Proceedings of the 27th IEEE International Symposium
on Robot and Human Interactive Communication,
Nanjing, China, August 27-31, 2018

Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder

Guangtao Nie, Zhi Zheng, Jazette Johnson, Amy R. Swanson, Amy S. Weitlauf, Zachary E. Warren and Nilanjan Sarkar

Abstract— **Autism Spectrum Disorders (ASD) are characterized by deficits in social communication skills, such as response to joint attention (RJA). Robotic systems have been designed and applied to help children with ASD improve their RJA skills. One of the most important goals of robot-assisted intervention is helping children generalize social interaction skills to interact with other people. Thus predicting children's human-human interaction (HHI) performance based on their human-robot interaction (HRI) process is an important task. However, to the best of our knowledge, little research exists exploring this topic. The Early Social-Communication Scales (ESCS) test is a measurement of nonverbal social skills, including RJA, for young children. We conducted two longitudinal user studies with a robot-mediated RJA system in young children with ASD, followed by HHI sessions consisting of ESCS administration. In this paper, we present findings regarding how to predict participants' RJA performance in HHI based on their head pose patterns in HRI, under a semi-supervised machine learning framework. As a three-class classification problem, we achieved a micro-averaged accuracy of 73.5%, which indicates the potential effectiveness of the proposed method.**

I. INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), 1 in 68 children in United States has Autism Spectrum Disorder (ASD) [1]. Deficits in social communication skills, such as joint attention skill, are among the primary symptoms of ASD [2]. Joint attention is defined as the shared focus of two individuals on the same object, whereas response to joint attention (RJA) refers to one's ability to follow the direction of another person's gaze and gestures [3].

Evidence suggests that early detection and intensive behavioral intervention are critical to improved outcomes for young children with ASD [4]. Nevertheless, it has also been shown that many families are not able to access evidence based intervention and standardized assessment [5]. There is strong evidence that technology-assisted interventions are promising for children with ASD due to their controllability, flexibility, replicability and cost effectiveness [6]. In particular, although underlying causal mechanisms are as of yet unclear, existing literature supports the specific use of robots for interventions with children with ASD, who many show preferences for humanoid robots and non-biological motion over other types of stimuli [25]. In a previous study, we developed a robot-mediated system to help young children with ASD learn RJA skills [7], which was well tolerated and elicited promising RJA performance [7, 8]. Subsequently we conducted another study where some human-robot interaction (HRI) sessions were followed by the Early Social-Communication Scales (ESCS) test, which is a videotaped structured observation measurement of nonverbal communication skills for young children, administered by a human expert [9], to assess human-human interaction (HHI) skills. In this paper, we present a method to predict HHI RJA performance of children with ASD based on their RJA skill observed during HRI.

During the interaction with the robotic system, the participant needed to turn his/her head to look at the joint attention (with the robot) target to share their focus [7]. Head pose is an important social cue to establish shared reference in both HRI [10, 11] and HHI [12]. Thus, we believe that the participants' head pose pattern in an HRI session can be used to predict their RJA performance in HHI. This idea is similar to the work of Liu et al. [13], in which gaze pattern in human-computer interaction was used to predict whether a participant had ASD or not. We apply a semi-supervised machine learning framework to investigate how well head pose in HRI can predict RJA performance in HHI. This framework allows us to utilize data from both HRI sessions with subsequent ESCS tests and those without ESCS tests. Under the semi-supervised machine learning framework, the results of RJA performance prediction in HHI using head pose in HRI is promising. We achieved 73.5% micro-averaged accuracy for a three-class classification problem. This work extends the functionality of our robotic system by providing a reference RJA performance level corresponding

G. Nie is with the Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN 37212 USA (e-mail: guangtao.nie@vanderbilt.edu).

Z. Zheng is with Biomedical Engineering Department, University of Wisconsin-Milwaukee, Milwaukee, WI 53211 USA (e-mail: zheng36@uwm.edu).

J. Johnson is with the Mechanical Engineering Department, Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN 37212 USA (e-mail: jazette.johnson@vanderbilt.edu).

A. R. Swanson is with the Treatment and Research Institute for Autism Spectrum Disorders (TRIAD), Vanderbilt University Medical Center, Nashville, TN 37212 USA (e-mail: amy.r.swanson@vanderbilt.edu).

A. S. Weitlauf is with the Treatment and Research Institute for Autism Spectrum Disorder (TRIAD), Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN 37212 USA (e-mail: amy.s.weitlauf@vanderbilt.edu).

Z. E. Warren is with the Treatment and Research Institute for Autism Spectrum Disorders (TRIAD), Departments of Pediatrics and Psychiatry, Vanderbilt University Medical Center, Nashville, TN 37212 USA (e-mail: zachary.e.warren@vanderbilt.edu).

N. Sarkar is with the Mechanical Engineering Department, Electrical Engineering and Computer Science Department, Vanderbilt University, Nashville, TN 37212 USA (e-mail: nilanjan.sarkar@vanderbilt.edu).

978-1-5386-7980-7/18/$31.00 ©2018 IEEE

to the ESCS test, which is the standard and universal measurement but not easily accessible for most young children with ASD.

Although other robotic systems have been designed for children with ASD to ameliorate joint attention skills [14], or induce the joint attention behaviors [15], none has reported any work on predicting HHI performance based on HRI process.

The remainder of this paper is organized as follows. In Section II, we describe the system, task design and the protocols for data collection. In Section III, feature generation and fusion methods for machine learning are introduced. In Section IV, the semi-supervised machine learning framework is introduced. Results are presented and analyzed in Section V followed by a brief discussion of the current work and future work in Section VI.

II. Problem Background

A. System and Task Design

The experimental environment and setup of the robot-mediated RJA system [7] are shown in Fig. 1 and Fig. 2, respectively. Four web cameras, Cam 1 to Cam 4, composed the gaze tracking module, encircling the center of the participant chair with a radius of 90 cm. This web camera array detected the real-time head pose of the participants, using the supervised gradient descent algorithm [16]. Monitor 1 and Monitor 2 were used to display both targets and reward video for successful task completion. A NAO robot stood in front of the participant chair, performing deictic gestures to initiate the joint attention task. The system environment was symmetric about the line connecting the center of the participant chair and the NAO robot. Detailed information about system setup and the environment can be found in [7].

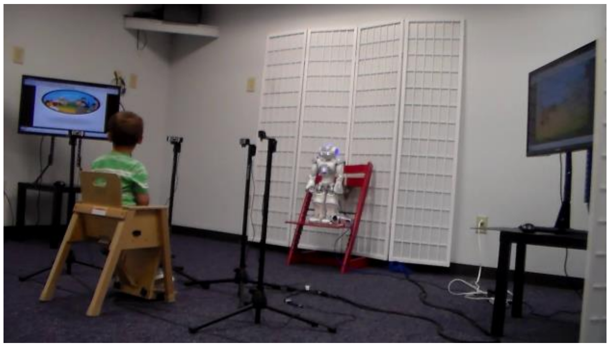


Fig. 1 Experiment video screenshot (back view)

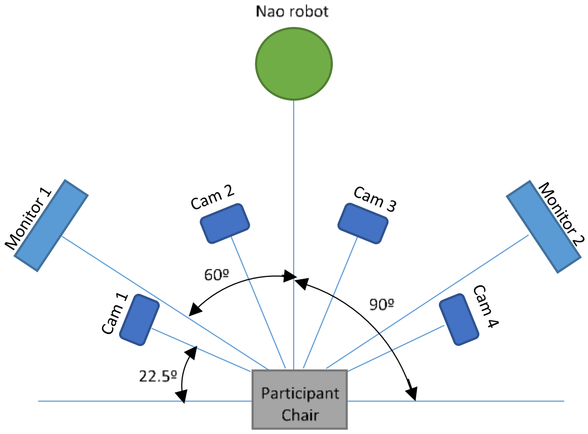


Fig.2 System setup (top view)

An HRI session in this paper is outlined in Fig. 3: There were 8 trials in an HRI session. Throughout each trial, a target was presented on either the left monitor (left target trials) or the right monitor (right target trials). Four right target trials and four left target trials were arranged in a random sequence. The prompt hierarchy for each trial was the same, as shown in Fig. 3. When the participant looked at the target, the current trial finished and the next trial started.

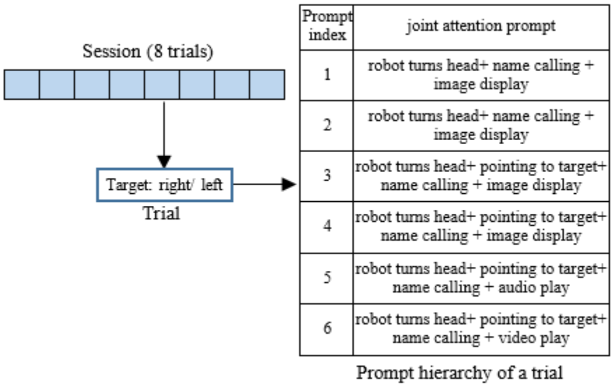
B. Data Collection

We conducted two longitudinal studies using this system: The first study included 14 participants [7] and the second study included 20 participants. All participants were young children with ASD. The age range of these 34 participants at recruitment was 1.65 to 4.57 years old (M=2.66, SD=0.54). In the first study, each participant finished 4 HRI sessions. In the second study, 20 participants were equally and randomly assigned into two groups (immediate group and control group), with the experiment timeline shown in Fig. 4. A0, A1, A2 were assessment visits while S1, S2, S3 and S4 were intervention visits. Each assessment visit contained an HRI session and a following ESCS test. In contrast, intervention visits contained only the HRI session.

We completed 56 and 130 HRI sessions in the first and second studies, respectively. 50 of the 130 HRI sessions in the second study were followed by an ESCS test. The RJA part of all ESCS test videos were coded by a human expert. As 16 ESCS test videos could not be coded, there were ultimately 34 HRI sessions with ESCS scores and 152 HRI sessions without ESCS scores available in total.

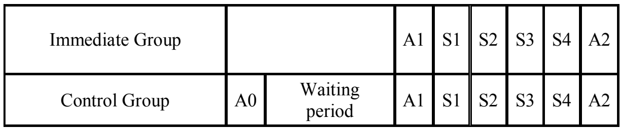
During each HRI session, the real-time head pose, denoted as a 3D vector $[yaw, pitch, roll]$, was calculated in the participant's coordinate system to indicate the participant's attention. The sampling frequency of head pose detection was 15 fps on average [7].

There are four RJA tasks in an ESCS test with a picture



Prompt index	joint attention prompt
1	robot turns head+ name calling + image display
2	robot turns head+ name calling + image display
3	robot turns head+ pointing to target+ name calling + image display
4	robot turns head+ pointing to target+ name calling + image display
5	robot turns head+ pointing to target+ name calling + audio play
6	robot turns head+ pointing to target+ name calling + video play

Fig. 3 HRI session structure



Immediate Group			A1	S1	S2	S3	S4	A2
Control Group	A0	Waiting period	A1	S1	S2	S3	S4	A2

Fig. 4 HRI session timeline for the second study

978-1-5386-7980-7/18/$31.00 ©2018 IEEE

(i.e., the target) placed on both the left and right side of the participant. The ESCS test was videotaped and participant skills was coded offline by a human expert. RJA skills were scored based on how many times a participant could successfully look at the target and are recorded as an integer number in the range [0, 4].

III. Features generation

To model head pose distribution in an HRI session, three feature generation methods were adopted here. Inspired by the area-of-interest (AOI) method [17], which is usually used to analyze gaze distribution on different semantic areas, we first adopted a three component Gaussian Mixture Model (GMM) as a feature generation method to model the head pose distribution based on three semantic components in the system, which were the left and the right monitors and the robot. Subsequently, to model the head pose distribution in a finer way, hard histogram and soft histogram feature generation methods were adopted, which are inspired by the Bag-of-Words (BoW) representation [18], a method popularly used to analyze content of documents and pictures.

Hard histogram and soft histogram methods were also applied to model head pose motion distribution.

A. Gaussian Mixture Model (GMM) of Head pose

Inspired by the AOI method to count the gaze distribution on different manually labeled semantic areas of interest, a GMM model was adopted to explicitly model the distribution of head pose on different areas of interest in this system. GMM is believed to properly model the head pose data falling on the border of two AOIs by softly classifying these head pose data points.

In order to interact with this robot-mediated RJA system, the participant needed to pay attention to three critical components: left monitor, right monitor and the NAO robot. These three components were the three AOI in this system. Therefore, a three-component GMM was adopted to model the distribution of head pose vectors ($[yaw, pitch, roll]$) in an HRI session. The parameters of GMM were estimated using EM algorithm, with the centers of the three critical components (in the participant's coordinate system, in the form of $[yaw_i, pitch_i, roll_i], i = 0,1,2$) initialized as the three mean vectors of the GMM.

After fitting the head pose vectors of an HRI session into the GMM, a general 3 by 3 covariance matrix Σ_i was calculated for each component. Due to the symmetry of matrix Σ_i, 6 of its entries are independent. The size of the mean vector μ_i is 1 by 3. Along with the mixing coefficient π_i, one component has 10 independent parameters, and thus, the three-component GMM has 30 independent parameters. Therefore, fitting the head pose vectors of an HRI session into the three-component GMM leads to a 30-dimensional feature vector.

The way GMM models the distribution of head pose in an HRI session is meaningful but coarse. From Fig. 1, we can see that there were a camera array, desks, cables and a divider in the experimental environment, which may distract the attention of these young children with ASD. Their possible interest on these sub-areas of interest (sub-AOI) was neglected by the GMM. Therefore, to model the head pose distribution in a finer way, hard histogram and soft histogram methods were adopted.

B. Hard Histogram of Head Pose

The calculation of the hard histogram of the head pose data of an HRI session can be divided into two steps, according to the training and predicting steps of a K-means classifier. First, a K-means clustering algorithm was adopted to find several cluster centroids on all of the head pose vectors in 186 sessions (training step). Second, this K-means classifier was utilized to classify the head pose vectors of each HRI session into a hard histogram, with the number of bins the same as the number of clusters in this K-means classifier. The value of each bin was the normalized number of head pose vectors classified into the corresponding cluster (predicting step). For both hard and soft histogram, K was set as 8, 16, 32, 48, 64, 80, 96, 112, 128, 144 and 160.

C. Soft Histogram of Head Pose

Compared to hard histogram method, soft histogram is believed to benefit the classifying of head pose vectors falling on the borders of clusters found by the K-means classifier [13]. The calculation of soft histogram can also be divided into two steps. The first step was the same as that of the hard histogram. The second step predicted a head pose vector's membership of every cluster in a soft way. First, we used K-means algorithm to find cluster centroids. Second, for each head pose vector in an HRI session, the inverse of squared Euclidean distance between this vector and each cluster centroid was calculated and normalized as a membership vector, as shown in Equation (1).

$$MV_{n,i}^{k} = \frac{\frac{1}{\left\| H_{n,i} - C_k \right\|_2^2}}{\sum_k \frac{1}{\left\| H_{n,i} - C_k \right\|_2^2}}, \tag{1}$$

where $MV_{n,i}^{k}$ represents the k-th dimension of the membership vector of head pose vector i in session n, where $n = 1,2,\ldots,186. i = 1,2,\ldots,M. k = 1,2,\ldots,K.$ M is the number of head pose vectors in an HRI session and K is the number of clusters we set for K-means classifier. $H_{n,i}$ is the head pose vector $[yaw, pitch, roll]$ and C_k is the k-th cluster centroid, in the same size as $H_{n,i}$.

Then, the membership vectors calculated for all head pose vectors in an HRI session were added and normalized to get a soft histogram for this session, as shown in Equation (2).

$$SH_n = \frac{\sum_i MV_{n,i}}{M}, \tag{2}$$

where SH_n represents the soft histogram of session n and $MV_{n,i} = \left[MV_{n,i}^1, MV_{n,i}^2, \ldots, MV_{n,i}^K \right]$.

D. Features of Head Pose Motion

Head pose motion was calculated as:

$$\Delta H_{n,i} = H_{n,i+1} - H_{n,i} \tag{3}$$

We adopted the aforementioned hard histogram and soft histogram to model the distribution of the head pose motion in each HRI session.

IV. Machine learning framework

For all of the 186 sessions, 34 of them were successfully labeled with ESCS score and the other 152 sessions left unlabeled. In this machine learning framework, each HRI session was viewed as a data point: a session with ESCS score was a labeled data point and a session without ESCS score was an unlabeled data point. To take advantage of both 34 labeled and 152 unlabeled data points, we adopted Transductive Support Vector Machine (TSVM) [19] with radial basis function (RBF) kernel to test the effectiveness of the features generated. In TSVM, unlabeled data points work to provide density information which helps the classifier set the decision boundary in low density area to improve the classification performance [20]. In this work, each input data point of the TSVM included a feature vector consisting of the 30 independent parameters of GMM and/or the K dimensional hard/soft histogram. The ESCS performance levels were used as labels. For this work, we extended the binary TSVM implementation with QN-S3VM [21] using one-vs-one scheme.

A. Label Thresholding

There were 34 labeled data points with an integer score in the range [0, 4]. The distribution of the original score is shown in Table 1.

Table 1. Distribution of ESCS score

Score	Number of session
0	14
1	7
2	4
3	2
4	7

Here we set the following thresholds to classify these scores into three performance levels:

$$\text{Label=}\begin{cases} \text{low if score=0} \\ \text{medium if score=1, 2} \\ \text{high if score=3, 4} \end{cases}$$

According to this rule, 14, 11 and 9 sessions were labeled as low, medium and high performance levels, respectively. Two rationales of setting such thresholds included: 1), by turning a 5-class classification problem into a 3-class classification problem, the distribution of data points in each class was more balanced, which could benefit the training of the TSVM; and 2), compared to binary labeling, the 3-class setting provided a finer resolution for the psychological analysis in the future.

B. Feature Preprocessing

We did normalization and feature selection for every input vector. First, each dimension of input vector was normalized to [0, 1] by

$$f = \left(f - f_{min}\right) / \left(f_{max} - f_{min}\right) \tag{4}$$

Then, for each input vector with more than 8 dimensions, we set a variance threshold to eliminate dimensions with variances lower than this threshold [22].

C. Feature Fusion

Besides each single feature generated, we also fused some of these features, hoping to improve the performance of machine learning. Four pairs of features were concatenated to do feature fusion:

1. GMM of head pose + hard histogram of head pose motion
2. GMM of head pose + soft histogram of head pose motion
3. Hard histogram of head pose + hard histogram of head pose motion
4. Soft histogram of head pose + soft histogram of head pose motion

D. Evaluation methods and metrics

We ran 10-round 5-folds cross validations (CV) to test the effectiveness of every generated feature and fused feature. In each round, 34 labeled data were randomly shuffled and stratified into 5 folds. In the CV of TSVM, each training set consisted of not only 4 training folds, but also the 152 unlabeled data. In each round both micro- and macro- averaged metrics were measured. All results reported in the next section were the arithmetic means and standard deviations of micro- and macro-averaged metrics on the 10-round CV. The parameters of TSVM were empirically set.

Accuracy and F1 score were the metrics we used to report the performance of our machine learning framework.

V. Results

Since the distribution of labeled data was only slightly unbalanced, the micro-averaged metrics were quite close to macro-averaged metrics. Due to space limitations, only the best micro-averaged metrics were selected and reported in Tables 2-6. In Table 7, both micro-averaged and macro-averaged metrics are reported. $F1_0$, $F1_1$ and $F1_2$ are the F1 scores of low, medium and high performance levels, respectively. For each table, the best evaluation metrics are marked bold and the last row is reported to show the effectiveness of the feature selection procedure.

A. Features of Head Pose

The best micro-averaged classification accuracy and F1 scores of GMM of head pose are reported in Table 2. Table 3 and Table 4 display the best micro-averaged classification

Table 2 Metrics mean (std. dev) of GMM of head pose

OIVD[1]	VT[2]	DaFS[3]	Micro-averaged			
			accuracy	$F1_0$	$F1_1$	$F1_2$
30	0.018	17	**0.560(0.038)**	0.623(0.029)	0.474(0.111)	**0.535(0.084)**
30	0.020	16	0.550(0.023)	**0.625(0.031)**	0.492(0.06)	0.482(0.05)
30	0.007	27	0.556(0.028)	0.613(0.032)	**0.523(0.047)**	0.494(0.075)
30	0	30	0.464(0.037)	0.557(0.036)	0.415(0.093)	0.355(0.067)

1: OIVD—Original Input Vector Dimension
2: VT—Variance Threshold (of Feature selection)
3: DaFS—Dimension after Feature Selection
These abbreviations will also be used in the following tables

accuracy and F1 scores using the hard and soft histogram of the head pose, respectively.

Table 3 shows that, in general, the hard histogram of head pose performs the best when K=64. From Table 4, we can observe that K=32 was the optimal choice when using the soft histogram of head pose. The hard histogram was better than soft histogram when modeling the distribution of head pose.

Table 3 Metrics mean (std. dev) of hard histogram of head pose

OIVD	VT	DaFS	Micro-averaged			
			accuracy	$F1_0$	$F1_1$	$F1_2$
64	0.014	49	**0.626(0.098)**	0.657(0.071)	0.579 (0.082)	0.646(0.1)
112	0.017	51	0.615(0.046)	**0.685(0.058)**	0.526(0.072)	0.598(0.067)
32	0.019	23	0.574(0.058)	0.606(0.06)	**0.652(0.089)**	0.363(0.073)
64	0.017	42	0.621(0.05)	0.629(0.054)	0.576(0.089)	**0.666(0.104)**
64	0	64	0.485(0.059)	0.580(0.044)	0.420(0.095)	0.400(0.124)

Table 4 Metrics mean (std. dev) of soft histogram of head pose

OIVD	VT	DaFS	Micro-averaged			
			accuracy	$F1_0$	$F1_1$	$F1_2$
32	0.014	31	**0.603(0.038)**	0.550(0.057)	0.740 (0.073)	0.499(0.074)
8	NA	8	0.579 (0.023)	**0.683(0.029)**	0.560(0.073)	0.401(0.052)
32	0.019	28	0.600 (0.035)	0.512(0.063)	**0.779(0.036)**	0.481(0.091)
48	0.019	30	0.553(0.075)	0.501(0.091)	0.628(0.1)	**0.539(0.121)**
32	0	32	0.571(0.035)	0.545(0.040)	0.670(0.062)	0.458(0.074)

Table 5 Metrics mean (std. dev) of hard histogram of head pose motion

OIVD	VT	DaFS	Micro-averaged			
			Accuracy	$F1_0$	$F1_1$	$F1_2$
48	0.017	41	**0.703(0.031)**	**0.703(0.045)**	0.699 (0.057)	**0.716(0.01)**
48	0.014	42	0.674(0.031)	0.679(0.057)	**0.738(0.055)**	0.600(0.065)
48	0	48	0.618(0.039)	0.648(0.064)	0.632(0.072)	0.568(0.064)

Table 6 Metrics mean (std. dev) of soft histogram of head pose motion

OIVD	VT	DaFS	Micro-averaged			
			accuracy	$F1_0$	$F1_1$	$F1_2$
48	0.014	42	**0.647(0.037)**	**0.632(0.055)**	0.718 (0.036)	0.579(0.086)
112	0.019	93	0.591 (0.031)	0.539(0.044)	**0.782(0.062)**	0.397(0.093)
128	0.019	103	0.635(0.027)	0.590(0.069)	0.722(0.052)	**0.603(0.074)**
48	0	48	0.553(0.032)	0.542(0.048)	0.603(0.077)	0.517(0.059)

B. *Features of Head Pose Motion*

The classification results of hard and soft histogram of head pose motion are reported in Table 5 and Table 6. The optimal K for hard and soft histogram of head pose motion were both 48, in general. Both hard and soft histogram of head pose motion worked better than that of head pose, which may indicate that head pose motion may contain more discriminative features than head pose itself, at least for young children with ASD in this study.

C. *Feature Fusion*

The four different feature fusion methods were validated and reported in Table 7. The aforementioned results (Table 2 through 6) show that a feature with the best micro-averaged accuracy generally performed better than one with the best F1 score. Therefore, for each feature fusion method, the result with best micro-averaged accuracy is selected and reported while macro-averaged metrics are reported as reference.

We can see that by concatenating a 32 dimensional hard histogram and a 48 dimensional hard histogram, the best micro-averaged classification accuracy could be achieved was 73.5%. Considering the fact that it is a three-class classification problem, this is a promising result.

Table 7 shows that the micro-averaged metrics are close

Table 7 Metrics mean (std. dev) of feature fusion

Feature fusion method	OIVD	VT	DaFS	Micro- averaged				Macro-averaged			
				accuracy	$F1_0$	$F1_1$	$F1_2$	accuracy	$F1_0$	$F1_1$	$F1_2$
GMM of head pose + hard histogram of head pose motion	30+128	0.04	20	0.712(0.037)	0.782(0.055)	0.677(0.073)	0.654(0.107)	0.710(0.031)	0.806(0.06)	0.695(0.058)	0.674(0.111)
GMM of head pose + soft histogram of head pose motion	30+128	0.032	65	0.688(0.044)	0.706(0.056)	**0.737(0.055)**	0.611(0.091)	0.688(0.041)	0.725(0.066)	**0.770(0.063)**	0.628(0.107)
Hard histogram of head pose + hard histogram of head pose motion	32+48	0.030	35	**0.735(0.053)**	**0.807(0.055)**	0.710(0.069)	0.660(0.082)	**0.734(0.059)**	**0.823(0.057)**	0.718(0.08)	0.696(0.1)
Soft histogram of head pose + soft histogram of head pose motion	48+48	0.038	18	0.694(0.044)	0.749(0.075)	0.564(0.082)	**0.756(0.063)**	0.697(0.047)	0.762(0.083)	0.576(0.083)	**0.799(0.06)**
Hard histogram of head pose + hard histogram of head pose motion	32+48	0	80	0.579(0.042)	0.800(0.055)	0.405(0.071)	0.442(0.090)	0.569(0.048)	0.800(0.070)	0.398(0.083)	0.440(0.114)

to the corresponding macro-averaged metrics. This could be a consequence of the label thresholding. According to these thresholds, although the data set was not completely balanced, it was not skewed severely either. Also, this shows that in Table 2-6, using only micro-averaged metrics is reasonable.

VI. Conclusion

In this paper we have presented our work on how to predict HHI (ESCS) RJA performance level based on the classification of participants' head pose patterns in HRI RJA process using TSVM. The results were promising, with the best micro-averaged accuracy of 73.5%, for a three-class classification problem. We believe this work can extend the functionality of our system by predicting the ESCS RJA performance based on the HRI session.

Comparing our results with those presented in [13], we find out that hard histogram worked in general better than soft histogram. We believe this is a practical difference which depends on the specific problem and its data distribution.

The main limitation of this work comes from the small sample size. The 186 HRI sessions from two longitudinal user studies were used as 186 data points in the machine learning framework. We believe that in the future, with more completed HRI sessions, either with or without ESCS score, the performance of this machine learning framework can be improved further, as more data can be used for training the TSVM.

There are different implementations of TSVMs besides QN-S3VM [21] that we used in this work, such as SVMlight [23], VS^{3}VM [24] and semi-supervised classification by low density separation [20]. As different implementations influence the practical performance of TSVM, these implementations may also be explored in the future.

Despite these limitations, we believe this is the first work to our knowledge that seeks to predict human-human social communication performance of children with ASD based on their skill observed in HRI. We believe that such an approach will lead to the assessment and generalization of HRI based intervention of children with ASD.

Acknowledgement

We gratefully acknowledge the support provided by the National Institutes of Health grants 1R33MH103518-01 and 1R21MH111567 that partially supported this research.

References

[1] Autism Spectrum Disorders Prevalence Rate, Autism Speaks and Center for Disease Control (CDC), 2011.

[2] C. Lord, E. H. Cook, B. L. Leventhal, and D. G. Amaral, "Autism spectrum disorders," Autism: The Science of Mental Health, vol. 28, p. 217, 2013.

[3] C. Moore and P. Dunham, Joint attention: Its origins and role in development: Psychology Press, 2014.

[4] Z. Warren, W. L. Stone, "Best practices: Early diagnosis and psychological assessment." Autism spectrum disorders, 2011: p. 1271-1282.

[5] Z. Warren, A. Vehorn, E. Dohrmann, C. Newsom, and J. L. Taylor, "Brief report: Service implementation and maternal distress surrounding evaluation recommendations for young children diagnosed with autism," Autism : the international journal of research and practice 17.6 (2013): 693–700. PMC. Web. 21 Feb. 2018.

[6] Z. Zheng, Q. Fu, H. Zhao, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Design of a Computer-Assisted System for Teaching Attentional Skills to Toddlers with ASD," Cham, 2015, pp. 721-730.

[7] 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," in IEEE Transactions on Human-Machine Systems, vol. PP, no. 99, pp. 1-11.

[8] Z. Zheng, G. Nie, A. Swanson, A. Weitlauf, Z. Warren, and N. Sarkar, "Longitudinal Impact of Autonomous Robot-Mediated Joint Attention Intervention for Young Children with ASD," Cham, 2016, pp. 581-590.

[9] P. Mundy, C. Delgado, J. Block, M. Venezia, A. Hogan, and J. Seibert, "Early social communication scales (ESCS)," 2003.

[10] C. Breazeal, "Social interactions in HRI: the robot view," IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 34, pp. 181-186, 2004.

[11] A. Gaschler, S. Jentzsch, M. Giuliani, K. Huth, J. de Ruiter, and A. Knoll, "Social behavior recognition using body posture and head pose for human-robot interaction," in Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on, 2012, pp. 2128-2133.

[12] M. Argyle, Social interaction vol. 103: Transaction Publishers, 1973.

[13] W. Liu, X. Yu, B. Raj, L. Yi, X. Zou, and M. Li, "Efficient autism spectrum disorder prediction with eye movement: A machine learning framework," in Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, 2015, pp. 649-655.

[14] R. S. De Silva, K. Tadano, M. Higashi, A. Saito, and S. G. Lambacher, "Therapeutic-assisted robot for children with autism," in Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on, 2009, pp. 3561-3567.

[15] S. M. Anzalone, E. Tilmont, S. Boucenna, J. Xavier, A.-L. Jouen, N. Bodeau, K. Maharatna, M. Chetouani, D. Cohen, and M. S. Group, "How children with autism spectrum disorder behave and explore the 4-dimensional (spatial 3D+ time) environment during a joint attention induction task with a robot," Research in Autism Spectrum Disorders, vol. 8, pp. 814-826, 2014.

[16] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, 2013, pp. 532-539.

[17] D. C. Yates and W. C. McCoy, "Device for seeking an area of interest within a body," ed: Google Patents, 1993.

[18] Z. S. Harris, "Distributional structure," Word, vol. 10, pp. 146-162, 1954.

[19] V. N. Vapnik and A. Sterin, "On Structural Risk Minimization or Overall Risk in a Problem of Pattern Recognition," Automation and Remote Control, Vol. 10, No. 3. (1977), pp. 1495-1503.

[20] O. Chapelle and A. Zien, "Semi-Supervised Classification by Low Density Separation," in AISTATS, 2005, pp. 57-64.

[21] F. Gieseke, A. Airola, T. Pahikkala, and O. Kramer, "Sparse Quasi-Newton Optimization for Semi-supervised Support Vector Machines," in ICPRAM (1), 2012, pp. 45-54.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," Journal of machine learning research, vol. 12, pp. 2825-2830, 2011.

[23] T. Joachims, "Transductive inference for text classification using support vector machines," in ICML, 1999, pp. 200-209.

[24] G. Fung and O. L. Mangasarian, "Semi-superyised support vector machines for unlabeled data classification," Optimization methods and software, vol. 15, pp. 29-44, 2001.

[25] J. J. Diehl, L. M. Schmitt, M. Villano, and C. R. Crowell, "The clinical use of robots for individuals with autism spectrum disorders: A critical review," Research in Autism Spectrum Disorders, vol. 6, pp. 249-262, 2012.