Engagement Estimation During Child Robot Interaction Using Deep Convolutional Networks Focusing on ASD Children

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

***Abstract*— Estimating the engagement of children is an essential prerequisite for constructing natural Child-Robot Interaction. Especially in the case of children with Autism Spectrum Disorder, monitoring the engagement of the other party allows robots to adjust their actions according to the educational and therapeutic goals in hand. In this work we delve into engagement estimation with a focus on children with autism spectrum disorder. We propose deep convolutional architectures for engagement estimation that outperform previous methods, and explore their performance under variable conditions, in four databases depicting ASD and TD children interacting with robots or humans.**

I. Introduction

Social robots have been increasingly involved in our daily lives, while they have also been introduced in the educational process of children [1], [2]. Besides typically developing (TD) children, social robots have been employed to help children with special needs acquire knowledge and skills, especially children with Autism Spectrum Disorder (ASD) [3], [4]. Social robots have shown many advantages in educational and therapeutic purposes for children with ASD [5], [6], [7]. When interacting with robots, children with ASD have shown more interest and experienced more elevated attention. Also they were more likely to maintain a calm and active mood, showed themselves to be more comfortable with emotional response modification and were less likely to display repetitive behaviors compared to interacting with people. These findings indicate that children with ASD might profit significantly from their interaction with robots [8], [9].

In order to achieve qualitative interaction between children and social robots it is highly important that robots can adapt their behaviour to the children cognitive state [10]. Engagement is a key characteristic of human response to an interaction, and various definitions of it can be found in bibliography. In [11] engagement is defined as the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake. Poggi in [12] added that engagement is the level at which a participant attributes to the goal of being together with other participants within a social interaction and how much they continue this interaction. Lemaignan et al. [13] described engagement as the measure of "with-me-ness",meaning the extent to which the human is "with" the robot over the course of an interactive task.

Additionally, there are numerous challenges in estimating children engagement, as it is a complex, multi-faceted cognitive mechanism that is only indirectly observable [13]. With engagement being an inherently internal mental state of the human interacting with the robot, observers (human or robot) have to resort to the analysis of external cues (vision, speech, audio) to estimate its level [14]. Furthermore, search results show that engagement is easier to predict for TD children than for ASD children [15]. Cues like eye-gaze, blinking and head-pose, which are shown to be indicative of the engagement level in TD children, do not appear so directly connected with it in ASD children[16].

Our main goal is estimating the engagement level for ASD children interacting with social robots. We propose a method that estimates at a high level children engagement so that the robot adapts its actions to establish common ground and shared goals with the child. We have been working with several different deep neural network architectures and have concluded to the ones that outperform previous methods when they use either 3D or 2D coordinates of pose keypoints. Moreover, in order to evaluate the generalization of our models, we test them on four different datasets with different situations and participants during the interaction. We have tested our method on interaction of ASD children playing at home with their mothers, a situation that differs significantly than children interacting with robots in laboratory conditions. In addition, we have also tested our method on data of TD children, in order to check its efficacy and compare it with previous published results [17]. Our resulting models achieve important success in engagement estimation for children facing autism spectrum disorders in a variety of conditions, situations and interactions.

II. Related Work

Recently, several studies proposed methods for estimation of child engagement during Child-Robot Interaction (CRI), with each method varying as far as used data, engagement classes, features and machine learning algorithms are concerned. Earlier studies focused mainly on features such as head [18], gaze direction[19], face expression [20], and distance from partner [21]. Hadfield et al. [17], in a previous work of our laboratory, designed an LSTM based network to estimate level of engagement of TD children participating in a joint attention task with a robot. They used children poses as well as other features like head and body direction. Rudovic et al. [22] employ a ResNet which

[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

Proc. of IEEE International Conference on Robotics & Automation (ICRA_2021), Xi'an, China, June 2021

receives directly the RGB data, followed by a deep neural network that implements reinforcement learning: the model decides whether to estimate the engagement level or seek for human help. Moreover, Baxter et al. [14] in order to estimate engagement designed a network with two modules: the convolutional module extract features from the RGB frames, while the recurrent module uses these features to extract a temporal feature vector, from which a fully connected layer estimates the engagement level. Filntisis et al. [23] showed that by using pose keypoints we can get a more accurate estimation of children's emotional state than by using only the rgb information of children faces. Thus, they proposed a model that fused two estimation scores to produce the final estimation: a score given by a convolutional ResNet 50 fed with the children face cropped images and a score given by a dense neural network fed with pose information.

A handful of studies have proposed models for engagement estimation of children with autism spectrum disorders. Recently, Javed et al. [24] proposed a method to estimate the engagement level of ASD children. They used pose keypoints that correspond to eyes and lips, as well as three extra features designated by Laban Movement Analysis [25]. They trained a one-dimensional convolutional deep neural network with the feature vectors. In CultureNet [16], Rudovic et al. propose a method for ASD children engagement estimation personalized in different culture backgrounds. Their method uses directly the RGB data, which it feeds in a ResNet-50 network.

III. Data and Data Processing

A. Data Description

We have been working with four different groups of data. Our main set of data consists of seven sessions in which participate seven children, two girls and five boys, facing autism spectrum disorder (mean age 10.6 years old). The kids were recruited from the Special School for Children with Autism in Piraeus, Greece, and were randomly assigned in two levels of severity, mild and moderate, which were assessed by two experienced clinical child psychologists.

During each session, which lasted approximately 20 minutes, one child participated and played five different games with two robots, NAO [26] and Furhat [27]. Each child entered in an especially adapted laboratory accompanied by the researcher and he/she was introduced to the robots and vice versa. The purpose of this introduction is to help the children to assume that the robots can be regarded as a playmate. After the introduction, while the interaction were proceeding, the researcher didn't interfere at all.

The games that children played were: Show me the Gesture, Express the Feeling, Pantomime, Guess the Object and Joint Attention. During the *Express the Feeling* Furhat asks the child to imitate using their faces emotions appearing on the touch screen. In *Show me the Gesture*, Furhat prompts the child to make specific gestures (e.g. wave). During the *Pantomime* game, the Nao robot asks the child to imitate some actions appearing on the touch screen. In the *Guess the Object*, Furhat describes objects that lie in the room and asks the child to find them and place them in a specific box. Finally, during the *Joint Attention* game the child interacts with Nao and is free to move around the room as they please. With a series of motions the robot attempts to capture the child's attention and prompts the child to hand over a brick, that lies on the floor. For comparison reasons, we refer to the first four games as ASD-Games data (Fig. 1b), and the last game as ASD-Joint Attention (Fig. 1a).

Another dataset that is used for our experiments consists of 25 sessions in which participate equal number of typically developing children (mean age 8.6 years old) participating in *Joint Attention* task. Once again, in each session, the child engages with Nao robot in the joint attention activity. We refer to these data as the TD-Joint Attention data set as child and robot interact in a simple activity that acquires their mutual attention.

Finally, the fourth group of data consists of three sessions in which participate three younger children (mean age 5 years old) with their mothers, facing autism spectrum disorder. The interactions take place in each child's home. Each child participates in two sessions. In the first one child and mother play the *Explorer* game, in which they explore together a box with many different toys. In the second one the mother pretends that she has been hurt and she cries so that the child's response to a situation like this can be recorded. We refer to these data as the BabyAffect data set (Fig. 1c)

We use these four different data groups in order to ensure that our models are dealing with data of significant variety. There are TD and ASD children taking part in the same interaction under the same conditions, ASD children taking part in more different from each other interactions and finally ASD children taking part in activities along with their mothers in their home environment.

B. Data Annotation and Analysis

We regard the level of engagement as the level at which the child is both attentive and cooperative with the robot towards their common goal. We designate three distinct levels of engagement: the first (class 1) signifies that the child is disengaged, meaning that they are paying limited or no attention to the robot and that they do not act towards their common goal in any way; the second (class 2) refers to a partial degree of engagement, where the child either acts relatively to the common goal or pays attention to the robot but not both simultaneously; the final level (class 3) means that the child is actively cooperating with the robot to complete their common goal. The data were annotated by laboratory members according to a set of instructions containing groups of visual and acoustic cues that correspond to each engagement level provided by an expert psychologist. For example, if the child looks at a relevant object, says something relevant to the common goal and has a neutral facial expression or if the child looks at the partner, manipulates an irrelevant object and does not speak, they both have medium engagement level.

There is significant imbalance between the different engagement levels in all four data groups, that we have to take

(a) JOINT-ATTENTION DATASET

(b) ASD-GAMES DATASET

(c) BABYAFFECT DATASET

Fig. 1: Instances from the different data sets environments. Children interact with social robots in laboratory. In front of them is placed a Furhat robot, Anna. On the right stands on the floor a Nao robot, Paris. Children are free to move in the room as they please. TD Joint Attention, ASD Joint Attention and ASD Games experiments take place in this environment. (a) A ASD-JOINT ATTENTION instance, child hands brick to Nao robot. (b)An ASD-GAMES instance, child plays 'Show me the Gesture' game (c) A BABYAFFECT instance, children play with their mother at home.

into consideration when designing our estimation method. Table I shows the distribution into classes for the different data groups. We have also included ASD-Joint Attention-Human and ASD-Games-Human, which consist of the same interactions (same children and games/goals) as the ASD-Joint Attention and ASD-Games, but conducted with human partner instead of robots and TD-BabyAffect in which participate typically developing children in BabyAffect context.

Class discrimination in Table 1 shows that during interaction with their mother in a semi-naturalistic setting, children with ASD are approximately half time disengaged compared to TD children (14.72% vs. 31%). This finding seems to contradict with one of the core features of ASD, i.e., withdrawn. However, microanalysis of the interactions shows that compared to the TD group mothers in the ASD group make more intense effort to initiate interaction and impose responses to their child more often, not allowing her to be disengaged [28], [29]. On the contrary, in the experimental joint attention situation with a humanoid robot, children with ASD spent twice as long time disengaged compared to the TD children (7.80% vs. 17.68%). This time was doubled in the human condition (34.62%). Similarly, children with ASD spent approximately 1.5 more time disengaged in the Game – Human condition compared to the Game – Robot condition (31.98% vs. 19.35%). On the other hand, children with ASD spent approximately 1.5 less time cooperating with their mother in a home setting, than TD children (20.22% vs. 13.8%). Moreover, it should be noted children with ASD spent approximately the same time cooperating with a robot almost in all structured conditions in the laboratory. These findings accord with studies demonstrating that in structured situations children with ASD may devote even as much time as TD children in certain forms of collaboration, although qualitative differences are still observable [30].

C. Pose Estimation

As mentioned before the child's pose can be very informative for recognizing the child's engagement level. Pose contains concentrated information that exists on an image of people interacting. The problem of detecting human pose keypoints in images is a challenging one, due to occlusions and widely varying articulations and background conditions.

Data	Distribution (%)			Total # Frames
	Class1	Class2	Class3	
TD-Joint Attention-Robot	7.80	83.42	8.78	108,408
ASD-Joint Attention-Robot	17.68	66.07	16.22	50,869
ASD-Games-Robot	19.35	70.09	10.56	109,381
ASD-Joint Attention-Human	34.62	50.00	15.38	4,680
ASD-Games-Human	31.98	52.69	15.33	75,150
ASD-BabyAffect	14.72	71.48	13.80	27,207
TD-BabyAffect	31.00	48.78	20.22	26,830

TABLE I: Distribution of engagement levels for the different datasets.

Only recently has the problem been solved to a satisfactory degree, especially with the introduction of the Open Pose library [31]–[32] for 2D keypoint detection.

We use the Open Pose library to extract the 2D pose keypoints. For the Joint Attention data groups we also possess depth information and multi-camera views. Using these, we obtain 3D coordinates following the method proposed in "A deep learning approach for multi-view engagement estimation of children in a child-robot joint attention task" [17]. Afterwords, in every frame we have to decide which of the detected poses corresponds to the child's pose. This need is created because Open Pose detects poses of other people entering and exiting the camera's view. Especially, in the BabyAffect data this is very important because child and mother are really close to each other in almost every frame. To do that, we choose the child's pose on the first frame by comparing torso lengths of the detected poses. In subsequent frames we choose the pose that is closest to the previous child's pose, while on the same time their distance is smaller than a suitable threshold. If there is no such pose we mark the frame as missing. If the previous frame is marked as missing we choose the child's pose comparing torsos length with the child's torso length. Again torso length must lie between two thresholds. We also apply linear interpolation in order to produce missing body parts values.

In BabyAffect data we also produce mother's poses in the same way. In interactions where the partner is the Nao robot (and not Furhat which is stable) we locate Nao's head in every frame. Due to the fact that children in all data sets interact with partners, we are not interested in their independent poses, but in their poses regarding the partner. Therefore, we subtract the coordinates of partner's head from

the child's pose keypoints coordinates.

Afterwards we subtract from each pose's point coordinates the left hip coordinates. We have produced the missing hip values by applying linear interpolation. In this way the vector's values are relative to each other, they reflect the relation between pose parts. Finally, we normalize values in space [0,1].

IV. Method

A. Network Architectures

Engagement significantly depends on temporal information, on the progress of the ongoing interaction. Simultaneously, pose features resemble images and therefore could be successfully processed by convolutional networks. We have to combine the ability of original two dimensional convolutions to extract patterns from images with the need to exploit temporal data information in engagement estimation.

Inspired by '2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning' [33], we rearrange the pose keypoints vectors to structures that resemble images as follows: Vertical axis represents time, horizontal axis represents skeleton parts, while the coordinates of each part represent the images channels. For example, for 3D coordinates of pose keypoints, the x, y, z coordinates correspond to r, g, b channels. Respectively, for 2D coordinates of pose keypoints, the x, y coordinates correspond to two channels. This rearrangement is shown in Fig. 2. In this way, we can exploit the rich information of pose data and the time continuity with convolutions.

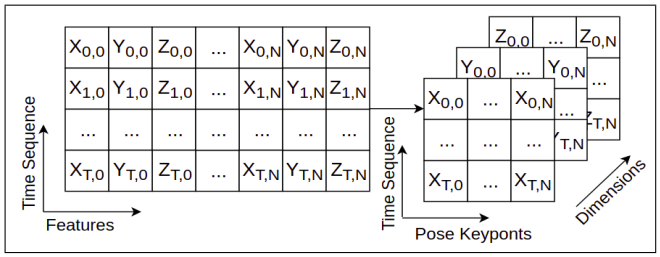

Fig. 2: Rearrange pose keypoints vectors to structures that resemble images.Vertical axis represents time, horizontal axis represents skeleton parts, while the coordinates of each part represent the images channels.

Therefore, we design two convolutional networks that receive data in the form described above. The first one, is a convolutional network that resembles the classic AlexNet architecture [34]. It consists of convolutional layers with suitable characteristics to suit and process our inputs.

In order to achieve greater computational efficiency, both in terms of training time and space required, we also designed a simpler 2D CNN that consists of three convolutional layers (one with kernel size 5 and two with kernel size 3) with ReLU activation and max-pooling, an adaptive average pooling layer, a fully connected layer with dropout and ReLU activation and a final fully connected layer, with a softmax function, applied to the output as before. Number of channels of the first convolutional layer equals the data dimensionality. The network's architecture can be seen in Fig.3.

We compared the results of our network architectures with recurrent neural networks based on LSTM layers that had been used before in the TD-Joint Attention data set[17]. Long Short-Term Memory networks are a special kind of RNN, capable of learning long-term dependencies and are well-suited to classifying based on time series data. The network consists of three fully connected layers with dropout and ReLU activation, one or two Long Short-Term Memory (LSTM) layers and a final fully connected layer, with a softmax function, applied to the output to produce a probability score for each class.

Furthermore, we compared our results with an one-dimensional multi-channel convolutional network inspired by the network in [24] with the addition of pooling layers. Specifically, the network consists of two one-dimensional convolutional layers with ReLU activation, max-pooling and dropout layers, an average pooling layer, two fully connected layers with dropout and ReLU activation, and a final fully connected layer with a softmax function, applied to the output as before.

B. Implementation

We use PyTorch library [35] to implement all networks described in the previous section. We divide the data used to training and validation set leaving in each experiment videos of some children for validation set. Deep neural networks generally require a large amount of data to train successfully and avoid overfitting. Thus, we use two ways of data augmentation. We flip vertically the feature vector with a 0.3 probability and we add a small amount of Gaussian noise to the feature vector with 0.5 probability.

We used a batch size of 128 and a learning rate of $3*10^{-4}$ after we experimented with different values. We also used a sequence length of 200 frames. The sequence length corresponds to the time window that the neural network "sees" every time. With a frame rate of 30fps 200 frames correspond to 6 to 7 seconds.

To update network weights we employ as optimization algorithm the Adam Optimizer [36]. We also choose ReduceLROnPlateau scheduler in order to decrease learning rate when running loss has stopped decreasing by a factor of 2-10 once learning stagnates.

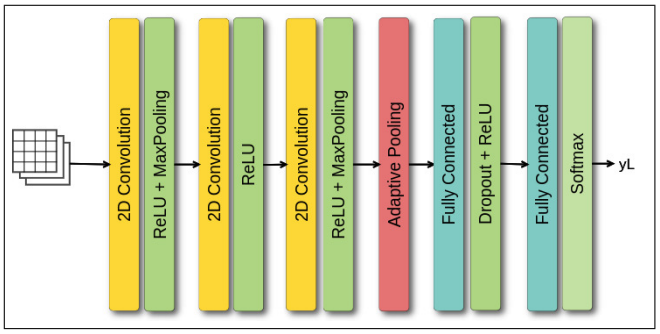

Fig. 3: 2D CNN network for engagement estimation. Network receives as input a sequence of child's poses in form of an image and estimates the child's engagement level.

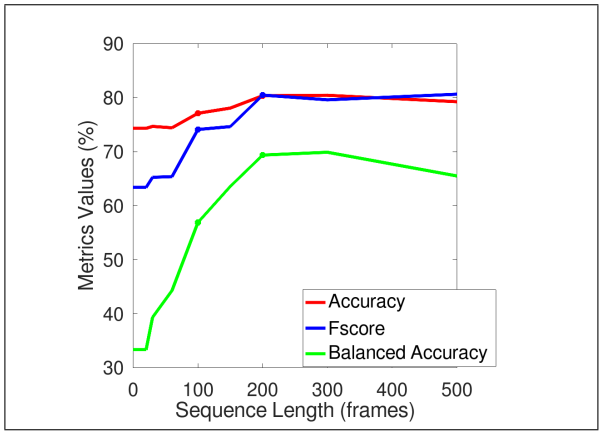

Fig. 4: Change of evaluation metrics for different sequence lengths given to the convolutional AlexNet network to estimate engagement. Sequences larger than 100 frames (approximately 3 seconds) allow the network to train and estimate engagement. Sequences of 200 frames (approximately 6 to 7 seconds) lead to the best results.

As loss function for training our networks we use a weighted Cross Entropy Loss Function, so that our networks take into consideration the significantly unequal distribution of frames into classes in our data sets and attach greater importance to the less common classes. We produce a vector containing the weights for each class, by dividing the number of frames belonging to the majority class with the number of frames belonging to each class. We pass the weight vector as an argument to the Cross Entropy Loss Function. Because Cross Entropy Loss implementation in PyTorch also embodies the softmax function, we remove the final softmax layer from all networks.

To evaluate our models besides standard accuracy, we use two more metrics to account for the great class imbalance. The first one is weighted *F-score*, which take high values only when both precision and recall of each class are high, while the second one is *weighted precision* (w.precision), which is the weighted mean value of all classes precision.

V. Results & Discussion

In Table II we present estimation results for the TD-Joint Attention data set. We use the tag 'majority class' to refer to the a network that would always produce as output the most common class. Besides the four networks described above we also include an LSTM network as in [17], an one dimensional CNN network as in [24] (here as input features we use our features and not the features proposed in [24]) and a network based on ResNet-50 as in [22]. The estimation results, especially these produced by the AlexNet model, are truly satisfying. Our network achieves accuracy, F-score and precision higher than 80% , which are significantly higher than scores of other networks.

In Table III we present estimation results for the ASD-Joint Attention data set. We train for a few epochs the pretrained on TD-Joint Attention data set models using the ASD-Joint Attention data. If we estimate ASD children engagement level directly using the TD- Joint Attention pretrained networks we get relatively poor results with highest accuracy not outreaching 60%. However, using the TD pretrained networks in a fast training with small amount of data we obtain networks that succeed in estimating ASD engagement. This, is important as ASD data are fewer and more difficult to collect. Although the metrics values are not so high as for the TD children, our models manage to estimate ASD children engagement to a large extent. As we mentioned above, ASD children do not express themselves in so direct ways as TD children, making it more challenging to estimate their engagement. Therefore, the fact that the convolutional AlexNet achieves nearly 80% accuracy (78.73%), with relatively high F-score and precision, is truly encouraging. The convolutional AlexNet performs considerably better than all other networks.

Network	Accuracy	F-score	W. Precision
majority class	74.32	63.38	55.25
1D CNN	72.98	60.15	56.32
ResNet50	74.52	64.52	63.285
LSTM (one layer)	76.23	74.15	74.35
1D CNN	77.44	75.15	76.30
2D CNN	78.93	76.46	77.43
LSTM	79.47	76.88	78.04
AlexNet	**80.36**	**80.48**	**80.71**

TABLE II: Engagement estimation results of different network architectures for the TD-Joint Attention data set.

Network	Accuracy	F-score	W. Precision
majority class	57.44	41.15	32.26
1D CNN	67.91	67.98	68.10
2D CNN	70.88	71.32	73.63
LSTM	71.76	71.55	71.49
AlexNet	**78.73**	**73.39**	**82.38**

TABLE III: Engagement estimation results of different network architectures for the ASD-Joint Attention data set.

Network	Accuracy	F-score	W. Precision
majority class	57.44	41.15	32.26
2D CNN (2D coordinates)	**75.82**	**75.39**	**76.96**
2D CNN (3D coordinates)	70.88	71.32	73.63
LSTM (2D coordinates)	61.53	60.48	62.05
LSTM (3D coordinates)	**71.76**	**71.55**	**71.49**
AlexNet (2D coordinates)	67.18	64.26	67.83
AlexNet (3D coordinates)	**78.73**	**73.39**	**82.38**

TABLE IV: Engagement estimation results for ASD-Joint Attention data set using three dimensional (3D) and two dimensional (2D - without depth) keypoint coordinates.

For the TD and ASD-Joint Attention data sets we worked with the 3D coordinates of pose parts. This was possible because all sessions were recorded with three RGB-D cameras, enabling the 3D coordinates extraction. However, not all children engagement observation experiments are intended for computational utilization from the beginning. Therefore, they may be recorded with a simple conventional camera, not allowing depth extraction. ASD-Games and BabyAffect data sets consist of two dimensional data. Therefore, in Table IV we present a comparison between training with 2D and

3D data the ASD-Joint Attention data set. 3D models are provided with more essential information, as the depth coordinate reveal information about the different image layers, the relative positions of people and objects in the image. Thus, as it was expected models learn more accurately when 3D coordinates are available. The 1D CNN model could not be trained at all being provided only with 2D coordinates. On only one occasion (2D CNN) the 2D model achieved better estimation than the 3D one. This may be due to overfitting, but more probably our convolutional neural network (shown in Fig. 3) fitted with the two dimensional input vectors. This network achieves relatively high estimation results for the two remaining data sets which are two dimensional.

In Table V we present estimation results for the ASD-Games data set. We fine-tuned our models and have managed to produce satisfying engagement estimations for these interactions too. Even though, the estimation is less accurate than the estimation on the Joint Attention data sets, the results still are relatively good. Specifically, convolutional AlexNet network achieves accuracy and F-score that approach 70% (68.34% and 67.57% respectively). These results show that our method can successfully be used for ASD children engagement estimation in CRI interactions for a variety of such interactions during which children are asked to talk, gesture, move around the room or play before a screen. The reason for the accuracy difference between ASD Games and ASD Joint Attention data set is that in ASD Games we estimate ASD children engagement level in much more varying interactions than the simple joint attention interaction with the robot asking for a toy.

Finally, in Table VI we present estimation results for the BabyAffect data set. Once again, we achieve satisfying engagement estimation scores, in different and difficult conditions. Here, ASD children play with their mothers in their home environment, meaning that they feel entirely free and at ease. Both convolutional networks achieve high engagement estimation accuracy results. The 2D CNN convolutional network achieves accuracy that once again approaches 80% (77.59%), with over 70% F-score and weighted precision respectively. The results for both convolutional networks are significantly high, given the difficulty of the task's conditions, proving that our method can successfully estimate engagement in a totally different environment and for children-adult interactions too.

It is important to emphasize that all above results have been achieved by training the networks (regardless of network architecture) with sequence length of 6 seconds. This time interval proves to be both necessary and sufficient in order to allow the network to learn to identify a child's level of engagement. We have experimented with a variety of lengths before choosing 6 seconds. In all different kind of networks, we came into the conclusion that the network needs to see an interval of at least three seconds in order to estimate the engagement level successfully. There is an improvement from 3 to 7 seconds, but beyond this sequence length the network does not further improve, while the training becomes much slower. Fig. 4 shows the change of values of the evaluating metrics for different sequence lengths for the convolutional AlexNet network for the ASD-Joint Attention data set. It can be seen that the network is trained and estimates satisfactory the engagement when the sequences are larger than 100 frames, i.e. approximately 3 seconds, while sequences of 200 frames, i.e. approximately 6 to 7seconds, lead to the best results. The above results confirm the effectiveness of the network on estimating the engagement and are also in accordance with corresponding psychologists conclusions. Human beings are endowed with the ability to express feelings and intentions in a shared social space through movements organized in a time frame ranged 0.3 to 7 seconds. Especially the time frame 3 to 6 seconds is considered fundamental to human motoric and perceptual functions, such as early mother – infant interactions, language, and music [37], [38], [39].

Network	Accuracy	F-score	W. Precision
majority class	62.28	47.81	38.80
LSTM	64.77	55.55	62.80
2D CNN	67.28	56.09	51.65
AlexNet	**68.34**	**67.57**	**65.07**

TABLE V: Engagement estimation results of different network architectures for the ASD-Games data set.

Network	Accuracy	F-score	W. Precision
majority class	71.48	59.79	51.10
LSTM	70.55	62.68	56.20
2D CNN	**77.59**	**71.73**	**74.67**
AlexNet	74.24	69.51	68.03

TABLE VI: Engagement estimation results of different network architectures for the BabyAffect data set.

VI. Conclusion

In this paper we focused on engagement estimation for children with autism during interaction with robots. We proposed deep convolutional architectures trained with pose features for the task at hand and our extensive experiments showed the superiority of our method to previous ones. The greater challenge is to create a model that can relatively easily adapt from its training conditions to different ones and continue to successfully estimate ASD children engagement. An important direction for future work will be to test and generalize our method to different kind of interactions, such as interactions during which children are sited and therefore their pose presents much less variety, as well as enrich our approach with more input characteristics, such as facial expressions and speech.

ACKNOWLEDGMENT

The authors wish to thank Dr. Asimenia Papoulidi for conducting and evaluating the majority of the recorded experiments with the children and P.P. Filntisis for his useful remarks on the research. We would like also to thank the personnel of the Special School for Children with Autism in Piraeus, for their collaboration in the experimental procedure and the members of the NTUA IRAL who organized and recorded the experiments.

REFERENCES

[1] N. Efthymiou, P. P. Filntisis, P. Koutras, A. Tsiami, J. Hadfield, G. Potamianos, and P. Maragos, "Childbot: Multi-robot perception and interaction with children," *arXiv preprint arXiv:2008.12818*, 2020.

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

[3] 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.

[4] 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.

[5] 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 *International Conference on Social Robotics*. Springer, 2016, pp. 561–570.

[6] 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.

[7] 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.

[8] Y. Zhang, W. Song, Z. Tan, H. Zhu, Y. Wang, C. Man Lam, Y. Weng, S. P. Hoi, H. Lu, B. S. M. Chan, 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.

[9] 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.

[10] 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.

[11] C. L. Sidner, C. Lee, and N. Lesh, "The role of dialog in human robot interaction," in *International workshop on language understanding and agents for real world interaction*, 2003.

[12] I. Poggi, *Mind, hands, face and body: a goal and belief view of multimodal communication*, Weidler, 2007.

[13] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg, "From real-time attention assessment to "with-me-ness" in human-robot interaction," in *11th ACM International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016.

[14] F. Del Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? continuous engagement assessment from a robot's point of view," *arXiv preprint arXiv:2001.03515*, 2020.

[15] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *Proc. ICASSP*. IEEE, 2017.

[16] 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*. IEEE, 2018.

[17] 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*. IEEE, 2019.

[18] Mehdi Khamassi, Georgia Chalvatzaki, Theodore Tsitsimis, Georgios Velentzas, and Costas Tzafestas, "A framework for robot learning during child-robot interaction with human engagement as reward signal," in *3rd Workshop on Behavior Adaptation, Interaction and Learning for Assistive Robotics (BAILAR), in the 27th International Conference on Robot and Human Interactive Communication (ROMAN)*, 08 2018, pp. 461–464.

[19] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner, "Recognizing engagement in human-robot interaction," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010.

[20] 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 *Proceedings of the 2009 international conference on Multimodal interfaces*, 2009.

[21] 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.

[22] 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*. IEEE, 2019.

[23] 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, pp. 4011–4018, 2019.

[24] H. Javed, W. Lee, and C. H. Park, "Toward an automated measure of social engagement for children with autism spectrum disorder–a personalized computational modeling approach," *Frontiers in Robotics and AI*, vol. 7, pp. 43, 2020.

[25] E. Groff, "Laban movement analysis: Charting the ineffable domain of human movement," *Journal of Physical Education, Recreation & Dance*, vol. 66, pp. 27–30, 1995.

[26] 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*. IEEE, 2009.

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

[28] Martina Franchini, Vickie L. Armstrong, Marie Schaer, and Isabel M. Smith, "Initiation of joint attention and related visual attention processes in infants with autism spectrum disorder: Literature review," *Child Neuropsychology*, vol. 25, no. 3, pp. 287–317, 2019.

[29] Asimenia Papoulidi, Christina F. Papaeliou, and Stavroula Samartzi, "Rhythm in interactions between children with autism spectrum disorder and their mothers," *Timing & Time Perception*, vol. 5, no. 1, pp. 5 – 34, 2017.

[30] Papaeliou CF, Sakellaki K, and Papoulidi A, "The relation between functional play and other forms of cooperation and word learning in asd," *Int Arch Commun Disord*, 2019.

[31] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proc. CVPR*, 2017.

[32] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1812.08008*, 2018.

[33] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *Proc. CVPR*. IEEE, 2018.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, pp. 84–90, 2017.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.

[36] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[37] P. Fraisse, "Rhythm and tempo," *In D. Deutsch (Ed.), The Psychology of Music, New York: Academic Press*, pp. 149 – 180, 1982.

[38] K. Kaye and A. Fogel, "The temporal structure of face-to-face communication between mothers and infants," *Developmental Psychology*, vol. 16, pp. 454 – 464, 1980.

[39] J. Panksepp and C. Trevarthen, "The neuroscience of emotion in music," *In S. Malloch & C. Trevarthen (Eds.), Communicative musicality: Exploring the basis of human companionship*, vol. 16, pp. 105–146, 2009.