HAL
open science

A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task

Jack Hadfield, Georgia Chalvatzaki, Petros Koutras, Mehdi Khamassi, Costas S Tzafestas, Petros Maragos

To cite this version:

Jack Hadfield, Georgia Chalvatzaki, Petros Koutras, Mehdi Khamassi, Costas S Tzafestas, et al.. A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2019), Nov 2019, Macau, China. hal-02324118

HAL Id: hal-02324118

https://hal.science/hal-02324118v1

Submitted on 21 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Deep Learning Approach for Multi-View Engagement Estimation of Children in a Child-Robot Joint Attention Task

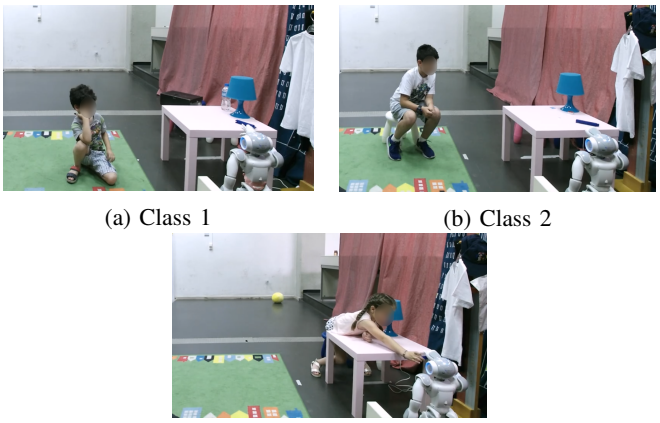
Jack Hadfield[1,2], Georgia Chalvatzaki[2], Petros Koutras[1,2],
Mehdi Khamassi[2,3], Costas S. Tzafestas[2] and Petros Maragos[1,2]

***Abstract*— In this work we tackle the problem of child engagement estimation while children freely interact with a robot in a friendly, room-like environment. We propose a deep-based multi-view solution that takes advantage of recent developments in human pose detection. We extract the child's pose from different RGB-D cameras placed regularly in the room, fuse the results and feed them to a deep neural network trained for classifying engagement levels. The deep network contains a recurrent layer, in order to exploit the rich temporal information contained in the pose data. The resulting method outperforms a number of baseline classifiers, and provides a promising tool for better automatic understanding of a child's attitude, interest and attention while cooperating with a robot. The goal is to integrate this model in next generation social robots as an attention monitoring tool during various Child Robot Interaction (CRI) tasks both for Typically Developed (TD) children and children affected by autism (ASD).**

I. Introduction

As robots become more integrated in modern societies, the cases of interacting with humans during daily life activities and tasks are increasing. Human-Robot Interaction (HRI) refers to the communication between robots and humans. This communication can be verbal or non-verbal, remote or proximal. A special case of HRI is Child-Robot Interaction (CRI) [1]. Robots enter children's lives as companions, entertainers or even educators [2]. Motivated by this need, in the context of BabyRobot EU project [3], we have developed and evaluated systems which employ multiple sensors and robots, for childrens' speech, gesture and action recognition during CRI scenarios [4]–[6], in a specially designed area, the "BabyRobot room". Children are very adaptive, quick learners with unique communication skills, able to easily convey or share complex information with little spoken language. A major challenge in the field of CRI, however, entails equipping robots with the ability to pick up on such information and adapt their behavior accordingly, in order to achieve a more fruitful interaction.

Robots assisting children is of particular importance in modern research, especially for mediating ASD therapy towards the development of their social skills, [7]. A review on social robots for education can be found in [8]. Children affected by ASD can benefit from interacting with robots, since such a CRI may help them overcome the impediments posed by face-to-face interaction with humans. Moreover, it is important that the robot's behavior can adapt to the special needs of each specific child and maintain an identical behavior for as long as needed in the intervention process [9].



(a) Class 1 (b) Class 2

(c) Class 3

Fig. 1: Examples of three levels of engagement: (a) Limited attention (class 1), (b) Attention but no cooperation (class 2), (c) Active cooperation (class 3).

One key issue for social robots is the development of their ability to evaluate several aspects of interaction, such as user experience, feelings, perceptions and satisfactions [10]. Human engagement in HRI according to [11] "is a category of user experience characterized by attributes of challenge, positive affect, endurability, aesthetic and sensory appeal, attention, feedback, variety/novelty, interactivity, and perceived user control". Poggi in [12] adds 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. Given this rich notion of engagement, many studies have explored human-robot engagement [10]. Lemaignan et. al [13] explored the level of "with-me-ness", by measuring to what extent the human is with the robot during an interactive task, for assessing the engagement level. Human engagement has been modeled in a number of works, using solely gaze [14], speech and gaze [15], and human pose with respect to the robot from static positions [10], [16].

Engaging children in CRI tasks is of great importance. The social characteristics that robots should have when performing as tutors were examined in [17], [18]. Specific

This research work has been supported the EU-funded Project BabyRobot (H2020, grant agreement no. 687831).

[1]Athena Research and Innovation Center, Maroussi 15125, Greece
[2]School of ECE, National Technical Univ. of Athens, 15773 Athens, Greece jack.hadf@gmail.com, {pkoutras,ktzaf,maragos}@cs.ntua.gr, gchal@mail.ntua.gr
[3]Sorbonne Université, CNRS, Institute of Intelligent Systems and Robotics, Paris, France. mehdi.khamassi@upmc.fr

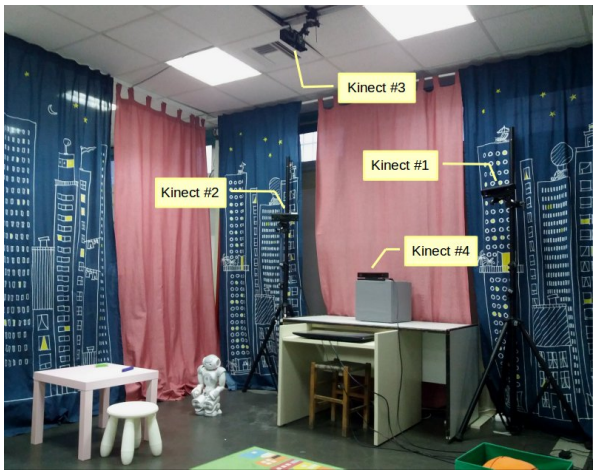


Fig. 2: Setup for recording joint attention experiments.

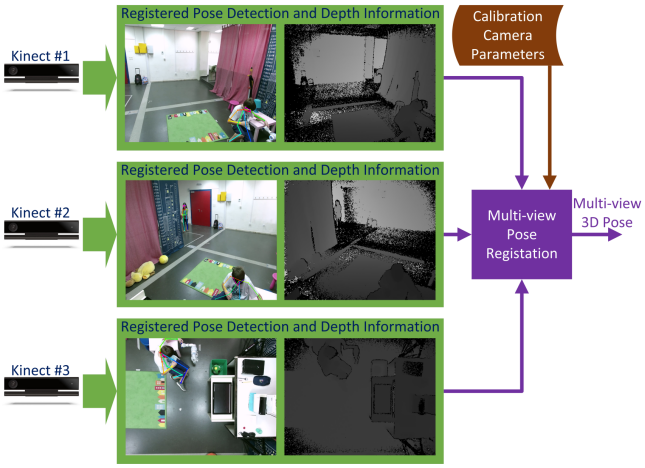


Fig. 3: Multi-view pose estimation overview.

focus is given in estimating the engagement of children with ASD interacting with adults [19] or robots [20]. A study analyzing the engagement of children participating in a robot-assisted therapy can be found in [21]. A method for the automatic classification of engagement with a dynamic Bayesian network using visual and acoustic cues and support vector machine classifiers is described in [22]. Another approach considers the facial expressions of children with ASD to evaluate their engagement [23]. A robot-mediated joint attention intervention system using as input the child's gaze is presented in [24]. A deep learning framework for estimating the child's affective states and engagement is presented in [25] in a static setting in front of the robot. Previous work from our research team focused on using engagement estimates in a reinforcement learning setting to adapt robot's motion and expressivity [26], [27]. Hence, estimating children engagement level is important for adapting the behavior of a robot companion.

In this paper we present an deep learning approach for estimating the engagement of children in a CRI collaborative task aiming to establish joint attention between a child and a robot. In the specific scenario examined, the robot's aim is to elicit child's attention through an experiment consisting of a handover task, that tests the child's attentiveness and social coordination skills, with the hope that the child will correctly understand the robot-agent's intentions and ultimately take part in a successful collaboration. In this setting, children engagement refers to rich information about their movement in the room w.r.t. the robot agent. Our method employs the children full body pose estimates along with high-level computed features, such as gaze and arm extension, to capture a wide range of movements and actions relevant to the child's engagement.

Our method incorporates a multi-view estimation of the child's pose, where the child is inside a specially arranged room with a network of cameras (Fig. 1). An LSTM-based algorithm classifies the engagement of the child to the task using the child's pose as input and observations by experts as control targets. We experimentally validate our algorithms exploiting the RGB-D data from recordings of children who participated in the experimental procedure.

A major contribution of this work is the proposal and validation of an engagement estimation method generalizable to a multitude of CRI tasks. The setup allows the child to freely move and interact with the robot, without being limited to a stationary position in front of a robot, sensor or adult. This is in contrast to prior works which are restricted in such a sense. The data-driven nature of our approach also contributes to its generalizability, as no prior hypotheses are made regarding the actions most relevant to the task. Instead, the algorithm learns to distinguish the movements and stances most informative of the child's engagement. Finally, the multi-view fusion helps achieve a more robust and accurate classifier by confronting cases of body part occlusions and providing better pose estimations than if a single camera was used. We believe that such a pose-related framework can also apply to ASD children engagement estimation who present a variability of behaviors and motions in order to adapt the behavior of robot mediation accordingly.

II. Method

The main problem addressed in this work is to estimate the engagement level of children from visual cues. The problem is cast as one of multi-class classification, where each class corresponds to a different level of engagement. Specifically, we designate three distinct levels of engagement: the first (*class 1*) signifies that the child is disengaged, i.e. paying limited or no attention to the robot; the second (*class 2*) refers to a partial degree of engagement, where the child is attentive but not cooperative; the final level (*class 3*) means that the child is actively cooperating with the robot to complete the handover task. The task details are described in Sec. III-A. During the course of an interactive session, the engagement level varies. The goal is therefore to perform this classification across a number of fixed-length time segments during the session, rather than producing a single estimate for the entire interaction. In the remainder of this section we describe the proposed method to perform this classification.

Fig. 4: The detected pose is shown along with the detected bounding box surrounding the robot's head.

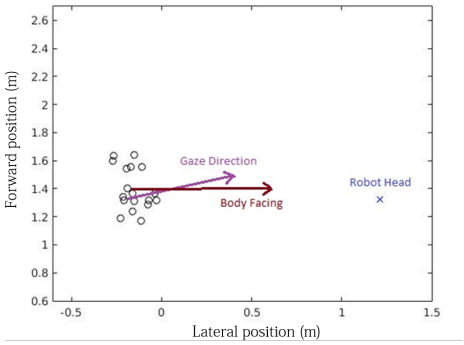


Fig. 5: Extracted features shown from an overhead view of the room. The detected keypoints are shown as black circles.

A. Multi-View child pose estimation

Perhaps the most informative data for recognizing the engagement level in joint attention tasks is that of the child's pose. The problem of detecting human pose keypoints in images is a challenging one, due to occlusions and widely varying articulations and background conditions. Only recently has the problem been solved to a satisfactory degree, especially with the introduction of the Open Pose library [28]–[30] for 2D keypoint detection.

In [31], an end-to-end 3D pose estimator from RGB-D data is proposed. The Open Pose library is used to detect 2D keypoint score maps from the color image. These maps are then fed to a deep neural network, along with a voxel occupancy grid derived from the depth image. The network is trained to produce 18 3D keypoint estimates: two for each wrist, elbow, shoulder, hip, knee and ankle, one for the neck and five facial points, consisting of the ears, the eyes and the nose. We employ this system in our work, to estimate the child poses during their interaction with the robot.

When using multiple cameras, the keypoints can be extracted for each view and fused to produce the final estimates (Fig. 3). The first step to achieve this is to register the points of each camera reference frame to a single common frame. The registration parameters were found using the ICP algorithm [32], which provides the transformation that best fits the point cloud of one camera to that of another, given an initial transformation that we set manually.

After transforming the keypoint coordinates of all cameras to a common reference frame, the next step is to determine which keypoints are valid from each view. The pose estimation algorithm occasionally fails, either when some of the child's joints are hidden, or when the pose differs substantially from those used to train the algorithm. In such cases, the system produces noisy estimates or no detections at all for certain keypoints. Another problem is that the algorithm sometimes outputs multiple poses, when another person is in view or occasionally when the network is confused by some background artifact. To tackle such problems, we only average the points that are sufficiently close to those detected in the previous frame, i.e. at a maximum distance of 0.5m. If no such points exist for a certain joint, we mark the joint as missing in the current frame.

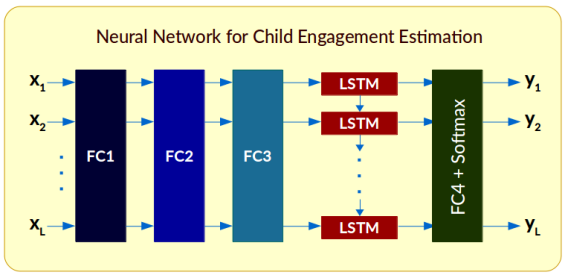
Having fused the pose detections of multiple views, we interpolate the missing values using the previous estimates and then smooth the output using a simple low-pass filter. The points then undergo a final rotation, so that the coordinate axes co-align with the edges of the room.

B. Feature Extraction

The child's pose is particularly useful when examined in relation to the robot's position. Therefore, we would like to estimate the 3D keypoints within the robot's frame. Since the Nao robot isn't equipped with any localization sensors, we must estimate its position with respect to the world coordinates through other means. To this end, we detect the robot in the color stream of one of the cameras, and infer its 3D position via an inverse camera projection. The robot's localization is thus performed fully autonomously, alleviating the need to provide an initial position and monitor its trajectory withing the rooms.

We fine-tuned the YOLOv3 detection network [33] to detect the robot's head on a set of 100 manually annotated images. Using this network, we then detected the robot position in all video frames. An example is shown in Fig. 4. Paired with the depth images, we converted the detections to 3D points. The robot detections also contained noise and missing values, and were subjected to a similar procedure as the keypoints, ie. missing value interpolation and smoothing. We also rejected erroneous detections if they lay outside a certain expected range, based on the limitations of the robot's movement. The body 3D keypoints are transformed to the detected robot frame each time instant, in order to capture the child's movements and transitions in relation to the robot.

Aside from the child's pose keypoints, we produce a number of high-level features that are expected to assist the classification process (Fig. 5). These include the angle between the child's gaze and the robot, the angle between the child's body facing and the robot and the distance of the child's hands from their respective shoulders. The gaze direction is calculated from the detected facial keypoints, by taking the ear-to-ear vector in the 2D plane and rotating it 90 deg. The body facing is calculated in a similar fashion from the shoulder keypoints. From the two resulting angles, we subtract the robot-to-child angle, which is calculated using the keypoint center of mass and the detected robot position. The high-level features are concatenated with the



(a) Network architecture: fully connected layers FC1, FC2, FC3, a single LSTM layer and a final FC4 layer coupled with a softmax function on the output.

Layer	Output	Includes
FC1	(N,L,2C)	Dropout + ReLU
FC2	(N,L,2C)	Dropout + ReLU
FC3	(N,L,2C)	Dropout + ReLU
LSTM	(N,L,C)	-
FC4	(N,L,K)	-

(b) Layer outputs sizes: N is the batch size, L is the sequence length, C is the hidden states size and K is the number of classes.

Fig. 6: (a) Architecture of neural network used to classify child engagement. (b) Network layer details.

keypoint values relative to the robot, mentioned above, to form the input data to the classifier.

C. Engagement Estimation

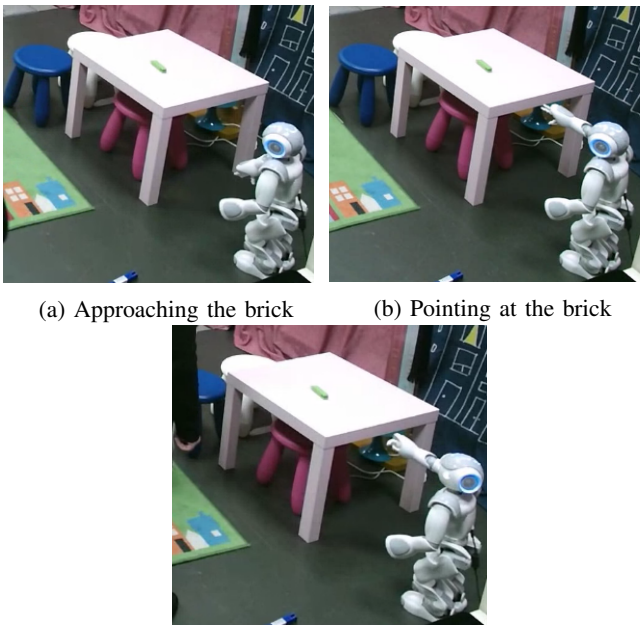
A key observation worth noting is that the degree of engagement heavily depends on temporal information. One reason for this is that the child tends to display the same level of interest over the course of a few-second interval. More importantly, however, the child's movement and actions contain rich information which can be exploited. For example, if the child is constantly shifting their gaze, this is usually an indication of disengagement, whereas a steady focus signifies a higher level of interest. By choosing suitable machine learning algorithms capable of capitalizing on temporal data, we can expect a notable improvement over simply classifying each segment individually.

We use a deep NN to classify the engagement level over time, the architecture of which can be seen in Fig. 6a. The network consists of three fully connected (FC) layers, a Long Short-Term Memory (LSTM) network [34] and a final FC layer, with a softmax function applied to the output to produce a probability score for each class. LSTM networks are a certain type of recurrent neural network that are known to be well-suited to dealing with time-varying data.

The network is fed a sequence of L inputs. We group the input features described earlier into segments of 5 frames, over which we compute the mean and standard deviation, thus further reducing noise and avoiding training on spurious data points, which can cause over-fitting. This gives us an input vector $\mathbf{x}_t$ for each segment $t \leq L$, with a dimension of

$$D = 2 \cdot [3 \cdot (18 \text{ keypoints}) + 4 \text{ high-level features}] = 116$$

The output $\mathbf{y}_t(\mathbf{x}_{1:t}, \mathbf{W})$ is a function of the previous t inputs in the sequence and the weights $\mathbf{W}$ of the network, with a dimension equal to the number of classes ($K = 3$). The FC layers produce linear combinations of their inputs, operating on each sequence point individually. The softmax layer ensures that the elements of $\mathbf{y}_t$ are positive and sum to one. The data are fed in batches of size N. The output dimensions of each layer are shown in the table in Fig. 6. The variable C is the size of the LSTM's hidden state. We set the output sizes of the first three FC layers to $2 \cdot C$ after trial and error. The first three FC layers are each followed by a dropout layer, with probability 0.5, and a Rectified Linear Unit (ReLU).



(a) Approaching the brick (b) Pointing at the brick

(c) Looking at the child

Fig. 7: Examples of the robot's attempts to elicit the child's attention.

III. Experimental Analysis & Results

A. Experimental setup

1) Setup: The experimental setup can be seen in Fig. 2. As shown, 4 Kinect devices are placed at various angles to capture multiple views of the room. Devices 1 and 2 capture the room from either side, device 3 from above and device 4 from the front. The Nao robot [35] was chosen for the task because it is capable of a wide range of motion, and its human-like features make it suitable for child-robot interaction. One or more bricks were placed either on a table in front of the robot, or on the floor close by. The child is free to move around the room as they please.

2) Session Description: The experiments evolved as follows. The robot approached one of the bricks and displayed the intention of picking up the brick, without being able to actually grasp it. With a series of motions it attempted to capture the child's attention and prompt the child to hand over the brick. These motions included pointing at the brick, opening and closing its hand, alternating its gaze between

Method	Mean F-Score	Accuracy	Balanced Accuracy
Majority class	27.90	71.97	33.33
Gaze LSTM	32.78	71.58	36.10
SVM	54.79	68.27	58.61
RF	56.41	68.60	61.78
3FC+LSTM	**62.18**	**77.11**	**61.88**

TABLE I: Performance results of different algorithms on the data. Results are averaged across folds of leave-one-out cross-validation.

the child and the brick and a combination of head turning and hand movement. Example snapshots of such motions are shown in Fig. 7. If after a certain length of time the child failed to understand the robot's intent, the robot proceeded to ask the child verbally. Once the brick was successfully handed over, the robot thanked the child and in some cases looked for another brick to grasp.

3) Data Collection: We recorded a total of 25 sessions, resulting in approximately 85 min of recorded RGB-D data for the four cameras. The children were aged 6-10 years old, 15 male and 10 female. The videos were then given to experts for annotation, according to the scheme described earlier.

B. Experimental Validation

We evaluate the method described above trained on the recordings of the children. Since we only have 25 videos, rather than splitting the set into training and testing subsets, we carry out the evaluation via leave-one-out cross-validation.

1) Implementation: We implemented the NN described in section II using the PyTorch library. The network was trained from scratch, with an initial learning rate of 0.1, momentum 0.5 and weight decay 10^{-6}. We used early stopping on a random subset of the training data, with a patience level of 10 epochs. When the training converged, we dropped the learning rate by a factor of 10. We chose a training batch size of $N = 16$ and set the hidden state size to $C = 560$. The sequence length L was set to 30. These values were chosen after an extensive hyper-parameter search.

Since LSTM networks generally require a large amount of data to train successfully and avoid over-fitting, we employ a few methods of data augmentation. Namely, we add a small amount of Gaussian noise to the mean value of each segment and randomly choose the starting point of each sequence within a range of 2 seconds. We observed a further improvement when training the FC layers first, and then freezing their weights, adding the LSTM module and training the remaining network. This forces the initial layers to produce informative outputs with regard to each individual segment, which the LSTM can then utilize to extract meaningful temporal information. Additionally, since we observed occasional spikes in the network's gradients, we performed gradient clipping on the LSTM layer by capping the gradient norms at the value of 0.1.

The final classification is performed on 1-second segments, by process of a majority vote within the segment. At a frame rate of 30 fps, each second contains 6 smaller segments. This seemed a logical compromise between over-sampling the data points and segmenting the temporal stream too crudely to be of use.

Net Architecture	Mean F-Score	Accuracy	Balanced Accuracy
3 · FC + LSTM	**62.18**	**77.11**	**61.88**
2 · FC + LSTM	56.23	71.86	58.46
3 · FC + 2 · LSTM	54.78	70.60	56.30
2 · FC + 2 · LSTM	54.45	69.71	56.91

TABLE II: Cross-validation results for different network architectures.

Parameters	Mean F-Score	Accuracy	Balanced Accuracy
N=8, L=30	58.75	74.41	58.40
N=16, L=30	**62.18**	**77.11**	**61.88**
N=32, L=30	55.36	69.34	58.68
N=16, L=10	47.21	61.68	51.58
N=16, L=60	48.25	57.32	56.86

TABLE III: Results for different hyper-parameter values.

The classes were notably imbalanced, with 281 segments belonging to class 1, 2578 to class 2 and 745 to class 3. Though we also experimented with under-sampling and over-sampling, the best results were achieved using a weighted cross-entropy loss during training:

$$\mathscr{L} = -\frac{1}{NL}\sum_{j}^{N}\sum_{t}^{L}\mathbf{w}_{c_{j,t}}\log\mathbf{y}_{c_{j,t}}(\mathbf{x}_{j,1:t},\mathbf{W}) \tag{1}$$

where y is the network output, c_j denotes the class of the j-th sample in the minibatch and $\mathbf{w}$ is a vector containing the weights for each class. We set $\mathbf{w} = (9.16, 1.00, 3.42)$, based on the appearance frequencies of each class in the dataset.

2) Evaluation Metrics: Due to the large class imbalance, the standard accuracy measure is not very informative. Therefore, we use two other measures of performance. The first is the average F-Score across all three classes, which is high only when both the precision and recall of each class is high. The second is the balanced accuracy of [36], given by:

$$\frac{1}{K}\sum_{c=1}^{K}\frac{TP_c}{TP_c + FP_c} \tag{2}$$

where TP_c and FP_c denote the true and false positives respectively of class c, and $K = 3$ is the number of classes.

3) Results: In Table I we compare the LSTM-based network against other popular classifiers, in particular a Support Vector Machine (SVM) and a Random Forest (RF). The SVM uses an RBF kernel with a regularization weight of $C = 100$ and a kernel coefficient of $\gamma = 0.01$. The RF consists of 10 trees with a maximum depth of 10. For fair comparison, the hyper-parameters of both classifiers were tuned via an equally extensive grid search as the LSTM network. We also include two baseline approaches. The first consists of the results if the majority class (class 1) is always predicted. The second is another LSTM-based network trained solely on the gaze direction feature, which is a commonly used feature in literature [37].

Notice that the proposed method outperforms all other classifiers, confirming our hypothesis that exploiting temporal relations in the input data can lead to better results. The SVM and the RF fail to capture temporal continuity

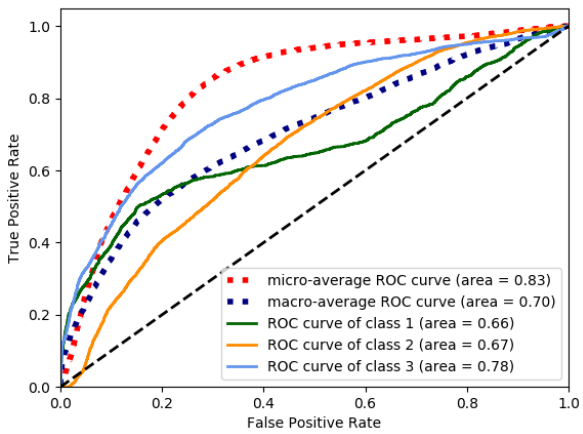


Fig. 8: ROC curves of network with optimal hyper-parameters.

and logical transitions between engagement states[1]. Both classifiers perform significantly better than simply predicting the majority class, however, meaning that even stationary pose information is partially descriptive of the engagement level. On the other hand, the network trained only on gaze direction performs rather poorly, since (a) the gaze estimate contains a lot of noise which the network can't deal with using redundant variables, and more importantly (b) it doesn't capture information such as arm extentions, walking and other movements.

In Table II we evaluate some other network architectures that we also tried. We experimented with the removal of the third FC layer (rows 2 and 4) and with the addition of a second LSTM layer (rows 3 and 4). As shown, the chosen architecture provides notably better results across all metrics. The additional LSTM layer causes over-fitting, rather than learning any deeper information in the data. The use of dropout allows a deeper network, with the addition of the third FC layer boosting the performance by a large margin.

In Table III we provide a comparison of different hyper-parameter values. The optimal sequence length is small enough to allow the training set to be divided into enough sequences to avoid over-fitting, but large enough to capture long term dependencies in the data. A batch size of 16 also provides a compromise between finely sampling the training data in each iteration and avoiding local minima while training. In Fig. 8 we can see the ROC curves for each class separately along with the micro and macro-average ROC curves. The macro-average number of True/False positives is simply the average of each class. The micro-averages, on the other hand, are weighted by the number of samples in each class. Therefore, the micro-average curve is better suited due to the large class imbalance, and our network achieves an Area Under Curve of 0.83 regarding this ROC.

Finally, in order to examine which features are of greater importance to the model, we calculate the gradient of the loss function with respect to each input variable. A higher gradient means the network is more sensitive to changes of the input variable, indicating a higher importance. Due to the high correlation between inputs, the network focused on different variables each time it is trained. Therefore, we trained the model 100 times and averaged the gradients calculated in each run. The results are shown in Fig. 9. We can see that both the high-level features and the keypoint positions contribute to the algorithm's output, justifying our pose-based approach to the problem. While the child's gaze is the most informative variable, a number of other inputs also score similarly. The child's body facing can sometimes complement the gaze, while the 2D position conveys how close the child is to the robot. The limb positions allow the network to exploit temporal differences to detect walking and reaching movements. Also note that although the segment standard deviations are less important than the mean values, they are not entirely uninformative, as they contain information regarding abrupt movements that can be indicative of the child's reactions to the robot's actions.

As we see from the results above, the proposed deep network architecture can learn and accurately track the child's engagement based on their pose variation during the proposed freely interaction task. The developed system can be further improved with the presence of more annotated data and can easily be adapted to a range of CRI tasks that require body movement on the child's behalf. The whole engagement module can be integrated alongside with child's speech, action or emotion recognition modules in order to create next generation social robots that can detect and recognize the children's behavior.

IV. Conclusions & Future Work

In this work we proposed, by taking advantage of recent progress in deep learning, a method of child engagement estimation during child-robot collaboration without restricting their movement. The use of child pose data, in conjunction with an LSTM-based neural network, proved to be effective towards this goal. This is especially important considering the difficulty of the problem. Differences in child behavior and personality, a wide range of possible motions and actions and various technical challenges all contribute to this difficulty. The concept of engagement is not rigidly defined, thus making the task hard even for humans. Despite this, we achieve relatively high evaluation metrics across a dataset of 25 children. An important direction for future work will be to test and adapt the system to children affected by ASD. Naturally, this imposes a further challenge, as ASD children act very differently to children in typical development.

Acknowledgment

The authors would like to thank psychologists Asimenia Papoulidi and Christina Papailiou for their annotations, and our colleagues Niki Efthymiou and Panagiotis Filntisis for helping organize and record the experiments.

References

[1] G. Gordon, C. Breazeal, and S. Engel, "Can children catch curiosity from a social robot?" in *HRI '15*.

[2] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, p. eaat5954, 2018.

[3] "BabyRobot project." [Online]. Available: http://babyrobot.eu

[1]See supplementary video for examples.

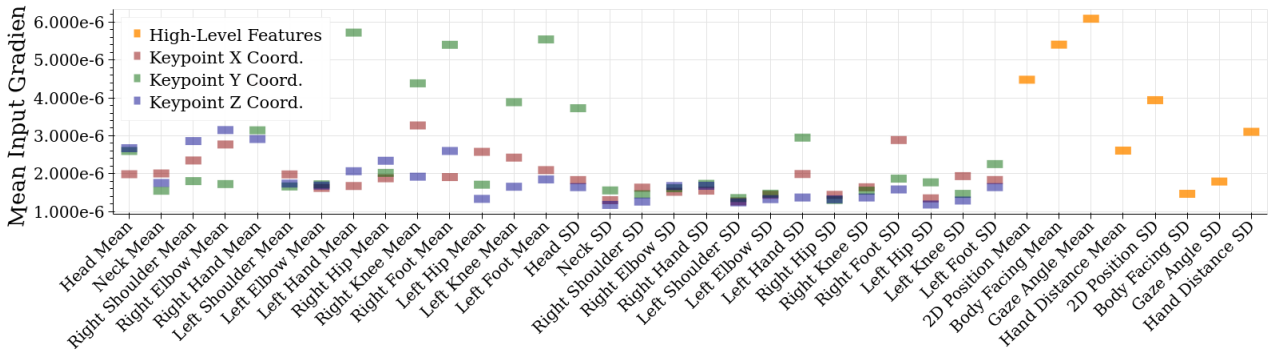


Fig. 9: Feature importance of input variables based on their average input gradient over 100 trained networks (SD=standard deviation).

[4] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos, "Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots," in *ICRA '18*.

[5] A. Tsiami, P. P. Filntisis, N. Efthymiou, P. Koutras, G. Potamianos, and P. Maragos., "Far-field audio-visual scene perception of multi-party human-robot interaction for children and adults," in *ICASSP '18*.

[6] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos., "Multi-view fusion for action recognition in child-robot interaction," in *ICIP '18*.

[7] A. Othman and M. Mohsin, "How could robots improve social skills in children with autism?" in *ICTA '17*.

[8] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, "Social robots for education: A review," 2018.

[9] C. A. G. J. Huijnen, M. A. S. Lexis, R. Jansens, and L. P. de Witte, "How to implement robots in interventions for children with autism? a co-creation study involving people with autism, parents and professionals," *Journal of Autism and Developmental Disorders*, 2017.

[10] S. M. Anzalone, S. Boucenna, S. Ivaldi, and M. Chetouani, "Evaluating the engagement with social robots," *IJSR*, 2015.

[11] H. L. O'Brien and E. G. Toms, "What is user engagement? a conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*.

[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 *HRI '16*.

[14] J.-D. Boucher, U. Pattacini, A. Lelong, G. Bailly, F. Elisei, S. Fagel, P. Dominey, and J. Ventre-Dominey, "I reach faster when i see you look: Gaze effects in humanhuman and humanrobot face-to-face cooperation," *Frontiers in Neurorobotics*, 2012.

[15] S. Ivaldi, S. Lefort, J. Peters, M. Chetouani, J. Provasi, and E. Zibetti, "Towards engagement models that consider individual factors in hri: On the relation of extroversion and negative attitude towards robots to gaze and speech during a human–robot assembly task," *IJSR*, 2017.

[16] M. E. Foster, A. Gaschler, and M. Giuliani, "Automatically classifying user engagement for dynamic multi-party human–robot interaction," *IJSR*, 2017.

[17] C. Zaga, M. Lohse, K. P. Truong, and V. Evers, "The effect of a robot's social character on children's task engagement: Peer versus tutor," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds., 2015.

[18] T. Schodde, L. Hoffmann, and S. Kopp, "How to manage affective state in child-robot tutoring interactions?" in *ICCT '17*.

[19] A. Chorianopoulou, E. Tzinis, E. Iosif, A. Papoulidi, C. Papailiou, and A. Potamianos, "Engagement detection for children with autism spectrum disorder," in *ICASSP '17*, 2017.

[20] A. Tapus, A. Peca, A. Aly, C. Pop, L. Jisa, S. Pintea, A. S. Rusu, and D. O. David, "Children with autism social engagement in interaction with nao, an imitative robot: A series of single case experiments," *Interaction Studies*, 2012.

[21] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, and R. W. Picard, "Measuring engagement in robot-assisted autism therapy: A cross-cultural study," *Frontiers in Robotics and AI*, 2017.

[22] 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*, 2017.

[23] H. Javed, M. Jeon, and C. H. Park, "Adaptive framework for emotional engagement in child-robot interactions for autism interventions," in *UR '18*.

[24] Z. Zheng, H. Zhao, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Design, development, and evaluation of a noninvasive autonomous robot-mediated joint attention intervention system for young children with asd," *IEEE Trans. on Human-Machine Systems*, 2018.

[25] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science Robotics*, 2018.

[26] M. Khamassi, G. Velentzas, T. Tsitsimis, and C. Tzafestas, "Robot fast adaptation to changes in human engagement during simulated dynamic social interaction with active exploration in parameterized reinforcement learning," *IEEE Trans. on Cognitive and Developmental Systems*, 2018.

[27] M. Khamassi, G. Chalvatzaki, T. Tsitsimis, G. Velentzas, and C. S. Tzafestas, "An extended framework for robot learning during child-robot interaction with human engagement as reward signal," in *3rd Workshop BAILAR, in ROMAN '18*.

[28] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR '17*.

[29] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR '17*.

[30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR '16*.

[31] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *ICRA '18*.

[32] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *Trans. on PAMI 1992*.

[33] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv:1804.02767*, 2018.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, 1997.

[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 *ICRA '09*.

[36] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *ICPR '10*.

[37] A. Saran, S. Majumdar, E. S. Shor, A. Thomaz, and S. Niekum, "Human gaze following for human-robot interaction," in *Proc. International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 8615–8621.