Robot-mediated Mixed Gesture Imitation Skill Training for Young Children with ASD

Zhi Zheng[1], *Student Member, IEEE*, Eric M. Young[2], Amy Swanson[3], Amy Weitlauf[3,4], Zachary Warren[3,4], Nilanjan Sarkar[1,2], *Senior Member, IEEE*
[1]Department of Electrical Engineering and Computer Science, [2]Department of Mechanical Engineering, [3]Vanderbilt Kennedy Center, Treatment and Research Institute for Autism Spectrum Disorders (TRIAD), [4]Department of Pediatrics
Vanderbilt University
Nashville, TN, USA
name@vanderbilt.edu

***Abstract*— Autism Spectrum Disorder (ASD) impacts 1 in 68 children in the U.S. with tremendous consequent cost in terms of care and treatment. Evidence suggests that early intervention is critical for optimal treatment results. Robots have been shown to have great potential to attract attention of children with ASD and can facilitate early interventions on core deficits. In this paper, we propose a robotic platform that mediates imitation skill training for young children with ASD. Imitation skills are considered to be one of the most important skill deficits in children with ASD, which has a profound impact on social communication. While a few previous works have provided methods for single gesture imitation training, the current paper extends the training to incorporate mixed gestures consisting of multiple single gestures during intervention. A preliminary user study showed that the proposed robotic system was able to stimulate mixed gesture imitation in young children with ASD with promising gesture recognition accuracy.**

Keywords— robot assisted intervention; autism spectrum disorder; gesture recognition; imitation

I. INTRODUCTION

Autism Spectrum Disorder (ASD) impacts 1 in 68 children in the U.S. and is associated with enormous costs [1]. This disorder is characterized by impairments in social communication and abnormal repetitive behaviors [2]. Imitation is a fundamental skill for neural development and social interaction [3]. However, the lack of ability to imitate is among the core vulnerabilities that are associated with ASD [3, 4]. Research indicates that early intervention focusing on core deficits lead to optimal treatment [5]. In this context, robot-mediated technologies have been shown to be promising as potential intervention tools for children with ASD due to the fact that many children with ASD prefer interacting with non-biological objects rather than with a real person [6, 7]. Therefore, this study focused on developing robotic technology to help young children with ASD learn imitation skills.

Several studies have shown that teaching imitation skills to children with ASD through the use of robotic technologies is feasible and has great potential [8, 9]. Dautenhahn et al. developed a humanoid robot KASPAR, which was able to interact with children with ASD using imitation games [10]. Fujimoto et al. designed techniques using wearable sensors for mimicking and evaluating human motion in real time to improve imitation skills of children with ASD [11]. Greczek et al. proposed a graded cuing mechanism to encourage the imitation behavior of children with ASD in a closed-loop "copy-cat" game [9]. Zheng et al. created a robotic system that provided imitation training for children with ASD with online feedback regarding the quality of gesture accomplished [7]. While the above-mentioned studies were important in establishing the feasibility and usefulness of robot-mediated systems for imitation skills training, they focus on simple, single-gesture based imitation skills. In reality, a child is expected to learn more complex gestures that can be combinations of a set of simple gestures. In this paper, we present a framework for robot-mediated imitation skill training for complex mixed gestures. Moreover, this current work utilizes a non-invasive setup that did not require the children to wear any physical sensors, since many children with ASD tend to reject body-attached hardware[12].

The current paper presents a new gesture recognition method capable of detecting mixed gestures, which are defined as simultaneous execution of multiple simple gestures from a participant, as well as identifying ("spotting") the start and the end of each detected gesture. A new intervention protocol was designed to test this algorithm and a preliminary user study with children with ASD and their typically developing (TD) peers was conducted to show the feasibility and potential of this robotic system.

The rest of this article is organized as follows. Section II describes the development of the robot-mediated intervention system. Section III features the experimental setup. Section IV presents the experimental results, followed by the authors' conclusions of this work in Section V.

II. SYSTEM DEVELOPMENT

A. System Architecture

The robot-mediated imitation skill training architecture (RISTA) consists of a robot module and a gesture tracking module that were operated based on the commands sent from a supervisory controller. The robot module utilized the humanoid robot NAO [13]. NAO is about 58cm high and has a childlike appearance. It is built with 25 degrees of freedom, flashing LED "eyes", speakers, multiple sensors, and a well maintained

This study was supported in part by the following grants: a grant from the Vanderbilt Kennedy Center (Hobbs Grant), a Vanderbilt University Innovation and Discovery in Engineering and Science (IDEAS) grant, National Science Foundation Grant 1264462, and the National Institute of Health Grants 1R01MH091102-01A1and 5R21MH103518-02. Work also includes core support from NICHD (P30HD15052) and NCATS (UL1TR000445-06).

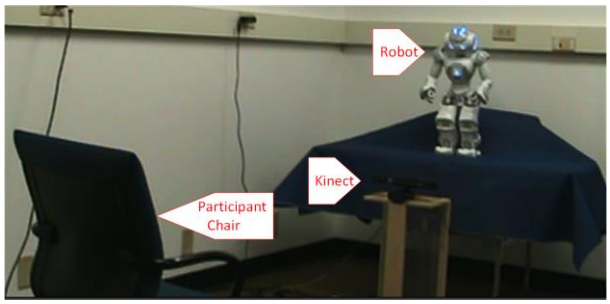

Fig. 1 RISTA system picture

software development kit (SDK). We chose this robot due to its attractive appearance to the children, simplified but adequate motion range and patterns, as well as the stability and flexibility of its software development environment. NAO communicated with the participants using both speech and motion. Its default text-to-speech functions and voice were used to provide verbal instructions. The physical motions needed in the experiments were preprogrammed and stored in a software library, and were called whenever needed.

The gesture tracking module used Microsoft Kinect [14]. Its SDK provides robust functions for real time skeleton tracking and head pose estimation. Skeleton data were used for imitation performance evaluation, and the head pose was treated as a coarse attention indicator which revealed how much attention the participant paid to the robot. The supervisory controller was in charge of the system execution logic, communication, and data logging. The robot module and the gesture tracking module were distributed and they communicated with the supervisory controller using different threads to achieve parallel operation. In this way, the system was able to both monitor the performance of the participant and provide different prompts to the child. The collected data were logged for offline analysis.

The participants were seated facing the robot about 2m away, and the Kinect was placed between the participant's chair and the robot. Fig. 1 shows the implemented system.

B. Single Gesture Recognition

The single gesture recognition method (SGR) proposed by Zheng et al. [7] was used as a basic component of the proposed mixed gesture recognition and spotting algorithm. In the SGR, the input is a temporal sequence of gesture variables (as listed in Table I), which are computed from the subject's arm skeleton tracking data. Fig. 2 (b-d) shows some of gesture variables of the right arm as examples.

A correct gesture is defined by trajectory constraints (TC) under preconditions (PC). PC describes the basic regional and positional constraints of a gesture. Since a participant was instructed to follow the robot's gesture, the TC was defined as multiple gesture stages in accordance with the order in which the robot presented a gesture. The recognition of a stage is triggered by the completion of the previous stage(s). The output of SGR is the gesture stage computed based on the input data.

Four gestures were studied in the original work: raising one hand (Gesture 1), raising two hands (Gesture 2), waving (Gesture 3) and reaching arms out (Gesture 4). Gesture 1 and Gesture 3 can be accomplished by either the right or the left hand/arm. If the imitation data satisfy the first n TCs, the performance is graded as $n \times 10$ / number of TC. Gesture 2 and Gesture 4 need to be accomplished by both hands to receive a full score. The grades for two hands were averaged for the final score. The PCs and TCs for each gesture applied in the current study are shown equations (1) – (6). These rules together with the gesture variables were preselected by experienced psychologists and engineers based on the analysis of 16 children's gesture performing data in the repository (8 children with ASD and 8 typically developing children, ages 2-5 years). This method utilizes the most important features of the selected gestures while keeping a low computational complexity of the gesture recognition module.

TABLE I. GESTURE VARIABLES

Symbol	Definition
$\overrightarrow{SW}$, $\overrightarrow{EW}$	Vector from shoulder to wrist, Vector from elbow to wrist
W_y, E_y, S_y	Y coordinates of wrist, elbow, and shoulder joint
$\angle A1$	Angle between $\overrightarrow{SW}$ and negative Y axis
$\angle A2$	Angle between $\overrightarrow{SW}$ and YZ plane, when arm pointing forward
$\angle A3$	Angle between $\overrightarrow{EW}$ and XY plane
$\angle A4$	Angle between $\overrightarrow{SW}$ and positive X axis (right arm) or negative X axis (left arm).
$\angle A5$	Angle between $\overrightarrow{SW}$ and XY plane,
$\angle WES$	Angle between upper arm and forearm
H	Wrist raised height in Y direction
D	Wrist movement in X direction
T_{item}	Threshold for distances or angles
$(R)_{nf}$	Condition R in the parentheses should be held for n consecutive frames

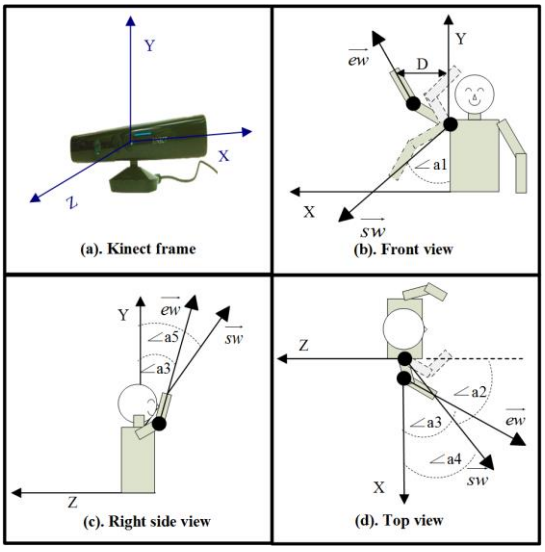

Fig. 2 Kinect frame and participant's gesture view

$$PC_{Wave} = \{\angle A1 < T_{ang1} \vee \angle A2 < T_{ang2} \vee W_y > S_y \vee \angle A3 < T_{ang3}\} \quad (1)$$

$$TC_{Wave} = \{(W_y > E_y)_{nf} \wedge \mathrm{H{>}T}_{up}, (\mathrm{W}_y > S_y)_{nf}, (\text{Only one hand})_{nf}, \quad (2)$$
$$\mathrm{D{>}T}_{dis} \text{ in one direction}, \mathrm{D{>}T}_{dis} \text{ in both directions }\}$$

PC_{Wave} Implies the gesture shall be started from raising an arm in front of the body. The first 3 constraints of TC_{Wave} requires that only one wrist shall be raised until higher than the shoulder. The last 2 constraints of TC_{Wave} indicate that the raised hand should be waved from side to side.

$$PC_{RaiseHand(s)} = \{\angle A1 < T_{ang3} \vee \angle A2 < T_{ang2}\} \quad (3)$$

$$TC_{RaiseHand(s)} = \{(W_y > E_y)_{nf} \wedge \mathrm{H{>}}T_{up}, (\mathrm{W}_y > S_y \wedge \angle WES > T_{ang4})_{nf}, (\angle WES > T_{ang5} \wedge (\pi - \angle A1) < T_{ang6})_{nf}, (\text{Only One Hand (raising one hand gesture)})_{nf}, (\mathrm{D{<}T}_{dis})_{nf}\} \quad (4)$$

These conditions represent that the hand(s) shall be raised from low to high in front of the body ($PC_{RaiseHand(s)}$) until they are gradually stretched straight and held still for a while ($TC_{RaiseHand(s)}$).

$$PC_{ReachArmsOut} = \{\angle A1 < T_{ang3} \vee \angle A4 < T_{ang3} \vee \angle A5 < T_{ang2} \vee \overrightarrow{SW}_y < 0 \vee (\overrightarrow{SW}_x > 0 \text{ (right arm) or } \overrightarrow{SW}_x < 0 \text{ (left arm)})\} \quad (5)$$

$$TC_{ReachArmsOut} = \{|\pi/2 - \angle A1| < T_{ang7})_{nf}, (|\angle A4| < T_{ang8})_{nf}, (\angle WES > T_{ang9})_{nf}\} \quad (6)$$

These rules imply that the arms shall be raised from a low position at the side of the body ($PC_{ReachArmsOut}$), and then stretched out evenly on each side ($TC_{ReachArmsOut}$).

In this work $n = rL$, where $r = 0.8$ and L is the length of the sliding time window. This indicates that the continuous constraint has to satisfy as least 80% of the sliding time window. This is important for gesture spotting discussed later. In the experimental study, psychologists and engineers found that the threshold values listed in Table II were suitable for the participants. Note that those values may need to be adjusted for other user studies with different participant groups.

TABLE II. GESTURE THRESHOLDS

Variable	value	Variable	value
T_{up}	10cm	T_{dis}	[10,15] cm
T_{ang1}, T_{ang2}	π	T_{ang3}	$\pi/2$
T_{ang4}	$[\pi/4, \pi/3]$* $\pi/2$**	T_{ang5}	$[\pi/5, \pi/3]$* $[\pi/5, \pi/4]$**
T_{ang6}	$[\pi/6, \pi/4]$* $[\pi/4, \pi/2]$**	T_{ang7}	$\pi/4$
T_{ang8}	$[\pi/5, \pi/4]$	T_{ang9}	$[\pi/4, \pi/3]$

* Raising one hand ** Raising two hands

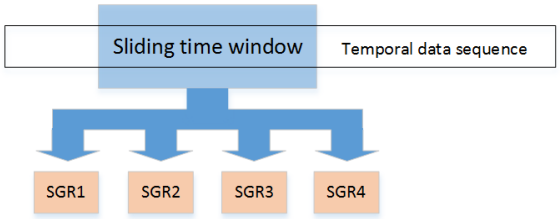

Fig. 3 Mixed gesture recognition data flow

A mixed gesture recognition algorithm is introduced within the following context. The robot demonstrates a continuous sequence of the four previously mentioned single gestures, and asks the child to imitate. The child might start and stop at any time and may or may not imitate all the demonstrated gestures. The task for the robot was to recognize when and what gestures were imitated as well as the quality of the imitation.

C. *Mixed Gesture Recognition and Spotting (MGRS)*

The newly proposed MGRS solves two problems in the mixed gesture prompting environment: a) how to recognize different gestures in parallel from the same input data sequence; and b) how to spot the start and the end points of each detected gesture. MGRS embeds the SGR as its components in a novel framework to address these two challenges. The four gestures described previously are presented as examples here; the proposed MGRS is not limited to those 4 gestures alone.

An initial imitation stage can evolve into different gestures. For example, both raising one hand and waving start with raising a hand from low to high. Therefore, the data subsequence extracted by a sliding time window is sent to all four SGRs for computing their stages. SGR1 to SGR4 represent the single gesture recognition algorithm for Gesture1 to Gesture4 in Fig. 3.

A gesture is detected if a data subsequence matches with its SGR's detecting stages. The hypothesis is that the more a data subsequence matches with a gesture's stages, the better it represents the corresponding gesture. This idea is similar to correlation based template matching in computer vision [15]. The start and end of this subsequence are the start and end of the corresponding gesture, respectively. This is formally defined as searching for T_{start} and T_{end} that satisfies

$$\arg\max_{T_{start}, T_{end}} stage = SGR(Data(T_{start}, T_{end})). \quad (7)$$

Given a gesture's SGR, we would like to find a data subsequence $Data(T_{start}, T_{end})$ which starts from T_{start} and ends at T_{end} that reaches a stage higher than the stage reached by any other subsequences overlapped with or adjacent to $Data(T_{start}, T_{end})$. This local optimum can be computed by updating the sliding window's position and length to refresh the stages detected by the SGR accordingly.

Consider the gesture of "raising one hand" as an example. It contains five stages in its TCs. In Fig. 4, blue blocks represent gesture imitation subsequences. The "raising one hand" gesture imitation subsequence starts from stage 1 (S1)

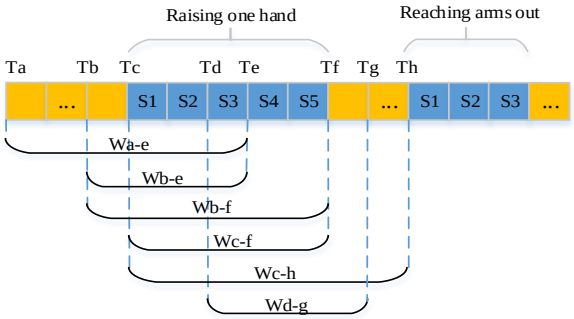

Fig. 4 MGRS algorithm demonstration

and ends with stage 5 (S5). The yellow blocks are non-imitation data, which can be unpurposive movements, resting postures, and so on. The sliding time window is updated in two nested loops: i) shifting its start point; and ii) with the same start point, adjusting its length from 1s to 5s. A gesture can be imitated within 1s to 5s. Ta to Th represent time points along the data sequence, and Wx-y means that the sliding window's start and end time points are x and y, respectively.

The following example illustrates how the algorithm computes the correct result as "the imitation reaches S5" and "the start and end point is Tc and Tf." The sliding time window updates from right (earlier) to left (later).

1) At Wa-e, the non-imitation data is over 20% of the window, and thus the subsequence violates the continuous constrains in the TCs. As a result, the current stage is set as 0;

2) The sliding window in then moved and at Wb-e, the window satisfies some of the TCs as only a small amount of non-imitation data is included. However, the window only contains data up to S3. So S3 is recorded as the current gesture stage, and the start and end time are noted as Tb and Te.

3) When the window length is extended to Wb-f, the window includes S5. In this case S5 is recorded as the current gesture stage, and the start and the end points are Tb and Tf.

4) When the window is moved at Wc-f, the window tightly cuts the data from S1 to S5, so the recorded state is still S5, but the start and the end points are refreshed to Tc and Tf.

5) At Wc-h, although the window includes the data from S1 to S5, but a large amount of non-imitation data after S5 prevents satisfaction of the continuous constraints. So the results of step 4 are kept.

6) At Wd-g, S1 and S2 are excluded. Because SRG does not jump previous gesture stages to reach later stages, the results of step 4 will not be refreshed.

Therefore, the final result is a gesture reaching stage 5 (S5) with a start point of Tc and end point of Tf.

This updating procedure can be executed using Algorithm1. Here *m* and *k* are time variables refreshed at each frame in the data sequence. *GestStage(m)* records the highest gesture stage that the subsequence reaches at time *m*. $T_{start}(m)$ and $T_{end}(m)$ mark the start and end time points of the whole gesture period where *GestStage(m)* belongs. Starting from the first frame of the data sequence, a 1 second length (LB, short for lower bound) data subsequence is used to compute an initial result using SGR. Then by appending frames onto the current sequence, 1 frame per update, SGR refreshes the results. If a higher stage is reached, it is recorded and its T_{start} and T_{end} are refreshed. This procedure is executed until the subsequence's length reaches the 5 second upper bound (UB). At that point, the sliding time window's start point is pushed forward by 1 frame and the above procedure is repeated. The iterations are executed until the end of the recognition period. Every gesture's iteration is updated in parallel as a new frame's data becomes available, and its results are recorded individually.

Note that this algorithm is an *example* of how to program the MGRS, but MGRS is not limited by it. Any computation procedure that reflects the goal in equation (7) can be applied.

Algorithm1

GestStage(1:DataLength) = 0;
T_{start} (1:DataLength) = 0;
T_{end} (1:DataLength) = 0;
for *k* = 1 : DataLength
    [*Stage1*] = **InitializeSGR**(Data(*k* : *k* + LB));
    GestStage(*k* : *k* + LB) = *Stage1*;
    T_{start} (*k* : *k* + LB) = *k*;
    for *m* = *k*+ LB +1 : *k*+ UB
        [*Stage2*] = **UpdateSGR**(Data(m), *Stage1*);
        if *Stage2* >= *Stage1*
            GestStage(*m*) = *Stage2*;
            T_{start} (*m*) = *k*;
            T_{end} (*m*) = *m*;
        end
        Stage1 = *Stage2*;
    end
end
Return (*GestStage*, T_{start} , T_{end});

III. Experimental Procedure

A. Participants

The MGRS algorithm should be able to successfully detect performed gestures as well as avoid giving false positive results when no targeted gestures are performed. Therefore, in this pilot study we selected participants with different imitation baseline levels, which helped us to collect imitation data ranging from good completion to non-completion. Two TD children (1 male and 1 female) and 4 children with ASD (3 males and 1 female) participated in this experiment. This group size is small due to the limited participant pool.

Table III lists the characteristics of the participants. Those in the ASD had received a clinical diagnosis of ASD based on

TABLE III. Participant Characteristics

Mean (SD)	ADOS CS	MSEL	SRS-2	SCQ	Age (Years)
ASD	8.50 (1.73)	51.00 (4.00)	76.50 (16.60)	20.50 (7.77)	4.61 (0.60)
TD	NA	NA	42.50 (3.54)	1.50 (0.71)	4.63 (0.01)

DSM-IV-TR [16] criteria. They met the spectrum cut-off of the Autism Diagnostic Observation Schedule Comparison Score (ADOS CS) [17], and had existing data regarding cognitive abilities from the Mullen Scales of Early Learning (MSEL) [18] in the clinical registry. Parents of participants in both groups completed the Social Responsiveness Scale–Second Edition (SRS-2) [19], and Social Communication Questionnaire Lifetime Total Score (SCQ) [20] to index current ASD symptoms.

B. Task and Protocol

This study was approved by the Vanderbilt Institutional Review Board (IRB). All the experiments were supervised by qualified clinicians and engineers. Videos of the experimental procedures were recorded for algorithm validation. The experiment had two steps:

Step1. Participant warmed up with the robot-administrated single gesture session "TrialA" of the previous work [7], for all 4 gestures. In this step, the robot first showed the participant a gesture, and then asked the participant to copy it. If the participant imitated the gesture correctly, the robot verbally praised the child. Otherwise, the robot provided a verbal explanation of what was wrong. This was intended to inform the participant that he/she was expected to copy the robot's gesture.

Step2. This was the mixed gesture imitation training part, consisting of 4 trials. Before each trial, the robot first mirrored the participant's physical motions for 15s to help the participant feel he/she was playing with the robot as a peer. Then, the robot asked the participant to now copy its gestures. The robot showed all 4 gestures twice in random order, all accompanied by background music. The "wave" gesture lasted for 4.4 seconds and the other 3 gestures lasted for 3.2 seconds. Two adjacent gestures were separated by a short transitional motion. In total, each trial lasted for about 49 seconds.

From the logged data we analyzed: 1) the accuracy of the MGRS algorithm; 2) the participant's attention on the robot; and 3) the participant's imitation performance.

IV. Experimental Results

A. Gesture Recognition and Spotting Results

To validate the accuracy of the MGRS algorithm, the results obtained from the MGRS were compared with an experienced therapist's ratings on the participants' imitation. From the videos recorded during the experiments, the therapist manually marked each gesture's start and end time as the ground truth. It is difficult for a therapist to identify and log different gesture stages similar to what a computer can do. Therefore, only gesture stages that were intuitively recognizable were marked. Accordingly, those marked stages were compared with those recognized by the MGRS.

We assessed two aspects of the MGRS: 1) could the algorithm identify a gesture correctly? 2) if the identification was correct, did the algorithm spot the start and end time of this gesture correctly? Table IV lists the number of gestures detected by the human therapist and the MGRS algorithm in the experiment. Seventy-three (5 detection misses) out of 78 human coded gestures (93.59%) were correctly detected by the MGRS, while 14 out of 87 MGRS detected gestures (16.09%) were false detections.

TABLE IV. Comparison between MGRS Recognition and Human Coding Results

Gesture	Human coded	Algorithm detected	False	Miss
Gesture1	17	14	0	3
Gesture2	19	24	5	0
Gesture3	19	23	6	2
Gesture4	23	26	3	0
Total	78	87	14	5

TABLE V. Start and End Time Deviation for Correctly Detected Gestures in Table IV

Mean (SD)	Start time deviation (s)	End time deviation (s)
Gesture1	0.25 (0.18)	0.50 (0.47)
Gesture2	0.46 (0.32)	0.48 (0.46)
Gesture3	0.88 (0.65)	0.97 (0.64)
Gesture4	0.33 (0.28)	0.54 (0.45)

For gestures that were successfully detected by the MGRS algorithm, the deviation between the human detected and MGRS detected start and end time was calculated. From Table V, we can see that the average deviation of the start and end time in all cases were smaller than 1s.

B. System Tolerance, Participants' Attention on the Robot and Imitation Performance

All participants completed the experiments except one child with ASD who did not complete 2 trials. Thus we had data for 14 trials from the children with ASD, and 8 trials from the TD children. The small sample was not sufficient for statistical significance testing. As a result, only the mean and standard deviation values are presented here.

The attention that the children paid on the robot was closely connected to his/her imitation performance. Attention can be coarsely estimated by head pose [21]. In this study, we utilized head pose estimation to infer the degree to which a participant was attending to the robot. A box of 85.77 cm × 102.42 cm around the robot (which covered the robot's full range of motion with a small margin) was set as the attention reference region. We assumed that a participant paid attention to the robot if his/her head faced toward this defined region. Table VI lists the total time that each group paid attention to the robot. The ratios are the percentage of the time spent facing the robot within a trial.

Due to the deficit of social communication, children with ASD usually pay significantly less attention to the social cues compared to the TD children. However, children with ASD still spent an average of 39.97% of trials facing the robot. This was less time compared to the single gesture sessions (60% in robot session, 42% in human session) as found in our previous study [7]. However, considering the increased task

TABLE VI. Participant's Attention Spent on the Robot

Mean (SD)	Time on Robot(s)	Ratio (%)
ASD	19.6 (9.98)	39.97 (20.27)
TD	39.34 (3.31)	80.41 (6.76)

complexity, this result was not surprising. TD children spent a majority of trials looking at the robot since they were interested in the robot as reported by their parents and experiment supervisors.

The participants' imitation performance was analyzed based on the MGRS algorithm. The scores of each gesture were normalized to [0, 10]. On average, children with ASD got a score of 8.71 (SD: 1.59) out of 10, while TD children got 9.58 (SD: 0.93). Children with ASD took 3.79s (SD: 1.91s) to finish one gesture on average, while their TD peers took 2.8s (SD: 1.37s). The results confirmed expected results that children with ASD would have a lower gesture imitation ability compared to that of TD children of similar ages.

V. Discussion and Conclusion

We proposed a robotic system with the RISTA architecture that aims to teach imitation skills to young children with ASD. This work utilized a humanoid robot for gesture prompting and a non-invasive setup for effectively evaluating the participants' performance. This system extends the previous robot-mediated intervention from single gesture to mixed, multiple gestures. Naturally, the participant was allowed to imitate different gestures in any order, and at any time during the intervention. In order to achieve mixed gesture imitation recognition, we developed a novel algorithm, the MGRS algorithm, which not only detects the imitated gestures, but also spots the start and end times of the performed gestures.

A preliminary user study showed that the MGRS algorithm achieved high accuracy in both gesture recognition and spotting. The RISTA system was well tolerated by the young children, attracted their attention, and showed great potential for extending the training of imitation skills for children with ASD.

There were some limitations in this study. The proposed MGRS algorithm was tested with only 4 gestures. Extensive tests with more gesture categories are necessary in the future to examine MGRS's scalability and robustness. Although the SGRs embedded in the MGRS are rule-based, the framework of MGRS is not limited to rule-based components. In fact, any single gesture detection algorithm can be embedded in this framework. If a large number of gestures are needed to be detected in parallel, then the correlation and similarity between those gestures can be used for pruning to avoid repeated computation. Finally, since the user group was small, any conclusion drawn based on this study requires further validation with larger samples and more pervasive analyses. Yet the current work may provide a beneficial preliminary framework for developing and evaluating multi-gesture imitation intervention for young children with ASD.

References

[1] "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, United States, 2010," *Morbidity and mortality weekly report. Surveillance summaries (Washington, DC: 2002)*, vol. 63, p. 1, 2014.

[2] American Psychiatric Association, *The Diagnostic and Statistical Manual of Mental Disorders: DSM 5*: bookpointUS, 2013.

[3] B. Ingersoll, "Brief report: Effect of a focused imitation intervention on social functioning in children with autism," *Journal of autism and developmental disorders*, vol. 42, pp. 1768-1773, 2012.

[4] B. Ingersoll, "Pilot Randomized Controlled Trial of Reciprocal Imitation Training for Teaching Elicited and Spontaneous Imitation to Children with Autism. ," *Journal of Autism and Developmental Disorders*, vol. 40, 2010.

[5] Z. E. Warren and W. L. Stone, "Best practices: Early diagnosis and psychological assessment," in *Autism Spectrum Disorders*, David Amaral, Daniel Geschwind, and G. Dawson, Eds., ed New York: Oxford University Press, 2011, pp. 1271-1282.

[6] B. Robins, K. Dautenhahn, and J. Dubowski, "Does appearance matter in the interaction of children with autism with a humanoid robot?," *Interaction Studies*, vol. 7, pp. 509-542, 2006.

[7] Z. Zheng, S. Das, E. M. Young, A. Swanson, Z. Warren, and N. Sarkar, "Autonomous robot-mediated imitation learning for children with autism," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, 2014, pp. 2707-2712.

[8] Z. Warren, Z. Zheng, S. Das, E. M. Young, A. Swanson, A. Weitlauf, *et al.*, "Brief Report: Development of a Robotic Intervention Platform for Young Children with ASD," *Journal of autism and developmental disorders*, pp. 1-7, 2014.

[9] J. Greczek, E. Kaszubksi, A. Atrash, and M. J. Matarić, "Graded Cueing Feedback in Robot-Mediated Imitation Practice for Children with Autism Spectrum Disorders," *Proceedings, 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014) Edinburgh, Scotland, UK* Aug. 2014.

[10] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, *et al.*, "KASPAR–a minimally expressive humanoid robot for human–robot interaction research," *Applied Bionics and Biomechanics*, vol. 6, pp. 369-397, 2009.

[11] I. Fujimoto, T. Matsumoto, P. R. S. De Silva, M. Kobayashi, and M. Higashi, "Mimicking and evaluating human motion to improve the imitation skill of children with autism through a robot," *International Journal of Social Robotics*, vol. 3, pp. 349-357, 2011.

[12] E. Bekele, U. Lahiri, A. Swanson, Julie A. Crittendon, Zachary Warren, and N. Sarkar, "A Step towards Developing Adaptive Robot-mediated Intervention Architecture (ARIA) for Children with Autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, pp. 289-299, 2013

[13] *Aldebaran Robotics*. Available: http://www.aldebaran-robotics.com/en/

[14] *Microsoft Kinect for Windows*. Available: http://www.microsoft.com/en-us/kinectforwindows/

[15] J. P. Lewis, "Fast template matching," in *Vision interface*, 1995, pp. 15-19.

[16] *Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the Diagnostic Criteria from DSM-IV-TR*, Fourth ed. Washington D.C.: American Psychiatric Association, 2000.

[17] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, "Autism Diagnostic Observation Schedule–2nd edition (ADOS-2)," ed: Western Psychological Services: Torrance, CA, 2012.

[18] E. M. Mullen, *Mullen scales of early learning: AGS edition*. Circle Pines, MN: American Guidance Service, 1995.

[19] J. N. Constantino and C. P. Gruber, "The social responsiveness scale," ed: Los Angeles: Western Psychological Services, 2002.

[20] M. Rutter, A. Bailey, and C. Lord, "The Social Communication Questionnaire," ed: Los Angeles, CA: Western Psychological Services, 2010.

[21] E. T. Bekele, U. Lahiri, A. R. Swanson, J. A. Crittendon, Z. E. Warren, and N. Sarkar, "A step towards developing adaptive robot-mediated intervention architecture (ARIA) for children with autism," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, pp. 289-299, 2013.