Identifying Engagement from Joint Kinematics Data for Robot Therapy Prompt Interventions for Children with Autism Spectrum Disorder

Bi Ge, Hae Won Park, and Ayanna M. Howard(✉)

School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, Georgia
ayanna.howard@ece.gatech.edu

Abstract. Prompts are used by therapists to help children with autism spectrum disorder learn and acquire desirable skills and behaviors. As social robots are more regularly translated into similar therapy settings, a critical part of ensuring effectiveness of these robot therapy system is providing them with the ability to detect engagement/disengagement states of the child in order to provide prompts at the right time. In this paper, we examine the various features related to body movement that can be utilized to define engagement levels and develop a model using these features for identifying engagement/disengagement states. The model was validated in a pilot study with child participants. Results show that our engagement model can achieve a recognition rate of 97 %.

Keywords: Robot therapy · Special needs · Kinematic assessment

1 Introduction

According to the Centers for Disease Control and Prevention, it is estimated that approximately 1 in 68 children in the US are diagnosed with autism spectrum disorder (ASD). For children with ASD, early intervention has been shown to be critical as the younger a child enters an early intervention program, the larger gain (s)he may have in developmental skills [1]. Furthermore, it has been shown that the effects of intervention (including acquisition of new skills) increases as the duration of treatment increases without diminishing returns [2]. However, therapy services and interventions offered by professionals are often expensive or inaccessible [3].

To increase the availability of intervention services offered to children with ASD, alternative technologies have been evaluated for their efficacy in the therapy setting in recent years. One such technology involves the inclusion of social robots with children with ASD. Examples of such systems include humanoids used in [4] as both therapist and interactive toy, designed to help children with ASD learn and practice social skills.

In the traditional therapy setting, prompts are a method of intervention in which cues or instructions are issued before or after a child's action in an effort to reengage his/her attention and help gain/eliminate desired/undesired behaviors [5]. For example, a therapist may issue prompts when a child loses his or her temper or stops concentrating on the task at hand during a therapy session. Prompts are essential since children

A. Agah et al. (Eds.): ICSR 2016, LNAI 9979, pp. 531–540, 2016.
DOI: 10.1007/978-3-319-47437-3_52

with ASD usually does not respond to social cues in the same manner as typically-developing children do. Therapists thus utilize prompts as extra stimulus that correspond to some particular response [5]. In [6], it was shown that when interacting with a therapist and a robot at the same time, children with ASD and typically-developing children both spend more time looking at the robot than the human therapist. As this is one of the primary conditions for providing effective prompts, it seems appropriate that an autonomous robot platform for ASD intervention should be able to use this type of intervention method. For example, as pointed out by [7], robots can provide consistent, repeatable and standardized stimuli. This inherent characteristics of a robotic system would also enable a robot to provide consistent, repeatable and standardized prompts.

In this paper, we look at the first step required to provide prompts - the ability to detect engagement/disengagement states of the child in order to provide prompts at the right time. In our experiment, we define engagement as "concentrating on the task at hand and willing to remain focused". We discuss our method for modeling engagement levels based on features extracted from body movements, namely we focus on detecting if a child is concentrating on a given task by extracting engagement levels from joint kinematics data. We show that by carefully selecting features from joint kinematics data, we can achieve good performance for detecting engagement levels.

2 Related Work

There has been a number of prior research efforts focused on developing algorithms to detect user disengagement or engagement states. Some features used by past research efforts include: body posture [8], gestures [8], facial expression [9], eye gaze [8, 9], EEG [10], contextual information [9] and spatial relationship between the robot and human [11].

In [8, 9], the feature sets included eye gaze directions, gestures and head directions, which were extracted from human observations. In such a scenario, a robotic system would need to employ modeling methods that can run autonomously, without any human intervention. In [9], researchers also discussed using contextual information obtained from the log file of a storytelling app on a tablet. This kind of information though may not always be available to the robot, since robot interventions should occur in real-time, during the therapy session, rather than after the child has completed the task.

With respect to EEG applications such as in [10], even though EEG is considered noninvasive, placing electrodes on some children with ASD might be intolerable. In addition, using EEG devices limits the naturalness of the session and thus behaviors learned in the session may not be transferrable to the child's natural environment.

The spatial model in [11] uses the relative location of robot and human for a receptionist robot to detect engagement levels. However, in the case of a therapy robot for children with ASD, relative location information gives very little information about the engagement level of the child since children are typically engaged within a local zone of proximity to the therapist during the therapy session.

Beyond the ones mentioned, there are a number of other challenges associated with the process of selecting features that enable the modeling of engagement levels in children with ASD. Using eye gaze to analyze engagement, such as in [12] and in the ASD study by [13], requires the tracked face to be directly aligned towards the sensor. In our previous studies [14], we have seen that in most therapy sessions, the face is not always orientated toward the sensor, and, in fact, is just as likely to be orientated toward the task, the therapist, or, in some cases, towards the floor/table.

Using voice and verbal recognition has also been shown as a feasible option, such as in [15], where acoustic features from speech were used to model and detect engagement levels in daily conversations. However, depending on the therapy, a child with ASD may or may not talk during the session and certain children with ASD are nonverbal.

Other work, such as in [16], monitors user input through tablet interaction in order to assess engagement. However, children with ASD may not always be interacting with a tablet during a therapy session, thus features for modeling engagement should be extracted from interactions outside the tablet.

3 Approach

In this paper, we discuss our approach for detecting child engagement and disengagement states. Our approach for modeling engagement levels is based on extracting features from body movements, namely we focus on detecting if a child is concentrating on a given task by extracting engagement levels from joint kinematics data. In this work, we utilized a RGB-D camera, namely the Kinect 2.0 by Microsoft, to detect a user's skeleton and analyze the skeleton's pose relative to the task. This provides decent objective measures to detect engagement levels. Another benefit of using the Kinect/RGB-D camera is that the camera can recognize the human skeleton without requiring attaching markers to the body, which may be intolerable to children with developmental challenges.

A typical autism therapy session usually involves a therapist, a child (patient) and a task. While the child works on the task, the therapist provides instructions or prompts to help the child complete the task. To mimic a typical therapy session, the task we employ is a turn-taking game played on a Samsung Android tablet involving the matching of cards (Fig. 1). During the therapy session, a Kinect camera mounted on a tripod is located in front of the therapist and child in order to capture as much movement associated with their interaction as possible. There is also another camera in the room which serves as a backup source and provides a different viewing angle of the session.

3.1 Body Movement Features

Given a typical therapy session, our first step is to extract the relevant features, which can be used to identity engagement state. As skeletal data is directly extracted from the RGB-D data set and thus represents the movement profiles associated with the

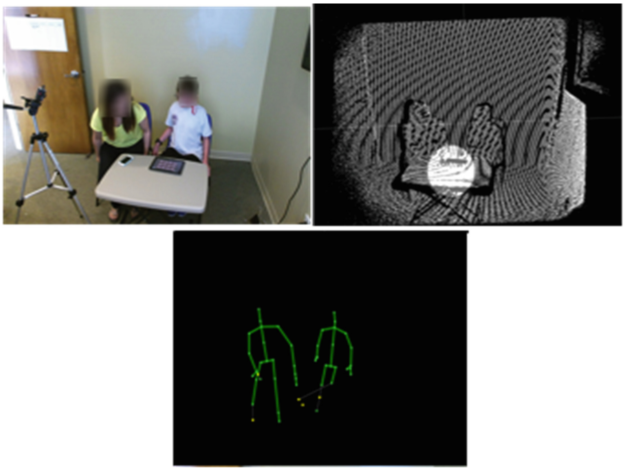

Fig. 1. Left: Child interaction session; Right/Bottom: Imaging and resulting skeleton of session using RGB-D camera

interaction, we first define a number of relevant features that can be derived from skeletal data. For this work, we classify these features as leaning angle, planar distance to therapist, mean joint to joint distance, distance of joints traveled within task ball, mean joint coordinates, mean joint distance to task and mean joint to joint distance.

a. Leaning angle

The leaning angle is the angle between the vertical y-axis in the camera's view and the vector constructed from the midpoint of the spine and neck of the child. Leaning angle is chosen to represent the scenario correlated with an individual concentrating on a task. In such cases, it has been observed that individuals tend to lean towards the object of interest associated with achieving a task. For example, in our experimental case, the tablet device becomes the object of interest. The leaning angle θ for the spine joint vector $\vec{j}_{midspine}$ and neck joing vector $\vec{j}_{neck}$ is calculated as:

$$\theta = \cos^{-1}\left(\frac{\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \cdot \left(\vec{j}_{midspine} - \vec{j}_{neck}\right)}{\left\| \begin{matrix} 0 \\ 1 \\ 0 \end{matrix} \right\| \left\|\vec{j}_{midspine} - \vec{j}_{neck}\right\|}\right) \quad (1)$$

b. Planar distance to therapist

The planar distance to therapist feature is one measure used to calculate the distance measured between two people interacting with a task's common object of interest. A plane is constructed by using the middle point of the spine, neck and head of the therapist. Distances between this plane and the child's skeleton joints are then calculated. For the therapist mid-spine joint vector $\vec{j}_{midspine}$, neck $\vec{j}_{neck}$ and head $\vec{j}_{head}$, the plane P and planar distance D can be derived using the following equations:

$$P = \left(\vec{j}_{midspine} - \vec{j}_{neck}\right) \times \left(\vec{j}_{midspine} - \vec{j}_{head}\right) \tag{2}$$

$$D = \frac{\vec{c}_i \cdot \mathrm{P}}{\vec{c}_i} \text{for child joint vector } \vec{c}_i \tag{3}$$

c. Mean joint to joint distance

The mean joint to joint distance records the average distance between each joint of the therapist and each joint of the child. This feature reflects the relative pose and distance between the therapist and child during an interactive session. For each child's joint i and their joint vector $\overrightarrow{c_i^f}$ and therapist's joint vector $\overrightarrow{t_i^f}$ at the f th frame of F frames, where F is the number of recorded Kinect frames associated with a movement profile, the mean joint to joint distance d_i is determined as:

$$d_i = \frac{1}{F}\sum\nolimits_{f=0}^{F} \left\|\overrightarrow{c_i^f} - \overrightarrow{t_i^f}\right\| \tag{4}$$

d. Distance traveled within task ball

The distance traveled feature measures the distance traveled by each joint inside a sphere whose center is located at the task's object of interest and has radius r. For a sphere with radius r and centered at $\vec{b}$, if joint vector $\vec{j}_i^f$ at the f th frame satisfies $\vec{b} - \vec{j}_i < r$, then $d_t = \sum\nolimits_{f=1}^{F} \left\|\vec{j}_i - \vec{j}_{i-1}\right\|$

e. Mean joint coordinates

The mean joint coordinates, associated with either an engagement or disengagement state, reflects the absolute pose of the child during a therapy session measured in the 3D world. It is obvious that in order to make this feature meaningful across various intervention sessions, it has to be normalized to eliminate overfitting to a specific session setup (for example how far the child sits away from the camera should not effect the performance of the system). As such, for each joint vector $\overrightarrow{j_i^f}$ and F frames, the mean joint coordinates m_i is calculated as:

$$m_i = \frac{1}{F}\sum\nolimits_{f=0}^{F} \left\|\overrightarrow{j_i^f}\right\| \tag{5}$$

f. Mean joint distance to task

The mean joint distance to task calculates the average distance between each joint and the location of the task object of interest. This feature reflects how far away a child is from the task during a therapy session. For each child joint vector $\overrightarrow{c_i^f}$ at the f th of F frames, the mean joint distance n_i to task T is:

$$n_i = \frac{1}{F}\sum\nolimits_{f=0}^{F} \left\|\overrightarrow{c_i^f} - \vec{T}\right\| \tag{6}$$

Once determined, these relevant features can then be utilized to identify engagement state. It is worth noting that some of these features require the presence of two skeletons during the therapy interaction, i.e. a child and a therapist/caregiver. However, the robot would not be autonomous if it requires the presence of a professorial therapist. Therefore, to classify engagement state in this paper, we will mainly focus only on those features that are derived solely from the child's skeletal data.

3.2 Classification of Engagement/Disengagement

In this work, we define two states – engagement/engaged and disengagement/disengaged. Engagement is defined as focused or concentration on the task at hand. Disengagement is therefore just defined as the contrasting state, i.e. not engaged. We therefore define our problem as a two-class pattern recognition problem and examine those classification methods that are most appropriate to these types of problem. For this work, we evaluate the performance of three classification methods that can be used to distinguish between the two different states, namely SVM, Random Forest, and AdaBoost [17]. Support vector machines (SVMs) is a supervised learning method in which, given a set of training examples, each sample is marked as belonging to one of two classes. In this paper, our two classes correspond to *Engaged* or *Disengaged*. Given a set of labeled examples, the SVM training algorithm is able to build a model that then assigns new examples to one of the two defined classes. Random forests, on the other hand, represent the classification problem as a group of decision trees in which a new example is first classified based on an input vector that, as it descends down the branches of the tree, gets parsed into smaller and smaller sets. Each tree then provides a classification, and the forest chooses the classification having the most votes (over all the trees in the forest). The other method examined was AdaBoost, which solves the two-class pattern recognition problem by weighting the outputs given a large pool of weak classifiers.

In this work, these methods were trained and tested using a "hold-one-child-out method" based on the Scikit-learn methodology [18]. In the "hold-one-child-out" method, we pick data recorded with one child as the test data, and train the classifier with data from all other children. In this process, the following procedure was followed for each of the classifiers:

- A model using one of the respective training algorithms (SVM, Random Forest and AdaBoost) was trained using clips from all the children except the test one.
- The resulting model was then validated for classification accuracy using the remaining hold out part of the data as the test set.
- This process was repeated for each set of children and the final accuracy value computed as the average of the computed accuracy values for each round.

4 Experimental Setup

The goal of the classifier is to accurately detect engagement/disengagement states of a child in order for a therapy robot to provide prompts at the right time. As such, given our set of possible body movement features and classifiers, our goal was to determine the accuracy rates for each model in order to select the set with the greatest ability to discriminate between states. For this experiment, a pilot study was conducted at the Kid's Creek Therapy Center. The parents of each participant signed the IRB (Institutional Review Board) approved consent form allowing their child to engage in the testing sessions. Children diagnosed with developmental disabilities were recruited for this experiment with 3 boys, $mean(\text{age}) = 12.3$ and $\sigma(\text{age}) = 1.5$ (Table 1).

The child study consisted of sessions where, in each session, the experimenter and the child played a turn-taking game on the tablet (Fig. 2). There was one session hosted per child. During interaction, the experimenter asked a series of questions to distract the child during the child's turn such as "Do you remember my name?" A total of three sessions of approximately 28 min in length were recorded and processed. During the experiment, the real world coordinates of all joints of the human upper body skeleton were recorded as well as color video and audio streams from the Kinect camera. For each participant, a total of 17 components (i.e. joints) were captured by the Kinect. Since the camera was fixed on a tripod across all sessions, the classifier trained using data from one session was able to be applied to another session without normalization. For training, we did not select the features, planar distance to therapist and mean joint to joint distance, since these features required two skeletons to be present. As the goal of a therapy robot is to allow children to receive intervention without a therapist closely present, we determined that our model should only be built from skeletal data based on the child's movement profile. As such, the skeleton was selected manually in order to ensure that the analysis was performed with the child's movement profile and not the therapist's.

Table 1. Demographic data of child participants

Participant	Primary diagnosis	Gender	Age
1	ASD	Male	12
2	Down syndrome	Male	14
3	ASD	Male	11

Lastly, when calculating the feature vectors associated with the child and task, we also needed to know the location of the task's object of interest (i.e. the tablet). To obtain this information, we calculated the real world coordinates of the tablet by first reconstructing the 3D scene using the depth image obtained by the Kinect camera and manually selecting the 3D point corresponding to the tablet in the image scene.

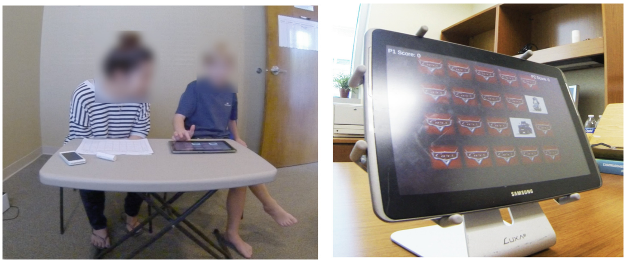

Fig. 2. Left: Experimenter interacting with child with ASD; Right: Turn-taking matching game on the tablet.

5 Results

5.1 Ground Truth

Once collected, the stream of data from the pilot study was annotated by a human annotator with timestamps indicating the start and end of both engagement and disengagement states. Timestamps were annotated based on the identified behaviors in the videos. For example, some typical disengagement behaviors included standing up and walking away from the tablet and talking to others about things unrelated to the session. Disengagement states were also associated to those instances of time when the experimenter asked the series of questions designed to expressly distract the child.

Once the timestamps associated with the start and end times for the different states were obtained, clips were then segmented into smaller ones, each lasting for 2 s with a 1 s overlap. This was done in order to provide us with a sufficient number of training and testing instances to validate the model. From this annotation process, 22, 14 and 47 clips were labeled as disengagement clips for the three sessions respectively and 2,14, and 21 clips were labeled as engagement clips.

Table 2. Percent accuracy of different classifiers with respect to body movement features

Classifier	Mean joint coordinates	Mean joint distance to task	Distance traveled within Task Ball	All three features
SVM	88 %	70 %	70 %	88 %
Random forest	96 %	65 %	60 %	96 %
AdaBoost	96 %	93 %	66 %	97 %

5.2 Performance

Data from the various child interaction sessions were used to evaluate the performance of the different classifiers and feature sets using the “hold-one-child-out” cross validation method as discussed previously. As shown in Table 2, the best performing body movement feature was identified as Mean Joint Distance. Figure 3 expands on the corresponding table results. Based on this assessment, using AdaBoost with Mean Joint

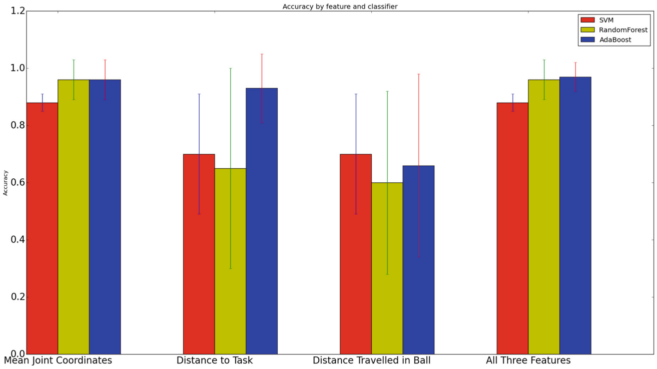

Fig. 3. Mean and standard deviation on performance accuracy associated with classifiers using different feature sets

Coordinates achieved the best single-feature performance at 96 % accuracy while AdaBoost with all the features gives an accuracy of 97 %. This combination appears to provide the best performing results.

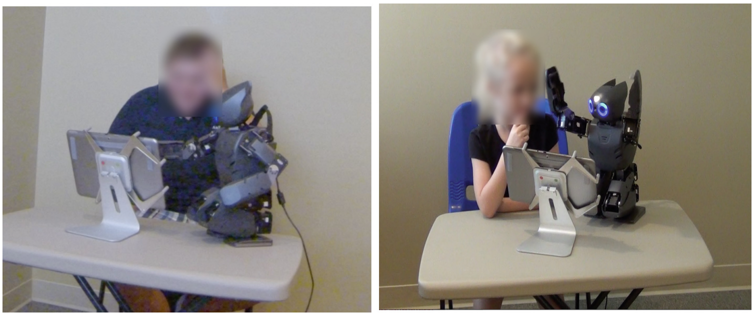

Fig. 4. Children with ASD interacting with therapy robot during a turn-taking matching game on the tablet. In future work, the results from this study will be used to evaluate if the same performance can be achieved for identifying engagement/disengagement in child-robot interaction scenarios such as these.

6 Conclusion and Future Work

We have discussed several features extracted from skeletal data recognized by a RGB-D camera, namely the Microsoft Kinect 2.0 that can be used for detecting engagement and disengagement states during therapy sessions. By carefully selecting the features, we demonstrated that without using contextual or voice features, we can still achieve decent performance.

An extension to this paper would be recruiting more participants and applying the algorithm to a real robot system to validate the engagement/disengagement model in a real robot-child therapy setting (as shown in Fig. 4) and/or compare the performance between the robot system and an actual ASD therapist.

References

1. Corsello, C.M.: Early intervention in autism. Infants Young Child. **18**, 74–85 (2005)
2. Granpeesheh, D., Dixon, D.R., Tarbox, J., Kaplan, A.M., Wilke, A.E.: The effects of age and treatment intensity on behavioral intervention outcomes for children with autism spectrum disorders. Res. Autism Spectr. Disord. **3**, 1014–1022 (2009)
3. Sharpe, D.L., Baker, D.L.: Financial issues associated with having a child with autism. J. Fam. Econ. Issues **28**, 247–264 (2007)
4. Robins, B., Dautenhahn, K., Te Boekhorst, R., Billard, A.: Robot assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills. Univ. Access Inf. Soc. **4**, 105–120 (2005)
5. MacDuff, G.S., Krantz, P.J., McClannahan, L.E.: Prompts and Prompt-Fading Strategies for People with Autism, Making a difference: Behavioral intervention for autism, Austin. TX, Pro-Ed (2001)
6. Bekele, E., et al.: A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children With Autism. IEEE Trans. Neural Syst. Rehabil. Eng. **21**, 289–299 (2013)
7. Eikeseth, S., Smith, T., Jahr, E., Eldevik, S.: Outcome for children with autism who began intensive behavioral treatment between ages 4 and 7. Behav. Modif. **31**, 264–278 (2007)
8. Leite, I., McCoy, M., Ullman, D., Salomons, N., Scassellati, B.: Comparing models of disengagement in individual and group interactions. In: ACM/IEEE International Conference on Human-Robot Interaction, Portland, Oregon, pp. 99–105 (2014)
9. Castellano, G., Pereira, A., Leite, I., Paiva, A., McOwan, P.W.: Detecting user engagement with a robot companion using task and social interaction-based features. In: International Conference on Multimodal Interfaces, Cambridge, Massachusetts, pp. 119–126 (2009)
10. Berka, C., et al.: EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. **78**, 231–244 (2007)
11. Michalowski, M.P., Sabanovic, S., Simmons, R.: A spatial model of engagement for a social robot. In: 9th IEEE International Workshop on Advanced Motion Control, pp. 762–767, Istanbul, Turkey (2006)
12. Nakano, Y.; Ishii, R.: Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In: Proceedings of the 15th International Conference on Intelligent User Interfaces, Hong Kong, China, pp. 139–148 (2010)
13. Bal, E., et al.: Emotion recognition in children with autism spectrum disorders: relations to eye gaze and autonomic state. J. Autism Dev. Disord. **40**, 358–370 (2010)
14. Park, H.W.; Howard, A.: Engaging children in social behavior: interaction with a robot playmate through tablet-based apps. In: Rehabilitation Eng. and Technology Society of North America (RESNA) Annual Conference, Indianapolis, IN, June 2014
15. Yu, C., Aoki, P.M., Woodruff, A.: Detecting User Engagement in Everyday Conversations. arXiv preprint cs/0410027 (2004)
16. Park, H.W., Coogle, R., Howard A.: Using a shared tablet workspace for interactive demonstrations during human-robot learning scenarios. In: IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, June 2014
17. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning. Springer, New York (2001)
18. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)