To Transfer or Not To Transfer: Engagement Recognition within Robot-assisted Autism Therapy

Nazerke Rakhymbayeva
Department of Robotics & Mechatronics
School of Engineering & Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
nazerke.rakhymbayeva@nu.edu.kz

Zarema Balgabekova
Department of Robotics & Mechatronics
School of Engineering & Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
zarema.balgabekova@nu.edu.kz

Mukhamedzhan Nurmukhamed
Department of Robotics & Mechatronics
School of Engineering & Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
mukhamedzhan.nurmukhamed@nu.edu.kz

Karina Burunchina
Department of Robotics & Mechatronics
School of Engineering & Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
karina.burunchina@nu.edu.kz

Wafa Johal
School of Computing & Information Systems
Faculty of Engineering & Information Technology
The University of Melbourne
Melbourne, Victoria, Australia
wafa.johal@unimelb.edu.au

Anara Sandygulova
Department of Robotics & Mechatronics
School of Engineering & Digital Sciences
Nazarbayev University
Nur-Sultan, Kazakhstan
anara.sandygulova@nu.edu.kz

***Abstract*—Social robots are increasingly being used as a mediator in robot-assisted autism therapy to improve children's social and cognitive skills. Engagement is one of the key measurements used to evaluate the therapeutic interventions' effect on children. While "engagement" is broadly used, it has been challenging to find a consensus about its definition in the community. With this paper, we explore the use of a data-driven approach to investigate the extent to which a model on engagement built on one dataset transfers to another. We utilized two publicly available datasets of engagement recognition, namely PInSoRo and Qamqor datasets, with an attempt to achieve a higher accuracy taking into account the transferred knowledge. The accuracy of 83.18% was obtained on the PInSoRo dataset of child-child interactions with face and body keypoints. We have used the methodology of transfer learning to improve the classification accuracy on the Qamqor dataset. The best result obtained is a 71.89% accuracy on the Qamqor dataset. This suggests that more data with similar keypoints is needed to achieve better accuracy when utilizing transfer learning from one dataset to another dataset.**

***Index Terms*—transfer learning, children with ASD, social robots, robot-assisted therapy, binary classification, multi-class classification.**

I. Introduction

The efficiency of integrating robots to improve and support traditional therapies for children with Autism Spectrum Disorder (ASD) has been investigated in the most recent human-robot interaction (HRI) studies [1]. These studies have shown that the majority of children with autism willingly engaged with social robots [2].

Human behaviour is naturally multi-modal. People use eye gaze, hand gestures, body posture, and tone of voice to manage social interactions. Machine Learning (ML) can recognize these behavioural cues using video and audio recordings. The behavioural cues are converted into input features before linked with engagement labels. These data pairings are then used to train models, which develop a functional form that translates input characteristics to engagement labels (output). In addition, the labelling of data with engagement is challenging as it is a complex and dynamic dimension [3].

In our research context, the term "engagement" indicates a child's purposeful participation during the interaction with the robot and/or the therapist. In the past studies with typically developing children, behavioral engagement was defined as "concentrating on the task at hand and willingness to remain focused" [4]. In the studies on robot-assisted therapy of children with autism by Kim et al. [5] and Rudovic et al. [6], engagement was assigned a label on a 0–5 Likert scale, with each level representing a series of pre-defined responses by the child to the task, robot's or therapist's prompts. To recognize engagement, Rudovic et al. (2018) [7] utilized OpenPose [8] keypoints, audio, labels of valence, demographic information and physiological measures (heart rate, electrodermal activity and temperature).

978-1-6654-0731-1/22/$31.00 ©2022 IEEE

Data-driven approaches are challenging to use in HRI as datasets are often small and specific to a particular context. Transfer Learning (TL) allows to re-train a model with another dataset with a different class distribution. In this paper, we analyze to what extent transfer learning can be helpful when using open HRI datasets to improve the accuracy of engagement labels recognition of autistic children with social robots in the Qamqor dataset.

II. Related Datasets

Due to ethical concerns, it is challenging to find HRI datasets featuring ASD children. For this reason, we choose to investigate how other openly available datasets could be used to improve the accuracy of engagement recognition in the Qamqor dataset (a dataset featuring children with ASD interacting with social robots). We use multimodal data (video, audio, depth, OpenPose keypoints [8]) as well as engagement labels.

A. Qamqor

The Qamqor project's goal was to develop a dataset using video recordings of 36 children with autism during robot-assisted autism treatment research. Qamqor dataset consists of extracted child's feature keypoints from video sequences: 2D data of 25 keypoints (joints) of the body and legs, 21 keypoints for each hand, and 70 facial keypoints. These features were extracted via the OpenPose library [8]. The dataset comprises 194 therapy sessions and about 48 hours of videos [9], [10]. Two independent raters annotated all videos, and engagement labels were coded from 1 to 5 relative to the timing of the applications. The annotators had an agreement score on 20% of cross-coded data computed from pair-wise ICC of the coders equal to 82.6%. The total number of hours of coded video was 48 hours and 34 minutes. The accuracy of engagement recognition was 73.62% on binary class classification.

B. MHHRI

The Multimodal Human-Human-Robot-Interactions (MHHRI) dataset [11] was collected during a controlled interaction study between two human participants and a robot. Participants were asking personal questions to each other. Sessions were recorded using two static and two dynamic cameras (mounted on the participants' heads) and two biosensors. The dataset also provides labels for personality traits and perceived engagement with their partners (self-labelled).

C. DREAM

Billing et al. [12] presented the DREAM dataset, which consists of more than 300 hours of video recording of 61 children with autism during two therapy conditions: with the robot and with humans. Each session was recorded using three RGB cameras and two depth (Kinect) cameras. The dataset's public release includes participants' age, gender, autism diagnostic (ADOS-2 scores), 3D recordings of the body motion, head position and orientation, and eye gaze characteristics.

D. PInSoRo

PInSoRo's data collection approach was based on an engaging but purposely under-specified free-play interaction. This way, they could catch a broad set of behavioural tendencies in everyday social interactions between children. The final dataset contains 45 hours of hand-coded interaction of 45 child-child and 30 child-robot pairs. The collection comprises thoroughly calibrated video frames, 3D records of the face, structural characteristics, entire audio files, and game engagements, in addition to annotations of social constructs [13].

E. MIT dataset

Rudovic et al. [7] proposed a multi-modal (audio, video, and autonomic physiology) dataset of 35 children (ages 3 to 13) with autism from two cultures (Asia and Europe) and achieved an average agreement of about 60% with human experts in the estimation of affect and engagement.

III. Methodology

In order to explore ways to improve the classification accuracy of engagement labels on the Qamqor dataset (initially at 73.62%), the idea was to transfer knowledge obtained on a related task.

After analyzing the content of the above datasets, we decided to use the PInSoRo dataset to train our initial model of engagement recognition. The PInSoro dataset was selected as it is publicly available and provides an engagement label for each frame as it is done in Qamqor. The features extracted from the PInSoRo are also very similar to the ones in Qamqor.

A. Pre-processing datasets

The first step was to perform pre-processing of the PInSoRo and Qamqor datasets. The PInSoRo dataset consists of .csv files made for each session: two of them differ in the interactive partner of the child (robot vs another child). We created three .csv files: 1) child-robot interaction, 2) child-child interaction, 3) combined 1 and 2 files. Next, we removed the rows with 'NAN' values of those keypoints which OpenPose did not recognize. As a result, several keypoints that were not shared by both datasets were deleted.

The Qamqor dataset was pre-processed by rearranging columns to achieve the same order of features as in the PInSoRo dataset.

Next, we created two modalities: M1 - face keypoints only, and M2 - both face and body keypoints. Additionally, we decided to recognize engagement as 1) a binary class and 2) a multi-class recognition task (ratings from 1-5). For the binary classification, labels 1 and 2 were classified as disengaged (class 0), whilst labels from 3 to 5 as engaged (class 1).

Finally, the datasets were split into test, validation, and train sets with face and body keypoints as input features and engagement labels as an output.

B. Neural Network architecture

After processing the data, we designed the neural networks with the following architecture: input, three hidden layers, and output layer. The hidden layers use the ReLu activation function, whilst the output layer uses the Softmax activation function for the PInSoRo dataset with multi-class classification and Sigmoid activation function for the PInSoRo dataset binary class classification. For the loss function, we used categorical cross-entropy for predicting the multi-class output (from 1 to 5) and binary cross-entropy for binary classification (0 or 1). Also, an RMSprop optimizer was used in both cases.

C. Transfer Learning

Each model (multi-class and binary) was trained on the PInSoRo datasets with different modalities ($M1$ and $M2$) and conditions (child-robot and child-child). The next step was to freeze the layers of the output model to avoid re-training. Then, the new models were compiled by transferring the knowledge from the PInSoRo models. The resulted neural network was trained on the Qamqor datasets with different modalities (Fig. 1). The Qamqor test set evaluated the achieved accuracies. Finally, the confusion matrices were plotted based on each transfer model.

IV. Results

We performed transfer learning from the PInSoRo dataset with different conditions: child-robot, child-child, and the combination of both. Furthermore, we checked the method using different modalities: M1 - face features, M2 - face and body features. Finally, we tested binary and multi-class classification of engagement recognition. Unfortunately, the multi-class results produced very poor performances; thus, we only report the binary classification results: engaged vs disengaged. Results are presented in Table I.

TABLE I: Results of accuracy with the binary classification that was obtained on the PInSoRo dataset (left) and the Qamqor dataset after transfer learning (right)

Conditions	PInSoRo		Qamqor after TL	
	M1	M2	M1	M2
Child-robot	69.00%	75.12%	62.26%	58.83%
Child-child	80.78%	83.18%	63.61%	64.50%
Both	71.65%	75.04%	71.89%	61.59%

A. Child-robot condition

1) M1 modality: In the case when only the keypoints of the face were used, the accuracy of the model for the child-robot condition of the PInSoRo dataset was 69.00%. The confusion matrix for binary classification is presented in Figure 2a and for multi-class classification in Figure 3a. As can be seen, the majority of labels were classified correctly. For the Qamqor dataset, after transferring knowledge from the PInSoRo dataset, the accuracy resulted in 62.26%.

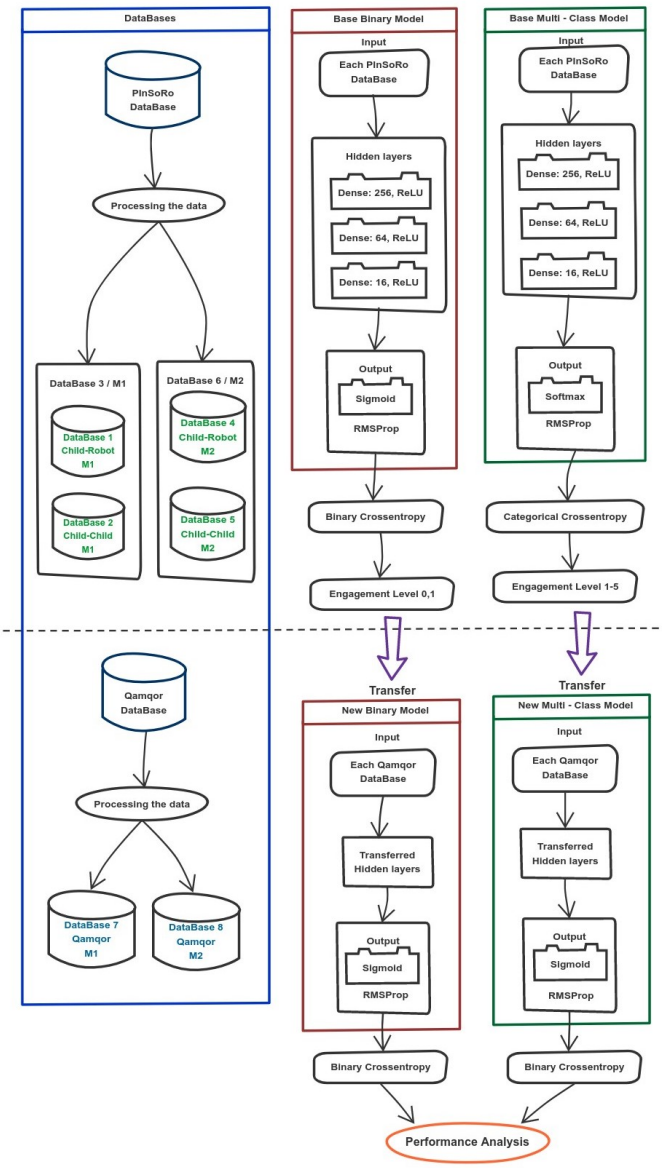
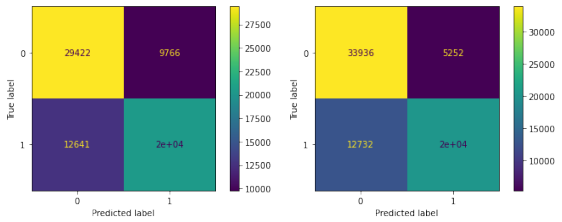


Fig. 1: Methodology schematics

2) M2 modality: When training the model for the PInSoRo dataset using data only for child-robot condition and keypoints of face and body as features, the model's accuracy for testing data was 75.12%. The confusion matrix for binary classification is presented in Figure 2b and for multi-class classification in Figure 3b. After transferring the weights of the obtained model to the Qamqor dataset, we obtained an accuracy of 58.83%.

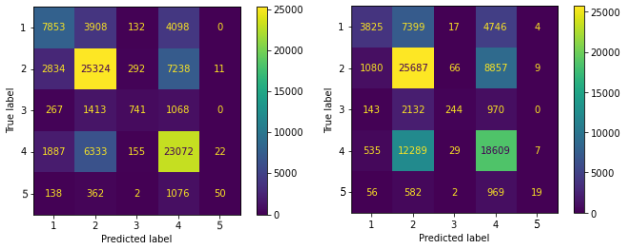
B. Child-child condition

1) M1 modality: When we used only the face keypoints, the accuracy of 80.78% was obtained for the PInSoRo dataset's model with the child-child condition. In the case of the Qamqor dataset, its accuracy became 63.61%.



(a) child-robot/M1 (b) child-robot/M2

Fig. 2: Confusion matrices for the model of PInSoRo dataset with binary classification



(a) child-robot/M1 (b) child-robot/M2

Fig. 3: Confusion matrices for the model of PInSoRo dataset with multi-class classification

2) M2 modality: Using data from child-child condition only and keypoints of face and body as features, we got the accuracy of 83.18% for the PInSoRo dataset's model. After TL, the accuracy for the Qamqor dataset was 64.50%.

C. Combination of child-robot and child-child conditions

1) M1 modality: When limiting features to the keypoints of the face only, we got the accuracy of 71.65% for the PInSoRo dataset with combined conditions.

2) M2 modality: After combining data for child-robot and child-child conditions, the model's accuracy for the PInSoRo dataset when keypoints of face and body are used as features became 75.04%. As for the Qamqor dataset, the accuracy of its model was 61.59% after transfer learning. The results are presented in Table I, the right two columns.

V. Discussion

As can be seen from Table I, there is no perfect correlation between the accuracy for the PInSoRo dataset's and Qamqor dataset's models. The best accuracy for PInSoRo was achieved with the child-child condition because of a larger amount of data used. However, when the data of both conditions were used, the accuracy did not increase. Additionally, in all conditions, the accuracy of the $M2$ modality is higher compared to the $M1$ modality. This suggests that a model with more keypoints provide better engagement recognition in the PInSoRo dataset.

In contrast, for the Qamqor dataset, the accuracy of the $M1$ modality is better than for the M2 modality for child-robot and combined child-robot with child-child conditions. Also, the best accuracy for the Qamqor dataset's model was achieved when both conditions were used. These findings suggest that more data with similar features is needed to improve accuracy for the transfer learning approach.

VI. Limitations

The PInSoRo dataset was collected from recordings of typically developing children, while the Qamqor dataset was collected from recordings of children with ASD.

Furthermore, there was a lot of missing data in the PInSoRo dataset, which could negatively affect the accuracy of the obtained model. For example, the best accuracy that we could obtain was 83.18% when data for only child-child condition and keypoints of face and body were used. Transfer learning, however, should be used when a very accurate model trained on high-quality data is available.

Also, since the PInSoRo dataset does not have features such as the keypoints of left and right hands and has only 18 keypoints of body, it was not possible to use the full capability of the Qamqor dataset, limiting its features to the keypoints of the face and 18 out of 25 keypoints of body. Therefore, there is still room for improvement regarding a dataset from which the knowledge is transferred.

VII. Conclusion and Future Work

In this work, we tested the assumption that transfer learning from a larger dataset (PInSoRo dataset) would improve the accuracy of the engagement recognition for the Qamqor dataset.

It turned out that our assumption failed as the best accuracy obtained without the use of transfer learning for binary classification was 73.13% for $M1$ modality and 73.62% for $M2$ modality on binary class classification. However, we could achieve the best accuracy of only 71.89% with transfer learning from the PInSoRo dataset.

Although transfer learning did not improve the accuracy of the Qamqor dataset, we cannot claim that it does not work at all due to the reasons presented in the Limitations section. The main contribution of the paper is in setting up a methodology for the formulated problem to conduct future research in this area. Given this, we have identified several suggestions for future work:

1) All other available datasets should be explored. For example, the DREAM and the MHHRI datasets can be used instead of the PInSoRo dataset. It is assumed that the DREAM dataset can provide more conclusive results as it was collected from the autistic children similarly to the Qamqor dataset. Ideally, two datasets between which transfer learning is applied should have as many of the same features as possible.

2) More neural networks architectures can be tried, varying the numbers of hidden layers and the number of nodes in each layer. The method might benefit from using Deep Learning and Recurrent Neural Networks.

Acknowledgements

This work was supported by the Nazarbayev University Collaborative Research Program grant (award number is 091019CRP2107).

REFERENCES

[1] B. Szymona, M. Maciejewski, R. Karpiński, K. Jonak, E. Radzikowska-Büchner, K. Niderla, and A. Prokopiak, "Robot-assisted autism therapy (raat). criteria and types of experiments using anthropomorphic and zoomorphic robots. review of the research," *Sensors*, vol. 21, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3720

[2] M. A. Saleh, F. A. Hanapiah, and H. Hashim, "Robot applications for autism: a comprehensive review," *Disability and Rehabilitation: Assistive Technology*, vol. 16, no. 6, pp. 580–602, 2021, pMID: 32706602. [Online]. Available: https://doi.org/10.1080/17483107.2019.1685016

[3] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.00092

[4] B. Ge, H. W. Park, and A. Howard, "Identifying engagement from joint kinematics data for robot therapy prompt interventions for children with autism spectrum disorder," vol. 9979, 11 2016, pp. 531–540.

[5] E. Kim, R. Paul, F. Shic, and B. Scassellati, "Bridging the research gap: Making hri useful to individuals with autism," *Journal of Human-Robot Interaction*, vol. 1, 08 2012.

[6] O. 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," *Front. Robot. AI*, vol. 4, no. 36, July 2017. [Online]. Available: https://doi.org/10.3389/frobt.2017.00036

[7] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Science*, vol. 3, 02 2018.

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[9] Z. Telisheva, A. Zhanatkyzy, A. Turarova, N. Rakhymbayeva, and A. Sandygulova, "Automatic engagement recognition of children within robot-mediated autism therapy," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 471–472. [Online]. Available: https://doi.org/10.1145/3371382.3378390

[10] N. Rakhymbayeva and A. Sandygulova, "Transfer learning of engagement recognition within robot-assisted therapy for children with autism," in *AAAI*, 2021.

[11] O. Celiktutan, E. Skordos, and H. Gunes, "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement," *IEEE Transactions on Affective Computing*, 2017.

[12] E. Billing, T. Belpaeme, H. Cai, H.-L. Cao, A. Ciocan, C. Costescu, D. David, R. Homewood, D. Hernandez Garcia, P. Gómez Esteban, H. Liu, V. Nair, S. Matu, A. Mazel, M. Selescu, E. Senft, S. Thill, B. Vanderborght, D. Vernon, and T. Ziemke, "The dream dataset: Supporting a data-driven study of autism spectrum disorder and robot enhanced therapy," *PLOS ONE*, vol. 15, no. 8, pp. 1–15, 08 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0236939

[13] 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, no. 10, pp. 1–19, 10 2018. [Online]. Available: https://doi.org/10.1371/journal.pone.0205999