism Diagnostic Observation ScheduleGeneric: A standard measure of social and communication deficits associated with the spectrum of autism," *Journal of autism and developmental disorders*, vol. 30, no. 3, pp. 205–223, 2000.

[42] Dream. [Online]. Available: https://www.dream2020.eu/

[43] J. Bradshaw, A. M. Steiner, G. Gengoux, and L. K. Koegel, "Feasibility and effectiveness of very early intervention for infants at-risk for autism spectrum disorder: A systematic review," *Journal of Autism and Developmental Disorders*, vol. 45, no. 3, pp. 778–794, 2015.