CultureNet: A Deep Learning Approach for Engagement Intensity Estimation from Face Images of Children with Autism

CultureNet: A Deep Learning Approach for Engagement Intensity Estimation from Face Images of Children with Autism

Ognjen (Oggi) Rudovic[1], Yuria Utsumi[1], Jaeryoung Lee[2], Javier Hernandez[1],
Eduardo Castelló Ferrer[1], Björn Schuller[3,4], Rosalind W. Picard[1]
[1]Massachusetts Institute of Technology, USA, [2]Chubu University, Japan, [3]Imperial College London, UK, [4]Augsburg University, Germany
{orudovic,yutsumi,javierhr,ecstll,picard}@mit.edu
jaeryounglee@isc.chubu.ac.jp, bjoern.schuller@imperial.ac.uk

***Abstract*— Many children on autism spectrum have atypical behavioral expressions of engagement compared to their neurotypical peers. In this paper, we investigate the performance of deep learning models in the task of automated engagement estimation from face images of children with autism. Specifically, we use the video data of 30 children with different cultural backgrounds (Asia vs. Europe) recorded during a single session of a robot-assisted autism therapy. We perform a thorough evaluation of the proposed deep architectures for the target task, including within- and across-culture evaluations, as well as when using the child-independent and child-dependent settings. We also introduce a novel deep learning model, named CultureNet, which efficiently leverages the multi-cultural data when performing the adaptation of the proposed deep architecture to the target culture and child. We show that due to the highly heterogeneous nature of the image data of children with autism, the child-independent models lead to overall poor estimation of target engagement levels. On the other hand, when a small amount of data of target children is used to enhance the model learning, the estimation performance on the held-out data from those children increases significantly. This is the first time that the effects of individual and cultural differences in children with autism have empirically been studied in the context of deep learning performed directly from face images.**

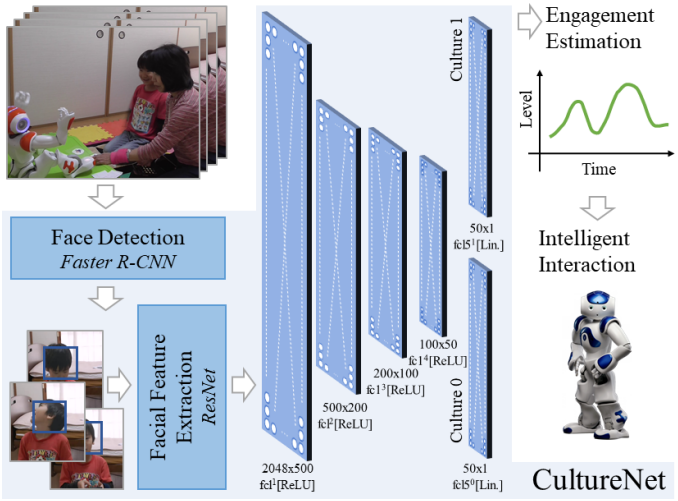

Fig. 1: Automated estimation of engagement directly from face images of children with ASC during a robot-assisted therapy, using the proposed "culturalized" deep models.

I. INTRODUCTION

Autism Spectrum Condition (ASC) is a complex neurodevelopmental condition characterized by socio-emotional communication challenges, as well as repetitive and stereotyped behaviours [5]. Some of the social challenges arise from different motor abilities and subsequent challenges producing speech and nonverbal expressions. These differences usually manifest early in life, prompting the need to detect them early and give the child enhanced opportunities to develop these skills. Technology that can engage the learner successfully can provide longer periods of practice, with opportunities to gain important knowledge for cognitive and social development [24], [26]. A challenge when working with population with ASC is that their personal displays of engagement also can vary largely across children and are usually perceived as of low intensity, compared to those of their neurotypical peers [33]. This makes engagement recognition extremely hard to perform. Another important aspect of perceived engagement displays are cultural differences among children with ASC. Yet, there is very little research examining ASC characteristics in cross-cultural settings [30], [37]. Nevertheless, cross-cultural analyses are particularly important for ASC since these may provide better insights into the perception of symptoms and expression of different behaviors, including engagement.

Measuring engagement of children with ASC during therapy offers opportunities to provide personalized interactions [13] and interventions [3] that, in turn, can help to optimize the learning outcomes. However, current methods for engagement measurement require experienced observers and/or self-reports, which are expensive and time intensive, and/or are not appropriate for non-verbal children. In addition, these approaches are subjective and may require a lot of time spent with the child before he/she can be accurately read, especially by a new person who is unfamiliar with the child and/or less experienced in performing the therapy. To help address these challenges, there is an ever-growing need for automated methods for measurement of engagement of children with ASC.

Most of the existing works on automated analysis of behavioral cues of children with ASC focus on child-therapist sessions [21], often assisted by robots to facilitate

the children's engagement [37]. These studies attempt to automate the engagement measurement from different behavioral modalities, including the children's facial expressions [4], body movements [14], autonomic physiology [21], and vocalizations [6]. This is usually conducted by applying machine learning algorithms to the sensory inputs (e.g., audio-visual) capturing different behavioral modalities [8], [18]. These inputs are then transformed into feature representations, which are used to train supervised models, where human experts provide labels for target states (e.g., engagement levels) of the children by examining each child's audio-visual recordings. To this end, most of the prior work exploits the use of non-parametric classification models, such as Support Vector Machines (SVM) [9]. Yet, traditional methods such as SVMs, Decision Trees, or Linear Discriminant Analysis [9], among others, require (i) a careful engineering of input features (hand-crafted), (ii) cannot deal efficiently with large feature dimensions, e.g., when using the pixel values from face images as the input. These have recently been addressed by deep learning frameworks [23], [27] that tackle (i)-(ii) by providing learning models that can make inference directly from high-dimensional face images by leveraging efficiently large amounts of training data.

So far, deep learning has shown great success in various machine learning tasks, such as object recognition and sentiment analysis [27], [31]. Furthermore, this has also been shown in the tasks of automated measurement of engagement and affective states directly from face images of neurotypical individuals [22], [32], where the proposed deep architectures were able to generalize well to previously unseen subjects. In the context of autism, deep learning has been applied to autism screening from brain images [20], and affect and engagement estimation from multimodal behavioral cues (facial landmarks, head pose, biosignals, and voice) [36]; yet, it has not been explored in the task of automated measurement of engagement directly from face images of children with ASC. This may particularly be challenging due to the large individual and cultural heterogeneity in image data of this population. Also, most of existing works on analysis of facial cues in autism focus on eye-gaze, blinking, and head-pose [12], [29], which are shown to be a good proxy of joint attention and engagement – the lack of which is pertinent to ASC. Extracting these cues from face images is usually done using detectors specifically built for each facial cue. While this may be cumbersome, it can also miss salient facial cues of importance for engagement estimation.

In this work, we explore the use of deep learning to estimate engagement levels from raw face images of children with ASC. Specifically, we use a cross-cultural dataset of 30 children with ASC participating in a robot-assisted autism therapy [37]. We focus on three main questions that, to the best of our knowledge, have not been explored before: (i) How robust is deep learning in the estimation of engagement levels from face images of children with ASC as we vary the amount of training data examples per child, and test the models on previously unseen data of these children? (ii) How well does it generalize within and across different cultures? (iii) How well can it generalize to new children, compared to when these children are also included in the training set? To this end, we perform a thorough analysis of different deep network architectures. Based on this, we propose a deep learning architecture that can efficiently leverage the multi-cultural data to improve the engagement estimation from face images. The contributions of this work can be summarized as follows:

- We provide the first study that explores the performance of deep learning in the context of automated engagement estimation of children with ASC *directly* from their face images. While deep learning has been explored before in the task of affect estimation from face images of neurotypical population (e.g., [32]), it has not been investigated in the context of autism.
- We perform in-depth analysis of the target deep learning architectures, providing insights into the ability of deep models to deal with highly heterogeneous face-image data, which can arise from both individual and cultural variation in children with ASC.
- We propose a novel personalized deep learning architecture (CultureNet – see Fig.1) that leverages efficiently the data from children with ASC and with a different cultural background. We show that for improving the engagement estimation from face images of children with ASC, it is critical to have access to data from the target culture and child during the model learning.

The rest of the paper is organized as follows. First, we provide a background on the data used in this work, cultural differences in autism, and engagement measurement. Then, we describe the proposed deep learning architectures and their experimental evaluation. Finally, we provide a discussion of the results and conclude the paper.

II. Autism, Culture and Engagement

A. Dataset: Robot-assisted Autism Therapy

In this work, we use the video data from a robot-assisted therapy (with the NAO robot) for children with ASC [37]. The goal of the therapy is to teach neurotypical emotion expressions to children with ASC: a therapist uses images of facial and body expressions of basic emotions (e.g., sadness, happiness, anger, and fear) as shown by typically developing children. To facilitate emotion recognition and imitation by the children, the robot then demonstrates the expression of these emotions. The therapy steps are adopted from the Theory of Mind (ToM) concept [7], designed to teach the perspective taking ("social imagination") – one of the key challenges for many children with ASC. However, the engagement estimation is quite challenging due to the large behavioral differences across the children with ASC from the two cultures examined here. Analysis of their facial expression allows for unobtrusive measures of engagement during the therapy. Success with this difficult task would, in turn, help enable assistive robots that can learn and recognize the children's engagement from their facial expressions and

adapt to them more intelligently (e.g., by changing the exercise, and/or providing feedback/prompts) [25], also providing effective ways to monitor the therapy progress [36].

B. Cultural Differences

The importance of cultural diversity when studying different populations has been emphasized in a number of psychology studies [17], [38]. Several cross-cultural studies highlight that culture-based treatments are crucial for individuals with ASC [15], [40]. For example, cross-cultural supports are argued to be needed for pervasive developmental conditions, including autism [16]. Perepa et al. describe how they investigated cultural context in interventions for children with autism and with a diverse cultural background – British, Somali, West African, and South Asian [34]. They found that the cultural background of the children's parents is highly relevant to their social behavior, emphasizing the importance of transcultural treatments for children with autism. Libin et al. showed that the children's background, such as culture and/or psychological profile, need to be taken into to account when designing the therapy [28]. Likewise, using the dataset employed in this work, our team previously showed that there are statistically significant cultural differences in engagement levels between children with ASC from Asia and Europe [37]. While these works provide evidence about cultural differences in autism, to date, it has not been explored how they impact potential deep learning models of engagement from face images.

C. Engagement Definition

There is a wide gamut of engagement definitions depending on their main focus [24]. Most of the prior work on engagement in human-robot interaction (HRI) relies on binary engagement (engaged vs. disengaged) mainly due to the difficulty in capturing subtle changes in engagement displays. For a thorough comparison of the different engagement definitions, we refer the reader to [2], [24]. We used continuous coding ($[-1, +1]$) of engagement levels that focuses on the task-response time [36]. The reference points were defined as follows: (-1) corresponding to cases when the child completely disengaged from the interaction with the robot and/or therapist, and/or refused to perform the task even after several prompts by the therapist, (0) when the child looked indifferent to the interaction with the therapist and/or robot, and $(+1)$ when the child was fully engaged in the task. For details about the coding process, see [36] and Sec.IV-A.

III. Deep Learning: The Model

We consider the following setting: we are given an image data set $S = \{S^0, S^1\}$ of subjects (children with ASC) from two cultures, $C0$ and $C1$. The data of subjects within each culture are denoted as: $S^c = \{s_1^c, \ldots, s_K^c\}$, where $c = \{0, 1\}$ and K is the number of subjects per culture[1]. Furthermore, the data of each subject are stored as $s_i = \{X_i, y_i\}$, where input images of the subjects $i = 1, \ldots, K$ are stored in $X_i \in \mathcal{R}^{D_x \times D_x}$, where $D_x = 256$ is the length (in pixels) of a detected face region in target images. Note that these examples may be temporally correlated (in the case of video data) or be randomly sampled from independent observations of the subject. Then, each image is associated with a ground-truth label y_i of the engagement level (provided by human coders). In what follows, we first describe how a general deep model is used to estimate target engagement states on held-out face images. Then, we introduce a culturalized model, devised to account for culture-specific differences, while leveraging the data from both cultures.

[1]For notational simplicity, we assume the same number of data per subject and culture.

A. General Deep Model (GenNet)

As the general deep model, we used a deep convolutional network architecture composed of the layers of the ResNet [19], a pre-trained deep network, in combination with additional network layers that we included for the target task, i.e., estimation of engagement levels from face images. The traditional ResNet-50 architecture is composed of multiple three-layer "bottleneck" building blocks, containing 50 (convolutional and dense) layers in total. However, the network's weights are optimized for object classification (e.g., the object categories such as "laptop" and "orange"). Our model learning thus consisted of two steps: (i) fine-tuning of the ResNet weights for faces, followed by (ii) learning of additional network layers for estimation of engagement levels. For (i), we used all of the ResNet layers but replaced the last (i.e., the softmax layer) with a data-uninformed fully-connected dense layer with linear activation. This network was then fine-tuned for face images using a rich dataset (500k+ images) of annotated faces – AffectNet [32], in terms of valence/arousal levels. Once we fine-tuned the network weights (W^{rnet}) for extraction of discriminative facial features, we removed the last regression layer and froze the ResNet weights, the role of which from this point was to perform efficient facial feature extraction. Then, in step (ii), we designed a deep network architecture containing five fully connected layers (fcl), which we added to the output of the fine-tuned ResNet to form a general deep model for engagement estimation in our experiments.

Formally, this architecture receives as input the face images of training subjects (X) and passes the most discriminative (deep) facial features (h_0) in the output of the ResNet. These are then passed through the fcls, where we used the rectified linear unit (ReLU) [27], defined as:

$$h_l = \max(0, W_l h_{l-1} + b_l), \tag{1}$$

where $l = 1, \ldots, 4$, and $\theta_l = \{W_l, b_l\}$. ReLU is the most popular activation function that provides a constant derivative, resulting in fast learning and preventing vanishing gradients in deep neural networks [27]. The last fcl ($l = 5$) is the standard linear dense layer, defined as:

$$\hat{y} = W_l h_{l-1} + b_l, \tag{2}$$

the output of which ($\hat{y}$) is the estimated engagement level. The optimization of the network parameters is obtained by

minimizing the loss α_c defined as:

$$\Omega^* = \underset{\Omega=\{\theta_1,..,\ \theta_5\}}{\arg\min}\ \alpha_c(\hat{y}, y) = \underset{\Omega=\{\theta_1,..,\ \theta_5\}}{\arg\min}\ \frac{1}{N}\sum_{i=1}^{N}\|y_i - \hat{y}_i\|^2, \tag{3}$$

where N represents the number of training face images from either of the datasets (S^0, S^1) or both, depending on the evaluation setting (see Sec. IV).

B. Culturalized Model (CultureNet)

The GenNet deep model does not offer flexibility to specialize its network parameters to each target culture. However, this is important, as the data coming from different cultures may have different variance at both the feature and label level. The former is likely because of differences in facial physiognomy in children with different cultural backgrounds, while the latter is likely because of their differences in expression and in the distributions of their expert-labeled levels of engagement (see Fig. 2). To this end, we formulate the CultureNet, a deep learning approach that allows us to leverage the data from both cultures (S) while also being able to have a culture-specific model (see Fig. 1). More formally, we start from the GenNet model architecture, initialized using the fine-tuned ResNet weights (W^{rnet}) and the ReLu/Linear fcls parametrized by Ω. Then, the learning of the CultureNet is performed in two steps:

- **Joint Learning.** We jointly learn the weights of (five) fcls as done in GenNet, while keeping the ResNet weights (W^{rnet}) frozen. Here, the data from both cultures (S^0, S^1) are used to fine-tune the network weights by solving the optimization problem in Eq.3.
- **Culturalization.** We freeze the network parameters $\{\theta_1, ..,\ \theta_4\}$ tuned in the previous step to both cultures and use the culture-specific data to additionally fine-tune the last layer of the network (θ_5), i.e., the linear fcl used for engagement estimation, effectively rendering the culture-specific models with the shared network structure (see Fig.1).

The learning in the second step ("culturalization") is attained through the last layer in the network (fcl $l = 5$), one for each culture. Then, before further optimization, the culture-specific layers are initialized as: $\theta_5^0 \leftarrow \theta_5$ and $\theta_5^1 \leftarrow \theta_5$, and then fine-tuned using the data from C0 (S^0) and C1 (S^1), respectively, as:

$$(\theta_5^c)^* = \underset{\theta_5}{\arg\min}\ \frac{1}{N}\sum_{i=1}^{N\in S^c}\|y_i - \hat{y}_i\|^2,\ \ c = \{0, 1\}. \tag{4}$$

The final network weights are then used to perform the culture-specific inference of engagement from target images.

C. Implementation Details

We implemented the proposed deep architectures using Keras API [11] with the Tensorflow [1] back-end. The network receives as input the cropped face images (256×256), resulting in 2048D activations in the output of the ResNet layers, which were passed to the fcls with the following structure: fc1 (2048×500), fc2 (500×200), fc3 (200×100), fc4 (100×50), and fc5 (50×1). The parameter optimization was performed using the standard back-propagation algorithm and Adadelta optimizer with the default parameters [27]. We used this shrinking network structure to reduce the number of network activations while still being able to exploit the network depth and different abstract representations produced by its activations in the output of each layer. Overall, this structure performed similar or better than other more shallow or deeper structures that we evaluated[2]. To avoid over-fitting of the deep models, we tried several strategies including batch-normalization and dropout with a varying portion of dropping nodes [27]; however, this did not improve the model's performance significantly on the target data. Therefore, we used only the early-stopping strategy based on the validation data with the maximum number of epochs set to 20, which turned out to produce the best performance. The details of the employed validation settings are provided in Sec. IV. The engagement inference from new face images is then performed by passing the raw face image through the described network architecture (ResNet+5 fcls), using the feed-forward algorithm. The python code and unidentifiable facial features, obtained in the output of the fine-tuned ResNet and used to train/evaluate the CultureNet, are made publicly available at: https://github.com/yuriautsumi/CultureNet-Autism.

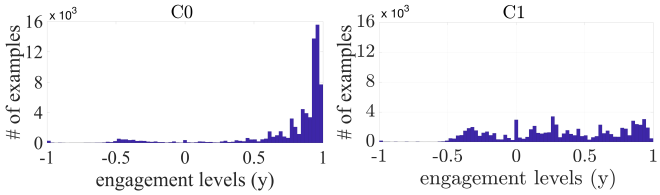

Fig. 2: Histograms of the children's engagement annotations, within the range [-1,1], for C0 (Japan) and C1(Serbia).

IV. Experiments

A. Experimental Setup

Data and Features. In our experiments, we used the cross-cultural dataset of children with ASC attending a single session (on average, 25 mins long) of a robot-assisted autism therapy [37]. During the therapy, an experienced educational therapist worked on teaching the children socio-emotional skills, focusing on recognition and imitation of behavioral expressions as shown by neurotypical population. To this end, the NAO robot was used to demonstrate examples of these expressions. It was also controlled by the therapist (the "Wizard-of-Oz" scenario), who used the NAO to keep the child engaged during the interaction. The data comprises audio-visual and autonomic physiological recordings of 17/18 children, ages 3-13, with Japanese (C0) / Serbian (C1) cultural background, respectively. All the children have

[2]The performance of the models was not affected significantly when using different network architectures, thus, we chose this one in order to have consistent settings across different experiments.

TABLE I: Overview of different evaluation settings.

	Evaluation setting	Data type	Model
M1	subject independent	within-culture	GenNet
M2	subject independent	cross-culture	GenNet
M3	subject independent	mixed-culture	GenNet
M4	subject independent	both-cultures	CultureNet
M5	subject dependent	within-culture	GenNet
M6	subject dependent	child-specific	GenNet
M7	subject dependent	both-cultures	CultureNet

a prior medical diagnosis of ASC, varying in its severity. In this paper, we used the video data of 15 children from each culture.[3]

The images of each child were processed using Faster-RCNN [35], an automatic face detection approach based on deep networks consisting of several convolutional layers and recurrent-neural networks and pre-trained for face detection[4]. In cases where the face was detected in the target frame, the output size was 256×256 pixels. To have a balanced dataset, we used 6k face images per child that were automatically detected (with a detection threshold > 0.95) using Faster-RCNN. Such cropped faces were then passed as input to the proposed deep-learning architectures (Sec. III).

To train the deep models, we used the annotations of children's engagement in target videos. The videos were coded on a continuous scale from -1 to $+1$ by five expert human coders, while watching the audio-visual recordings of the therapy sessions. The coders' agreement was measured using the intra-class correlation (ICC) [39], type (3,1). The ICC ranges from $0 - 100\%$ and is commonly used in behavioral sciences to assess the coders' agreement. The average ICC among the coders was $61 \pm 14\%$ (mean$\pm$SD). Their codings were temporally aligned to obtain the gold-standard labels [36], which were used to train the deep models. The histogram of the labels' distributions for both cultures is depicted in Fig. 2.

Evaluation Metrics. As evaluation measure, we used the following metrics: ICC (described above), concordance correlation coefficient (CCC), and Pearson Correlation (PC). We compare these three correlations scores as they allow us to quantify different aspects of the models: while ICC encodes the consistency between the model predictions ($\hat{y}$) and labels (y) as $y = \hat{y} + b$, CCC measures agreement as a departure from perfect linearity ($y = \hat{y}$), and PC captures the general linear relationship $y = a\hat{y} + b$, where a and b are the scale and bias terms, respectively. Note that CCC and PC range from [-1,1]; however, we report them in $\%$, i.e., [-100,100]. We also include Mean Absolute Error (MAE), encoding average absolute deviation from the labels ($|y - \hat{y}|$). We computed these scores on the pairs of the model estimates and gold-standard labels.

Evaluation Settings. To evaluate the different deep learning architectures, we randomly split the data of each child into non-overlapping partitions, containing $p_1 = 80\%$ (5k) and $p_2 = 20\%$ (1k) face images, and consider two evaluation settings: subject-independent (SI) and subject-dependent (SD). In the SI setting, a non-overlapping set of children was used to learn and test the models. From the children that were used for learning, we used p_1 and p_2 to train and validate the models, respectively. In the SD setting, however, a portion (p_2) of each child is used for learning the models. In both settings, the performance scores are computed on the held-out portion (p_1) of each testing child. We report the average results over 10 runs, each time starting from a different random initialization of the deep models/child data selection.

For the SI experiments, we evaluated the following settings: **M1** – within culture evaluation, where the training and test children from the same culture (C0 or C1) were used to evaluate the GenNet model. Specifically, 14 children were used to train the model and 1 child to test model. This was repeated for each child (thus, 15 times) from each culture. **M2** – cross-culture evaluation of GenNet, where the models were trained on C0 (15 children) and tested on C1 (15 children), and vice versa. **M3** – mixed-culture evaluation, where we used training children from both cultures (i.e., 29 children) to learn the GenNet, and tested on the left-out child. This, again, was repeated for all 30 children. Lastly, **M4** is the "culturalized" model (CultureNet), where the last-layer of the GenNet obtained by **M3** was further fine-tuned to each culture separately using data of 14 children from the target culture (see Fig.1). The testing was then performed in the leave-one-child-out fashion, as in **M3**.

For the SD experiments, we define **M5** as the within-culture GenNet model (**M1**); however, $p_2 = 20\%$ of data of each child (from culture C0 or C1) were used to train the model, and the remaining $p_1 = 80\%$ of the child-data for testing.[5] We also evaluated the child-specific models (**M6**), where only the $p_2 = 20\%$ of the target child data was used to train the model (thus, no data of the other children) using GenNet. This was repeated for each child from C0/C1. We also define model **M7** that performs personalization of the CultureNet model **M4** to each child. Specifically, we first trained a GenNet using p_2 data of all children from both cultures. We then fine-tuned its last layer to the target culture (C0 or C1) using p_2 data from each child from that culture (culturalization). Finally, this was followed by fine-tuning of the culture-specific layer using p_2 data of target child, effectively rendering 30 child-dependent models. Table I provides the summary of the proposed settings.

B. Results

Table II summarizes the average results computed for different deep models (**M1**-**M7**) and reported per culture, calculated as an average across the mean performance (over 10 runs) for each child. In the SI settings, it can be observed

[3]The videos of the remaining children were of a low quality and/or contained significant face occlusions.

[4]https://github.com/playerkk/face-py-faster-rcnn

[5]This strategy allows us to have direct comparisons with the SI models evaluated on the p_1 portion of target data. Also, as shown later in Fig.3, including more data of the other children did not improve the estimation performance.

TABLE II: Comparison of the proposed deep learning settings, showing the mean±standard deviation of 10 runs of subject-independent (**M1**-**M4**) and subject-dependent (**M5**-**M7**) cross-validation. We report Intra-class Correlation (ICC), Concordance Correlation Coefficient (CCC), and Pearson Correlation (PC), as well as Mean Absolute Error (MAE). The Mean Estimate (**ME**) is computed from the training portion (20%) of the engagement labels of each child. This constant child-specific estimate of his/her future engagement levels is compared against the test set as in the subject-dependent models.

Models	**C0**				**C1**			
	ICC [%]	CCC [%]	PC [%]	MAE	ICC [%]	CCC [%]	PC [%]	MAE
M1	**13.19±11.78**	10.63±8.72	**16.37±15.93**	**0.22±0.14**	**10.75±9.54**	**7.96±7.30**	**11.81±13.58**	**0.32±0.16**
M2	5.97±6.89	3.37±6.18	3.84±12.27	0.64±0.11	7.31±7.39	4.08±6.55	5.92±11.70	0.57±0.20
M3	10.56±10.24	8.29±8.01	12.61±12.69	0.28±0.10	9.40±8.14	6.48±6.47	10.04±10.44	0.37±0.12
M4	12.90±11.74	**10.92±9.50**	15.90±14.90	0.23±0.13	10.01±9.92	7.51±8.08	10.92±13.07	0.33±0.15
M5	38.91±25.90	36.71±26.28	41.48±26.37	0.15±0.07	38.68±20.99	36.33±21.17	41.24±21.32	0.18±0.07
M6	39.17±25.84	36.77±26.31	42.03±26.24	0.14±0.07	38.60±21.23	36.18±21.37	41.27±21.45	0.18±0.07
M7	**43.35±24.30**	**43.18±24.32**	**45.17±24.17**	**0.13±0.08**	**42.17±19.71**	**41.93±19.76**	**43.62±19.69**	**0.16±0.08**
ME	-	-	-	0.18±0.15	-	-	-	0.23±0.08

that **M1** (the within-culture training) performs the best overall. Compared to **M2**, which uses data from the other culture to train the deep model, we note a drop in the performance in both C0 and C1. This is expected as the label distribution in the two cultures are different, rendering models that are suboptimal for estimation of target engagement levels specific to each culture.

We also attribute this adverse performance to the difference in the facial physiognomy between the children from the two cultures, which may also bias the features (network activations) in the last layer of the ResNet. Interestingly, in this setting (**M2**), better results are obtained on C1 than on C0, which is again due to the label bias: in C0 mostly high levels of engagement were observed. This, evidently, led the model trained on C1 to predict low-level estimates of engagement in C0, as reflected in its high MAE. In terms of correlation scores, the relationship between the scores is consistent: PC>ICC>CCC, evidencing that the models' estimates are positively correlated with the ground-truth, but there are strong differences in scale and bias of the estimates. Overall, the MAE performance of the SI models is rather low, even below taht achieved by using the simple mean estimate (a constant engagement level) computed from data of each child (non-overlapping with test data) – see Table II .

On the other hand, we observe large improvements in the models' performance by the SD settings (i.e., when using 20% of the target child data). First, note that all three correlation measures produce similar performance, showing that these models attain more consistent estimates of engagement (ICC vs. PC), with an overall higher level of agreement (CCC vs. PC) with the labels. This is also reflected in their MAE. Interestingly, by comparing **M5** and **M6**, there is a small difference in their performance. This, in turn, indicates that the large heterogeneity in the individual data drives the models to focus mainly on the target child data during learning, while not benefiting from the data of the other children also used during the training (**M5**). In other words, while **M5** uses a simple addition of the target child data to its training set, **M7** exploits this data more effectively.

However, this seems to depend largely on the training strategy – by looking into the performance achieved by **M7**, which leverages the data of the other children and culture, we also see large improvements over the child-specific models **M6**. This is achieved through the proposed training strategy, designed to take advantage of (i) both cultures' data, followed by focusing on (ii) within culture data, and finally (iii) specializing to the target child (personalization). Thus, the network pre-training using steps (i)-(ii) plays an important role in learning robust initial network parameters, from which the child-adaptation can take full advantage. Note also the high error-bars, depicted in terms of one standard deviation (SD), of the performance of both settings (SI and SD). Namely, it turns out that for some children, the learned models simply fail due to the large variation in their individual data, which is especially pronounced in the SI experiments (see Fig. 3).

We next investigate the contribution of adding more data of target children for the model learning and generalization. Fig. 3 (*Left*) shows the performance of the **M1** (SI) model in terms of PC when varying the number of (i) training children, and (ii) portion of the data from the training children. Including more data during training did not seem to improve the generalization of the deep model. Instead, including more data sometimes even hurts the model's performance in the SI setting. We attribute this to the large heterogeneity in facial expressions features of target children: by adding more data, the model's variance increases and the model becomes even more uncertain when trying to generalize to the previously unseen children. Also, the facial features were extracted using ResNet, fine-tuned on the AffectNet data of faces of typical individuals, which could have also influenced the quality of the obtained mid-level feature representations. We also performed a similar analysis for the SD setting using the child-specific model **M6**. In this case, we found that including more data of target children improves largely the performance scores, with PC increasing from C0: $27.86\% \rightarrow 34.60\% \rightarrow 38.98 \rightarrow 42.03\%$, and C1: $25.14\% \rightarrow 31.97\% \rightarrow 36.98\% \rightarrow 41.27\%$, when $5\% \rightarrow 10\% \rightarrow 15\% \rightarrow 20\%$

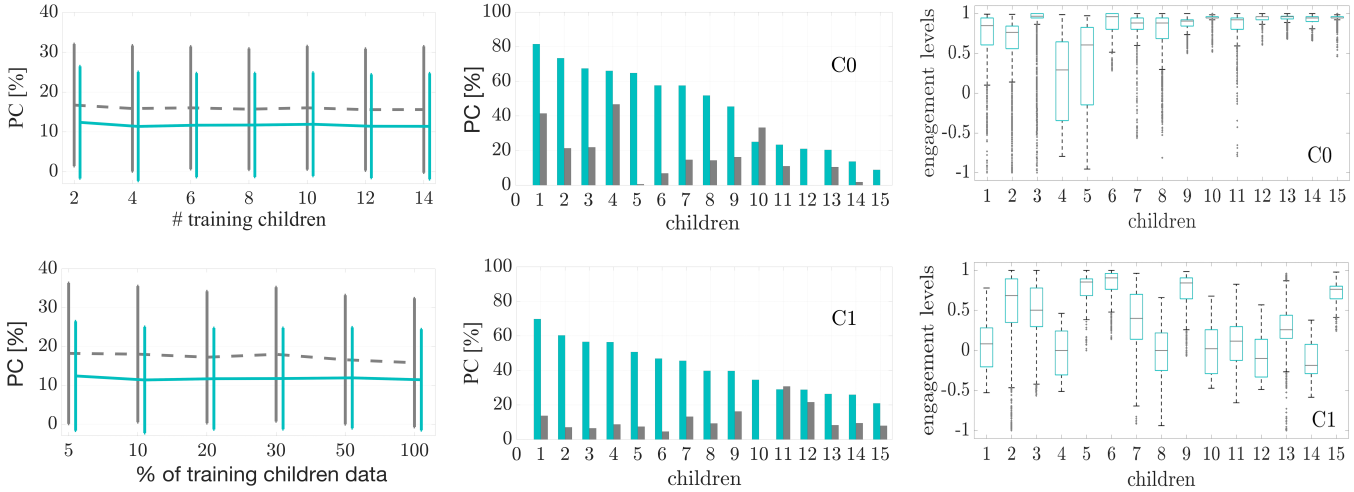

Fig. 3: *Left:* The performance of **M1** (the subject-independent within-culture model) w.r.t. the amount of training data: the number of children used to train the model (*top*), and the percentage of the data from each of the 14 children used to train the model (*bottom*). The reported PC scores are computed as average for the children from each culture, C0 and C1, depicted with dashed-gray and full-green lines, respectively. *Middle:* The performance per-child and culture by the "culturalized" models: **M4** (in gray) and **M7** (in green). The children are sorted based on the decreasing performance by **M7** in order to better assess the differences between the two models. *Right:* The boxplots indicating the median of enagement levels (provided by human coders) of the corresponding children. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively, and the whiskers extend to the most extreme data points.

of the target child data is used, respectively. While this is expected, here we show empirically that having access to the child data (even a small portion) is critical for achieving substantial improvements in the models' performance.

To obtain better insights into the models' performance at the individual level and across the cultures, Fig. 3 (*Middle*) shows the per-child PC attained by the culturalized/personalized models – **M4** and **M7** – evaluated under the SI and SD settings, respectively. We observe that in both cultures, personalizing the CultureNet by using 20% of the target child data during training allows **M7** to largely improve the estimation performance on almost all children. For instance, for child 5/C0, the SD model achieves PC above 70% compared to near 0% when tested in the SI manner. A similar trend is present for most children from C1, where **M7** outperforms **M4** by a large margin. Finally, note that even the model **M7** does not generalize with consistent performance across all the children/cultures. This, in turn, shows that for different children, more data may be needed to adapt the models in order to account for the within-child data variance. To illustrate this, in Fig. 3 (*Right*), we depict the distribution of the true enagement levels per child, where the children are sorted in the same order as in Fig. 3 (*Middle*). Note that in C0, most of the children are highly engaged; however, **M7** underperforms on children with low variation in their engagement levels (i.e., children 9-15). On the other hand, in culture C1, the engagement levels vary more largely between and within most of the children (note the longer whiskers in the plots of children from C1). This may explain the less steep drop in the **M7** perfromance on children from C1.

V. Discussion and Conclusions

In this work, we investigated different deep learning settings that we designed for automated estimation of engagement from face images of children with ASC, who participated in one-day robot-assisted autism therapy. Our results reveal important findings in terms of how deep learning can be used to leverage face-image data of children with ASC to attain better estimation of target engagement. More specifically, we analyzed the role of the cultural label as the driving factor in learning deep models for their generalization to different children with ASC: within and across two cultures (Asia vs. Europe). Furthermore, we performed this analysis in the child-independent and child-dependent manner. We showed that due to the large difference in the distribution of engagement levels in the two cultures, the deep models trained on only one culture have limited ability to generalize to the other culture. While we cannot say for sure the effect is due to any specifics in the underlying cultures (vs. due to some bias in our sample from the two cultures) it does appear to downgrade the estimation performance on these data. The difference in facial physiognomy between the two cultures and dynamics of their facial expressions, are also a potential cause. This indicates the importance of having access to data of the target culture and children when building deep models for engagement estimation.

Another important finding is that increasing the number of the children data to train the models did not lead to improvements in performance on previously unseen children. On the other hand, by including a relatively small portion of the target child data (1k examples), the models easily

adapt, largely improving the estimation performance. Note also that just adding more data of the target child during training will not necessarily increase the model's performance – the model's learning and inference has also to be designed carefully. We demonstrated this by personalizing the CultureNet to each child, which resulted in the best performance among the compared models. This poses a question to what extent the engagement estimation can be improved when working with face images of children with ASC? In our recent work [36], we achieved an ICC of 65% in the task of engagement estimation from multi-modal data, including face, body, voice and biosignals, of children with ASC. While this work achieved the ICC of 43% directly from face images (in contrast to [36] where facial landmarks and other high-level features were used), there is still a lot of room for improvement; however, more data and behavioural modalities (beyond raw face images) may be needed.

One of the technical limitations of the current approach is that the ResNet, used to extract mid-level feature representation in the proposed deep architectures, is fine-tuned on faces of typically developing individuals. In future, we plan to evaluate our approach using end-to-end learning of the deep models, hopefully rendering more discriminative facial features of children with ASC. Ideally, this approach would leverage data from multiple cultures and a larger number of children. This, however, raises concerns about data privacy and sharing of sensitive personal information such as face images. For this, secure data-sharing frameworks [10] can be explored to optimize the learning of deep models.

ACKNOWLEDGMENTS

This work was supported by Grant-in-Aid for Young Scientists B, grant no. 16K16106, Chubu University grant no. 27IS04I, and EU HORIZON 2020 grant nos. 701236 (EngageME), 751615 (BROS) and 688835 (DE-ENIGMA).

REFERENCES

[1] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *USENIX Symposium on OSDI*, 2016.
[2] L. B. Adamson *et al.*, "Early interests and joint engagement in typical development, autism, and down syndrome," *Journal of Autism and Developmental Disorders*, 2010.
[3] D. Almirall, C. Kasari, D. F. McCaffrey, and I. Nahum-Shani, "Developing optimized adaptive interventions in education," *Journal of Research on Educational Effectiveness*, 2018.
[4] S. M. Anzalone *et al.*, "Evaluating the engagement with social robots," *IJSR*, 2015.
[5] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub, 2013.
[6] A. Baird *et al.*, "Automatic classification of autistic child vocalisations: A novel database and results," *Interspeech*, 2017.
[7] S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a "theory of mind"?" *Cognition*, 1985.
[8] T. Belpaeme *et al.*, "Multimodal child-robot interaction: Building social bonds," *Journal of Human-Robot Interaction*, 2012.
[9] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
[10] E. Castelló Ferrer *et al.*, "Robochain: A secure data-sharing framework for human-robot interaction," *eTELEMED*, 2018.
[11] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: https://github.com/fchollet/keras
[12] A. Chorianopoulou *et al.*, "Engagement detection for children with autism spectrum disorder," in *IEEE ICASSP*, 2017.
[13] C. E. Clabaugh, "Interactive personalization for socially assistive robots," in *IEEE HRI*, 2017.
[14] M. B. Colton *et al.*, "Toward therapist-in-the-loop assistive robotics for children with autism and specific language impairment," *Autism*, 2009.
[15] D. Conti *et al.*, "A cross-cultural study of acceptance and use of robotics by future psychology practitioners." IEEE, 2015.
[16] T. C. Daley, "The need for cross-cultural research on the pervasive developmental disorders," *Transcultural Psychiatry*, 2002.
[17] H. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychological Bulletin*, 2002.
[18] P. G. Esteban *et al.*, "How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder," *Paladyn, Journal of Behavioral Robotics*, 2017.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016.
[20] A. S. Heinsfeld *et al.*, "Identification of autism spectrum disorder using deep learning and the abide dataset," *NeuroImage: Clinical*, 2018.
[21] J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, and R. W. Picard, "Using electrodermal activity to recognize ease of engagement in children during social interactions," in *Proc. of the ACM Int'l Joint Conf. on Pervasive and Ubiquitous Computing*, 2014.
[22] Y. Huang, E. Gilmartin, and N. Campbell, "Conversational engagement recognition using auditory and visual cues," in *Interspeech*, 2016.
[23] M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, 2015.
[24] D. Keen, "Engagement of children with autism in learning," *Australasian Journal of Special Education*, 2009.
[25] E. S. Kim, R. Paul, F. Shic, and B. Scassellati, "Bridging the research gap: Making HRI useful to individuals with autism," *Journal of Human-Robot Interaction*, 2012.
[26] Y. Kishida and C. Kemp, "The engagement and interaction of children with autism spectrum disorder in segregated and inclusive early childhood center-based settings," *Topics in Early Childhood Special Education*, 2009.
[27] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 2015.
[28] A. Libin and E. Libin, "Person-robot interactions from the robopsychologists' point of view: The robotic psychology and robotherapy approach," *IEEE*, 2004.
[29] K. B. Martin *et al.*, "Objective measurement of head movement differences in children with and without autism spectrum disorder," *Molecular Autism*, 2018.
[30] J. Matson *et al.*, "Examining cross-cultural differences in autism spectrum disorder: A multinational comparison from greece, italy, japan, poland, and the united states," *European Psychiatry*, 2017.
[31] R. Miotto *et al.*, "Deep patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, 2016.
[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE TAC*, 2017.
[33] S. L. Odom and P. J. Beckman, Eds., *Widening the Circle: Including Children with Disabilities in Preschool Programs*, ser. Early childhood education series. New York: Teachers College Press, 2002.
[34] P. Perepa, "Cultural basis of social 'deficits' in autism spectrum disorders," *European Journal of Special Needs Education*, 2014.
[35] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.
[36] 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.
[37] O. Rudovic, J. Lee, L. Mascarell-Maricic, B. Schuller, and R. W. Picard, "Measuring engagement in robot-assisted autism therapy: A cross-cultural study," *Frontiers in Robotics and AI*, 2017.
[38] Scherer KR and Wallbott HG, "Evidence for universality and cultural variation of differential emotion response patterning," *Journal of Pers. Soc. Psychol.*, 1994.
[39] P. E. Shrout and J. L. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, 1979.
[40] M. Tincani, J. Travers, and A. Boutot, "Race, culture, and autism spectrum disorder: Understanding the role of diversity in successful educational interventions," *Research and Practice for Persons with Severe Disabilities*, 2009.