Child Engagement Estimation in Heterogeneous Child-Robot Interactions Using Spatiotemporal Visual Cues

Dafni Anagnostopoulou[1], Niki Efthymiou[1], Christina Papailiou[2], Petros Maragos[1]

***Abstract*— Robots are increasingly introduced in various Child-Robot Interactions with educational, entertainment or even therapeutic goals. In order to achieve qualitative interactions, robots need to adjust their behavior according to children's response. A robot's ability to successfully estimate partner's engagement is of great importance towards this direction. In this research we propose a method to estimate the engagement level of children during heterogeneous and challenging child-robot interactions. Our method uses the spatiotemporal residual R(2+1)D blocks to simultaneously leverage the rich RGB and temporal information, which is crucial for the engagement estimation. We present results on three different groups of data, including the PInSoRo open dataset, proving our method's robustness and improvement over previous works.**[1]

I. Introduction

Interest in using social robots has increased over the years as more and more studies focusing on improving Child-Robot Interactions (CRI) are performed every year [1]. Social robots have been introduced in the educational process of children [2], [3] and they have also been employed to help children with Autism Spectrum Disorder (ASD) or learning difficulties to tackle consequences and challenges of such disorders [4], [5], [6].

In particular, there are promising results in the use of robots in supporting the social and emotional development of children with ASD [7], [8], [9]. Social robots proved to be a way to get through the social obstacles of children and make them involved in the interaction [1]. During interactions with social robots, children with ASD tend to preserve a calm and active mood and to display repetitive behaviors less frequently [10], [11]. Furthermore, research indicates that attention and engagement towards children's parents increased after a long-term of CRI [12]. These studies' results reveal that interacting with robots could significantly help children with ASD.

Robots' ability to adapt their behavior according to the children's cognitive state is of great importance, so that common ground between robots and children is established [13]. Engagement is a significant indicator of human response to interactions. It refers to a dyadic state constituted by mutual and extended interactions between a child and a partner (human or robot) about a topic in the environment. Engagement captures both partners' contributions to the maintenance of ongoing interaction periods comprised of extended interactional turns, when the child and the partner actively focus on shared objects and events [14].

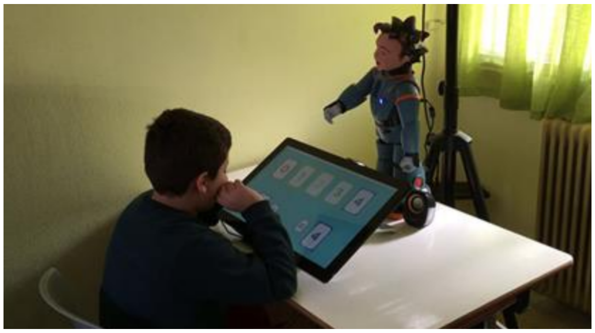

Fig. 1: Child interaction with Zeno robot in a Greek primary school.

Engagement estimation is not a simple task as it poses significant difficulties. First of all, engagement is a multifaceted cognitive mechanism that cannot be directly observed [15]. Moreover, although engagement is an internal mental state, the observing robot stays confined to the exploitation of external vision or audio cues to estimate its level [16]. While recognizing an action can be part of the information needed to estimate engagement, it is not enough as sometimes being fully engaged to an interaction means observing the partner without acting depending on the stage of the interaction.

Our purpose is to develop a reliable method of engagement estimation in diverse child-robot interactions. In our previous work [17], we proposed an engagement estimation method focusing on ASD children, taking part in different interactions both with social robots and with their mothers. Our model based mostly on children poses could successfully be trained to estimate children engagement in different sessions during which children where free to move around the room as they pleased. However, it could not tackle the engagement estimation problem successfully when dealing with interactions during which children where more confined, mainly sitting in front of a desk or a table. Therefore, we aimed at designing a different method based on raw RGB and optical flow data instead of pose, that operates on video clips instead of image frames and employs a different network architecture so that it can accurately estimate children engagement level in as many as possible, given the data we possess, conditions.

In this paper, we propose a method that can estimate engagement of children interacting with robots by using deep learning methods. The network we used for the machine learning experiments is based on the spatiotemporal ResNet

[1] School of ECE, National Technical University of Athens, 15773 Athens, Greece dafnianagno@hotmail.gr, nefthymiou@central.ntua.gr, maragos@cs.ntua.gr

[2] Department of Early Childhood Education and Care, University of West Attica, 12243 Athens, Greece cpapailiou@uniwa.gr

[1] Dafni Anagnostopoulou, Niki Efthymiou, Christina Papailiou, and Petros Maragos, "Child engagement estimation in heterogeneous child-robot interactions using spatiotemporal visual cues," in Proc. IROS, 2022

Proc. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2022), Kyoto, Japan, Oct. 2022.

(2+1)D block [18]. We have experimented with various data in order to be able to evaluate the generalization of our models. First of all, we have designed and developed our method on interactions of ASD children playing with Zeno [19] robot two different games in their school environment. Moreover, we have tested our method in data we used in [17], in order to compare it with previous published results. Finally, we applied our method on a part of the PInSoRo dataset [20]. We consider this very important due to the fact that, to the best of our knowledge, PInSoRo is the biggest publicly available dataset capturing both child-child and child-robot interactions and facilitating data-driven studies of social dynamics and CRI. All these experiments demonstrated that the proposing method can be used for child engagement estimation in completely different scenarios.

II. Related Work

In the past few years, numerous studies approached the problem of child engagement estimation. Earlier studies concentrated on features such as head pose [21], gaze [22], face expression [23], and distance between partners [24]. Hadfield et al. [25], in a previous work of our laboratory, proposed an LSTM neural network that uses skeletal pose, body and head direction as well as distance from the robot partner. Baxter et al. [16] proposed a two stage method for children engagement estimation. Firstly, a convolutional network is employed in order to extract useful representations from the RGB frames. Afterwards, a recurrent network leverages these representations to extract a temporal feature vector. Filntisis et al. [26] designed a method that estimated children's emotional state by combining estimations of two different networks for more accurate results. A ResNet-50 convolutional network that estimated emotional state using the RGB images of children faces and a fully connected neural network that estimated emotional state using children body poses. In [27], the method proposed to estimate engagement is based on thermal infrared imaging.

Some studies have also approached the problem of estimating engagement of children focusing on children with ASD or children with learning difficulties. In Anagnostopoulou et al.[17], we propose a deep convolutional network that receives as inputs children's poses transformed to resemble sequences of images and use it to estimate ASD children's engagement taking part in a variety of conditions and interactions. Rudovic et al.[28] introduced CultureNet for estimating engagement level of children with ASD. The network architecture is based on convolutional ResNet-50 but the training is personalized in different culture backgrounds. Papakostas et al. [29] proposed a method that estimates engagement of children with learning difficulties. Their method extracts various features like body and head orientation, emotion, response time and speaking duration and uses an AdaBoost decision tree to estimate engagement. Moreover, [30] employs the recurrent Legendre Delay Network and uses facial and skeletal landmarks to estimate engagement level on the PInSoRo child-child interactions.

Additionally, other recent studies focus on ASD children aiming to action recognition or emotion recognition. Zhang et al. [31] concentrate on action recognition of ASD children in order to facilitate stereotyped actions recognition process. Their approach employs OpenPose to detect children's poses which are fed to an LSTM neural network in order to recognize children's actions. In [32], the goal is to estimate ASD children's affect state and the method proposed leverages acoustic and visual cues in order to accomplish that. This approach employs speech emotion recognition to distinguish negative affect states and afterwards uses RGB data to distinguish between positive and neural affect states.

As previously shown while describing engagement, engagement estimation cannot be considered as a subproblem to the action recognition problem. However, efforts to estimate engagement can be reinforced by the ongoing progress in wider video recognition problems such as action recognition. Examples, of recent action recognition methods that engagement estimation could take advantage of are Temporal Shift Modules [33], which based on a convolutional network perform temporal shifts along channels in order to facilitate the exchange of information between frames as well as Video Transformer Networks [34], in which a temporal attention encoder is employed in a network for action recognition.

III. Data & Engagement Description

Since we are focusing on developing a robust system to estimate child engagement during challenging interactions with a robot, we use datasets with diverse conditions ans environments. In our experiments, we have been using three different groups of data that differ on: a) the participating children (TD or ASD), b) their posture (if they are seated or not), and c) the extent at which the designed interactions prompt children to cooperate with the robot towards a common goal. All of these, as mentioned before, affect the indications of the engagement level and, apparently, complicate engagement estimation. In Table I, we summarize important differences which have significant impact on engagement estimation across the different sets of data.

The first set of data consists of 13 sessions in which three children participate, one girl and two boys, facing autism spectrum disorder. Each child participated in two intervention sessions per week for three months in the school they studied, the Special School for Children with Autism in Piraeus, Greece. During each session, which lasted approximately five minutes, each child participated and played two games every time with Zeno robot [19]. Zeno was placed on a desk beside a flat touch screen, while children sat in front of the desk. For the recording, a Kinect camera was installed behind the game setup at an angle that captures the child's movements and facial expressions and the progression of the games.

The games that children played were: Sums Game and Emotions Game. During the *Sums Game*, Zeno asks the child to help him learn how to add up to number four. Simple sums appear on the screen, and Zeno and the child take turns to solve them correctly. In the *Emotions Game* Zeno asks the child to express the emotions of happiness, sadness

(a) BR-SCHOOL GAMES DATASET

(b) ASD-GAMES DATASET

(c) PINSORO DATASET

Fig. 2: Instances from the different data sets environments. (a) A BABYROBOT-SCHOOL GAMES Dataset instance, child plays *Sums Game* and helps Zeno robot find the answer to the sum question that appears on the screen. (b)An ASD-GAMES Dataset instance, child plays *Guess the Object Game* (c) A PINSORO Dataset instance, child plays with Nao robot around the touchscreen table.

DATASET	Child Development	Child Kinetic Behavior	Robot Social Behavior
BR-SCHOOL GAMES	ASD	Seated	Social
ASD-GAMES[17]	ASD	Move around	Social
PINSORO[20]	TD	Seated	Asocial

TABLE I: Most important differences among the used datasets.

and fear. In both games, the robots' prompts were graded in three levels, i.e. from more concrete to more abstract for the Sums Game, and from direct imitation of the robot to prompts for spontaneous expressions. We refer to these data as BABYROBOT-SCHOOL GAMES.

We also tested our method in some of the data we experimented with in our previous work [17]. We have seven sessions in which seven children participate facing autism spectrum disorder. During each session, which lasted approximately 20 minutes, each child participated and played four different games with two robots, NAO [35] and Furhat [36]. The games that children played were: *Show me the Gesture*, *Express the Feeling*, *Pantomime* and *Guess the Object*. Each child stood in front of the robots and interacted freely with them while moving in the room as they wanted. For comparison reasons, we refer to these data as ASD-GAMES.

Finally, we also experimented with the PInSoRo dataset [20], which is an open dataset of child social interactions designed to be used in data-driven research efforts. It consists of about 45 hours of social interactions between 45 child-child pairs and 30 child-robot pairs. All interactions are taking place around a large interactive table. Children are encouraged to play freely and are not directed to perform any particular task. Interactions are annotated on three different axis: task engagement, social engagement and social attitude. For our experiments, we have been using the child-robot interactions as we are mostly interested in empowering social robots to adjust their behavior according to children's engagement level. We have chosen not to use interactions with low inter-coder agreement for task engagement annotations. We have been using 23 out of the 30 child-robot interactions and specifically we cropped interactions between the 5th and the 15th minute. This time period was chosen because the density distribution of the duration is centered around 15 minutes [20]. For comparison reasons, for the rest of this text, we refer to these data as PINSORO dataset.

Engagement is expressed as a dynamic, multimodal, and temporally organized action generated in and referring to a socio – cultural context [37], [38]. It is a reciprocally motivated enactment involving whole body communication of agency based on intuitively regulated temporal contours of expressive sound and movement. Its coherence is based on the ways its parts are related temporarily and casually. Each partner monitors the other's attention and adapts his action accordingly, while maintains or exchanges perceiver – actor roles [39], [38]. Thus, from a psychological point of view we cannot monitor engagement by separately monitoring behavior towards the goal and social behavior towards the partner, although in the future we could explore this approach computationally in order to draw further conclusions. According to these we define three engagement levels:

- level 1: The child is disengaged,meaning they are paying limited or no attention to the robot or their common goal.
- level 2: This level is regarded when children pay attention to the robot but remain passive or act relative to the common goal but not pay attention to the partner.
- level 3: The child is engaged, referring that the child acts harmonically with the robot to complete their common goal.

Members of our laboratory annotated the data according to a set of instructions containing groups of visual and acoustic cues under psychologists supervision. These instructions can be found in our previous work [17]. The videos were annotated not based on a specified time interval but on the change of the engagement level with one second accuracy, while each one was annotated by one annotator.

For the PInSoRo dataset, four different engagement levels are used on the task engagement axis, close to our engagement annotations: no-play, adult-seeking, aimless and goal-oriented. By excluding the adult-seeking engagement level, which is extremely rare (about 2.5% for the particular sub-dataset we have been using), the engagement annotation schema of all the used datasets are in agreement.

IV. METHOD

A. Network Architecture

Previous research has shown that engagement estimation can be estimated adequately using children's pose. However, our empirical studies have shown that this approach is successful in scenarios where children are free to move around

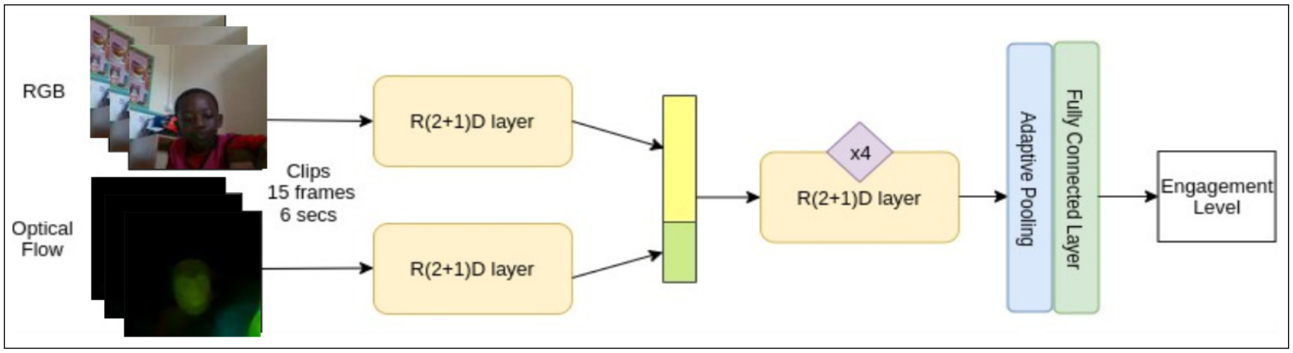

Fig. 3: Network for engagement estimation. Network receives as input a clip of RGB and optical flow data estimates the child's engagement level.

the room and therefore their pose (especially relative to the robot) presents significant variations. On the other hand, it is not successful when applied on interactions during which children are seated and thus their poses present much less variety and many parts of the children bodies are occluded. To tackle this, we need to leverage RGB information which contains much richer information.

Simultaneously, one other empirical finding is the fact that engagement heavily depends on the progress of the ongoing interaction. For example, during the *Sums Game* there are instances that the child needs to wait for the robot to answer the question. Waiting on these particular instances reveals higher level of engagement than immediately answering. These kind of parts of the interactions are really difficult to be correctly estimated. This is the reason why we have decided to estimate engagement on clips instead on separate frames. We split the videos into parts with constant engagement level and afterwards we split each part into clips with duration of 6 seconds each. We extract optical flow and we estimate depth data from RGB data. Video frames are scaled to 228× 128. Using a face detector, we crop frames into square images of size 112× 112 around the children faces.

We propose a network that is built using the spatiotemporal convolutional block R(2+1)D [18] which has been introduced for action recognition. This block is made of spatiotemporal convolutions which perform a two dimensional convolution over the spatial plane followed by a one dimensional convolution along the time axis. These spatiotemporal convolutions are repeated twice and combined in a residual block. Our network consists of five consecutive layers created from these spatiotemporal blocks. The spatiotemporal convolutional layers are followed by an adaptive pooling layer and finally fully connected layer.

We have been experimenting with different training data. We use raw RGB data, a fusion of RGB and depth data as well as a fusion of raw RGB and optical flow data. We experiment with fusing rgb and optical flow networks during a late or an early network stage. On the first occasion, we fuse the two channels before the last fully connected layer. On the second, we fuse the two channels after the first of the five spatiotemporal convolutional layers. The proposed architecture with early fusion of RGB and optical flow channels is depicted in Fig.3.

B. Implementation

We use PyTorch library [40] to implement our network. We use the ResNet (2+1)D pretrained on the Kinetics-400 dataset. We chose a batch size of 16 and a learning rate of 10^{-4} after experimenting with different values. We also used the Adam Optimizer [41] as well as ReduceLROnPlateau scheduler in order to update network weights and decrease learning rate when our metrics do not improve for 10 continuous epochs.

We create train and validation sets for our four data groups separately and we apply cross validation. In order to avoid overfitting danger as well as produce more training examples for the less common classes, we employ some methods of data augmentation. We create clips that consist of 15 frames that spread along six seconds. Thus, we use one every 12 frames given a fps of 30. For every six second interval we use different frames to produce different clips. As far as the PInSoRo dataset is concerned, this data augmentation method is not employed for the 'goal-oriented' engagement level which is by far the most common in our data, so that the imbalanced distribution between classes mitigates. In addition, we flip images vertically with a 0.5 probability and horizontally with a 0.2 probability.

In order to compute network loss, we employ the weighted CE loss function. We chose this loss function because in Child Robot Interactions data are not at all equally distributed among the different engagement levels that represent our training classes. Therefore, it is important that our network pays more attention to training examples of classes that are less common among the data.

Finally, we use three metrics to evaluate our models. These metrics are: standard accuracy, weighted *F-score* and *weighted precision* (w.precision). Generally, *F-score* improves when both precision and recall improve and not when one of them improves at the expense of the other, while weighted *F-score* is the weighted average of the *F-score* of different classes. Respectively, *weighted precision* is the weighted average of precision of different classes.

Network	Accuracy	F-score	W. Precision
majority class	38.03	20.95	15.12
pose CNN [17]	39.34	25.95	25.95
ResNet-50	38.45	25.54	26.30
R(2+1)D RGB	62.17	62.98	60.88
R(2+1)D RGB + Depth	62.52	63.32	63.18
R(2+1)D RGB + Flow (Late)	63.07	63.88	63.57
R(2+1)D RGB + Flow (Early)	**65.78**	**65.38**	**65.95**

TABLE II: Engagement estimation results for the BabyRobot-School Games dataset.

Network	Accuracy	F-score	W. Precision
majority class	62.28	47.81	38.80
pose CNN [17]	68.34	67.57	65.07
R(2+1)D RGB + Depth	68.22	67.47	67.32
R(2+1)D RGB + Flow (Late)	69.85	68.92	68.04
R(2+1)D RGB + Flow (Early)	**71.09**	**70.96**	**70.36**

TABLE III: Engagement estimation results for the ASD-Games dataset[17].

V. Results & Discussion

In this section, we present the results of our experiments on the various data. In Table II, we present estimation results for the BabyRobot-School Games data. We use the tag "majority class" to refer to a network that would always estimate the most common engagement level. Besides the R(2+1)D networks we also include the CNN network that uses children pose as input that we proposed in [17] for child robot interactions during which children where moving around the room as well as a network based on ResNet-50 as in [6].

During the BabyRobot-School Games interactions children's pose are quite limited and this is reflected on the efficiency of the CNN network that takes children's pose as input. Although, this network could successfully learn to estimate children engagement in interactions during which children were playing around the room it can not generalize on limited different poses. Moreover, the ResNet-50 based network, which does not take into account the sequence of frames - the process of the ongoing interaction - is not able to learn to accurately estimate engagement.

On the contrary, we see that the ResNet (2+1)D achieves *accuracy* higher than 60%, with correspondingly high values of *F-score* and *w. precision* even when training on the RGB data alone. Adding depth training data does not cause a significant estimation improvement. This is not unexpected at all, as during these interactions children are seated in front of a desk and the depth data of the videos does not display any significant divergence associated with children's action or engagement state. Estimation efficiency is growing when optical flow data are exploited along with raw RGB data, while the best estimation results are observed when RGB and flow channel are fused at an early stage of the model. These observations are in accordance with respective research conclusions on training CNN neural networks [42], [18]. Evaluation results on the other datasets lead to the same conclusions, as far as depth's, optical flow's, early and late fusion's contributions are concerned. Therefore, we settle on the R(2+1)D RGB + Optical Flow (Early Fusion) network, which achieves *accuracy* and *Fscore* around 65% both on the Sums and on the Emotions games.

Network	Accuracy	F-score	W. Precision
majority class	36.67	19.68	13.45
ResNet-50	40.52	34.32	35.05
R(2+1)D RGB + Depth	62.49	59.87	59.36
R(2+1)D RGB + Flow (Late)	66.78	65.30	65.32
R(2+1)D RGB + Flow (Early)	**68.40**	**67.11**	**68.50**

TABLE IV: Engagement estimation results for the PInSoRo dataset.

Ground Truth	Predictions		
Engagement Level	No-play	Aimless	Goal-oriented
No-play	**53.83%**	5.74%	40.43%
Aimless	9.02%	**48.62%**	42.36%
Goal-oriented	4.27%	9.63%	**86.10%**

TABLE V: Confusion matrix for engagement estimation results for the PInSoRo dataset.

In Table III, we present estimation results for the ASD-Games data from our previous work on the subject [17]. Here we can see that the proposed R(2+1)D based method outperforms the pose CNN method of [17] in all metrics, showing high estimation rates in these kind of interactions as well. We remind that these data contain a variety of different interactions during which children are asked to talk, gesture, move around the room or play before a screen.

In Table IV, we present estimation results for the PInSoRo dataset. Our model achieves *accuracy* and *weighted precision* higher than 68% as well as *F-score* around 67%. The other methods that we tested on these data failed to generalize and learn to estimate the ground truth, having low accuracy. ResNet-50 achieved accuracy 40.52% and pose CNN 57.07%. As far as we know, these are the first extensive results on the child-robot interactions of the PInSoRo dataset.

Finally, for comparison reasons with [30] in Table V, we present the confusion matrix for the PInSoRo dataset. We can observe that results are similar although we cannot have a direct comparison. Firstly, we refer to the child-robot interactions, whereas [30] refers to the child-child interactions. At the same time, training and testing clips are created in a different manner as we use specific parts of the interactions (5th-15th minutes), while [30] collects pieces with specific engagement level to form experimenting data.

VI. Conclusion

In this paper, our goal is to develop a method that successfully estimates children's engagement level during various and challenging child robot interactions regardless of the participating children's development state, their kinetic behavior or even the goal of the interaction. We propose a method that based on a spatiotemporal residual network estimates engagement on six seconds clips. The experiments with different data show that our method improves engagement estimation compared to previously proposed methods. Future work should explore the possibility of improving estimation leveraging state of the art action recognition architectures and look into the possible contribution of multimodal characteristics - e.g. audio, text or multimedia information of the interaction - for further improvement of the estimation. The final goal should be to incorporate the engagement estimation model to an educational robot so that it can be tested in real conditions during child-robot interactions.

References

[1] K. D. Bartl-Pokorny, M. Pykała, P. Uluer, D. E. Barkana, A. Baird, H. Kose, T. Zorcec, B. Robins, B. W. Schuller, and A. Landowska, "Robot-based intervention for children with autism spectrum disorder: A systematic literature review," *IEEE Access*, vol. 9, pp. 165433–165450, 2021.

[2] N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos, and P. Maragos, "Childbot: Multi-robot perception and interaction with children," *Robotics and Autonomous Systems*, 2021.

[3] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, 2018.

[4] J. Wainer, B. Robins, F. Amirabdollahian, and K. Dautenhahn, "Using the humanoid robot kaspar to autonomously play triadic games and facilitate collaborative play among children with autism," *IEEE Transactions on Autonomous Mental Development*, vol. 6, pp. 183–199, 2014.

[5] E. S. Kim, L. D. Berkovits, E. P. Bernier, D. Leyzberg, F. Shic, R. Paul, and B. Scassellati, "Social robots as embedded reinforcers of social behavior in children with autism," *Journal of autism and developmental disorders*, vol. 43, pp. 1038–1049, 2013.

[6] O. Rudovic, H. W. Park, J. Busche, B. Schuller, C. Breazeal, and R. W. Picard, "Personalized estimation of engagement from videos using active learning with deep reinforcement learning," in *Proc. CVPR Workshop*, 2019.

[7] S. Tariq, S. Baber, A. Ashfaq, Y. Ayaz, M. Naveed, and S. Mohsin, "Interactive therapy approach through collaborative physical play between a socially assistive humanoid robot and children with autism spectrum disorder," in *Proc. ICSR*. Springer, 2016, pp. 561–570.

[8] A.R. Taheri, M. Alemi, A. Meghdari, H.R. Pouretemad, and S.L. Holderread, "Clinical application of humanoid robots in playing imitation games for autistic children in iran," *Procedia-Social and Behavioral Sciences*, vol. 176, pp. 898–906, 2015.

[9] S. Ali, F. Mehmood, D. Dancey, Y. Ayaz, M. J. Khan, N. Naseer, R. D. C. Amadeu, H. Sadia, and R. Nawaz, "An adaptive multi-robot therapy for improving joint attention and imitation of asd children," *IEEE Access*, vol. 7, pp. 81808–81825, 2019.

[10] Y. Zhang, W. Song, Z. Tan, H. Zhu, Y. Wang, C. Man Lam, Y. Weng, et al., "Could social robots facilitate children with autism spectrum disorders in learning distrust and deception?," *Computers in Human Behavior*, vol. 98, pp. 140–149, 2019.

[11] I. Giannopulu, K. Terada, and T. Watanabe, "Communication using robots: a perception-action scenario in moderate asd," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, pp. 603–613, 2018.

[12] M. Otterdijk, M. de Korte, I. Smeekens, J. Hendrix, M. Dongen-Boomsma, J. Buitelaar, T. Lourens, J. Glennon, W. Staal, and E. Barakova, "The effects of long-term child–robot interaction on the attention and the engagement of children with autism," *Robotics*, vol. 9, 09 2020.

[13] Y. Feng, Q. Jia, M. Chu, and W. Wei, "Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation," *IEEE Access*, vol. 5, pp. 19494–19504, 2017.

[14] Roger Bakeman and Lauren B. Adamson, "Coordinating attention to people and objects in mother-infant and peer-infant interaction," *Child Development*, vol. 55, no. 4, pp. 1278–1289, 1984.

[15] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human-robot interaction," in *Proc. HRI*, 2016.

[16] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *Front. Robot. AI*, 2020.

[17] D. Anagnostopoulou, N. Efthymiou, C. Papailiou, and P. Maragos, "Engagement estimation during child robot interaction using deep convolutional networks focusing on asd children," in *Proc. ICRA*, 2021.

[18] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. CVPR*, 2018.

[19] "Zeno," https://www.hansonrobotics.com/zeno/.

[20] S. Lemaignan, C. E. R. Edmunds, E. Senft, and T. Belpaeme, "The pinsoro dataset: Supporting the data-driven study of child-child and child-robot social dynamics," *PLOS ONE*, vol. 13, 2018.

[21] M. Khamassi, G. Chalvatzaki, T. Tsitsimis, G. Velentzas, and C. Tzafestas, "A framework for robot learning during child-robot interaction with human engagement as reward signal," in *3rd Workshop BAILAR, Proc. ROMAN*, 08 2018, pp. 461–464.

[22] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *Proc. HRI*, 2010.

[23] G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. Mcowan, "Detecting user engagement with a robot companion using task and social interaction-based features," in *Proc. ICMI*, 2009.

[24] Y. Feng, Q. Jia, M. Chu, and W. Wei, "Engagement evaluation for autism intervention by robots based on dynamic bayesian network and expert elicitation," *IEEE Access*, vol. 5, pp. 19494–19504, 2017.

[25] J. Hadfield, G. Chalvatzaki, P. Koutras, M. Khamassi, C. S. Tzafestas, and P. Maragos, "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task," in *Proc. IROS*, 2019.

[26] P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos, "Fusing body posture with facial expressions for joint recognition of affect in child–robot interaction," *IEEE Robotics and Automation Letters*, vol. 4, 2019.

[27] C. Filippini, E. Spadolini, D. Cardone, D. Bianchi, M. Preziuso, C. Sciarretta, V. Cimmuto, D. Lisciani, and A. Merla, "Facilitating the child–robot interaction by endowing the robot with the capability of understanding the child engagement: The case of mio amico robot," *Social Robotics*, vol. 13, 2021.

[28] O. Rudovic, Y. Utsumi, J. Lee, J. Hernandez, E. C. Ferrer, B. Schuller, and R. W. Picard, "Culturenet: A deep learning approach for engagement intensity estimation from face images of children with autism," in *Proc. IROS*, 2018.

[29] G. Papakostas, G. Sidiropoulos, C. Lytridis, C. Bazinas, V. Kaburlasos, E. Kourampa, E. Karageorgiou, P. Kechayas, and M. Papadopoulou, "Estimating children engagement interacting with robots in special education using machine learning," *Mathematical Problems in Engineering*, 2021.

[30] M. E. Bartlett, T. C. Stewart, and S. Thill, "Estimating levels of engagement for social human-robot interaction using legendre memory units," in *Proc. HRI*, 2021.

[31] Y. Zhang, Y. Tian, P. Wu, and D. Chen, "Application of skeleton data and long short-term memory in action recognition of children with autism spectrum disorder," *Sensors*, vol. 21, 2021.

[32] J. Li, A. Bhat, and R. Barmaki, "A two-stage multi-modal affect analysis framework for children with autism spectrum disorder," 2021.

[33] Ji Lin, Chuang Gan, and Song Han, "Tsm: Temporal shift module for efficient video understanding," in *Proc. ICCV*, 2019.

[34] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann, "Video transformer network," *Proc. ICCVW*, 2021.

[35] D. Gouaillier, V. Hugel, P. Blazevic, C. Kilner, J. Monceaux, P. Lafourcade, B. Marnier, J. Serre, and B. Maisonnier, "Mechatronic design of nao humanoid," in *Proc. ICRA*, 2009.

[36] "Furhat Robotics," http://www.furhatrobotics.com/.

[37] Koichi Negayama and Colwyn Trevarthen, "A comparative study of mother-infant co-regulation of distance at home in japan and scotland," *Infant Behavior and Development*, vol. 68, pp. 101741, 2022.

[38] C. Trevarthen, "From the intrinsic motive pulse of infant actions to the lifetime of cultural meanings," in *In Philosophy and psychology of time*, Peter Øhrstrøm Bruno Mölder, Valtteri Arstila, Ed., pp. 225–265. Springer, Cham, 2016.

[39] Maya Gratier and Colwyn Trevarthen, "Musical narrative and motives for culture in mother-infant vocal interaction," *Journal of Consciousness Studies*, vol. 15, pp. 46–79, 10 2008.

[40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*. 2019.

[41] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. ICLR*, 2014.

[42] K. Gadzicki, R. Khamsehashari, and C. Zetzsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. FUSION*, 2020.