Recognizing Stereotypical Motor Movements in the Laboratory and Classroom: A Case Study with Children on the Autism Spectrum

Fahd Albinali[1], Matthew S. Goodwin[1, 2], and Stephen S. Intille[1]

[1]Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139 USA

[2]The Groden Center, Inc.
86 Mount Hope Avenue
Providence, RI 02906 USA

albinali | mgoodwin | intille @ mit.edu

ABSTRACT

Individuals with Autism Spectrum Disorders (ASD) frequently engage in stereotyped and repetitive motor movements. Automatically detecting these movements in real-time using comfortable, miniature wireless sensors could advance autistic research and enable new intervention tools for the classroom that help children and their caregivers monitor and cope with this potentially problematic class of behavior. We present activity recognition results for stereotypical hand flapping and body rocking using data collected from six children with ASD repeatedly observed in both laboratory and classroom settings. In the classroom, an overall recognition accuracy of 88.6% (TP: 0.85; FP: 0.08) was achieved using three sensors. Challenges encountered when applying machine learning to this domain, as well as implications for the development of real-time classroom interventions and research tools, are discussed.

Author Keywords

Autism, accelerometers, activity recognition.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Algorithms, Human Factors, Measurement.

INTRODUCTION

Health researchers in many disciplines lack effective tools for unobtrusively acquiring information about peoples' behavior in natural settings. Ubiquitous computing systems that detect certain behaviors might create new opportunities to improve scientific understanding of the interaction between context, behavior, and health. The goal of the current work is to use ubiquitous monitoring tools for the automated detection of stereotypical behavior observed in persons with Autism Spectrum Disorders. Autism Spectrum Disorders (ASD) affect as many as 1 in 150 children [1] and are characterized by deficits in socialization and communication, including stereotypical behavior [2]. Stereotyped behaviors are generally defined as repetitive interests and/or motor or vocal sequences that appear to the observer to be invariant in form and without any obvious eliciting stimulus or adaptive function [3]. The current work focuses on stereotypical motor movements. Several stereotypical motor movements have been identified [4], the most prevalent among them being body-rocking, mouthing, and complex hand and finger movements [5]. The majority of research in ASD focuses on social and communication deficits, rather than on restricted and repetitive behavior [4]. This is a potential problem given the high prevalence of stereotypical motor movements reported in individuals with ASD (e.g., [6]).

One reason why stereotypical motor movements may not be as thoroughly studied is because appropriate tools for measuring the behavior are not available to the research community. In this work, we present a case study on the automatic identification of stereotypical body rocking and hand flapping activity in children with ASD gathered from wireless accelerometers. Stereotypical body rocking and hand flapping are examples of movements that occur frequently in people with mental retardation and developmental disabilities [4], and less frequently in typically developing children and adults.

Impact of Stereotypical Motor Movements

When severe, stereotypical motor movements can present several problems for individuals with ASD and their caregivers. First, persons with ASD often engage in stereotypical motor movements for the majority of their waking hours. Second, if unregulated, stereotypical motor movements can become the dominant behavior in an individual with ASD's repertoire and interfere with the acquisition of new skills and performance of established skills (e.g., [7]). Third, engagement in these movements is socially inappropriate and stigmatizing and can complicate social integration in school settings and the community [8]. Finally, stereotypical motor movements can lead to self-injurious behavior under certain environmental conditions [9].

UbiComp 2009, Sep 30 – Oct 3, 2009, Orlando, Florida, USA.

Tools for Measuring Stereotypical Motor Movements

There are currently no tools for clinicians or caregivers to easily, accurately, and reliably monitor stereotypical motor movements. Traditional measures of stereotypical motor movements rely primarily on paper-and-pencil rating scales, direct observation, and video-based methods [10], all of which have limitations.

Paper-and-pencil rating scales typically involve a global impression of the frequency and/or severity of stereotypical motor movements based on general, non-specific observations. Several paper-and-pencil rating scales have been developed that ask an informant to give a global impression of an individual's stereotypical motor movements [4]. From a measurement standpoint, informant rating scales are subjective, can have questionable accuracy, and fail to capture inter-individual variations in the form, amount, and duration of stereotypical motor movement [11].

Direct observation also involves a rating but the focus is on the direct observation of specific behaviors. The observer watches and records a sequence of stereotypical motor movements. According to Sprague and Newell [10], the following factors, among others, can make direct observational measures unreliable: (a) Reduced accuracy in observing and documenting high-speed motor sequences; (b) Difficulty determining when a sequence has started and ended; (c) Limitations in the ability to observe concomitantly occurring stereotypical motor movements; and (d) Limitations in the ability to note environmental antecedents and record stereotypical motor movements at the same time.

Video-based methods involve video capture of behavior and off-line coding of stereotypical motor movements by an expert. The ability to view videos repeatedly and to slow playback speeds makes video-based methods more reliable than paper-and-pencil and direct observation methods. Video-based methods, however, are tedious and time consuming. The necessity to code videos off-line also precludes real-time monitoring. Combining video recording with other tagging technologies to permit practical, semi-automatic logging is an area of active research [12].

Goal: Explore the Possibility of Real-Time Recognition of Stereotypical Motor Movements

The aim of the current work is to explore whether wireless accelerometer sensor technology and pattern recognition algorithms can provide an automatic, real-time measure of stereotypical motor movements that may be more objective, detailed, and precise than rating scales and direct observation, and more time-efficient than video-based methods. An algorithm that achieves good recognition performance could operate for much longer periods of time than than a human observer.

In the remainder of this paper, we describe experiments we have performed to determine whether pattern recognition techniques using mobile wireless accelerometers that have shown promise in other domains of recognition of posture, mobility, exercise, and everyday activities can be adapted to create a real-time tool for stereotypical motor movement monitoring in children with ASD.

RELATED WORK

We are aware of only one published attempt to apply pattern recognition algorithms to this domain.

Automatically Detecting Stereotypical Motor Movements

Westeyn et al. used accelerometers and pattern recognition algorithms in pilot work to detect stereotypical motor movements [13]. While 69% of hand flapping events were automatically and accurately detected in this work using Hidden Markov Models, the data were acquired from individuals mimicking the actual behaviors – the work did not observe children with ASD actually performing the behaviors.

Using Pattern Recognition to Detect Other Physical Activities

A growing body of work shows that wearable accelerometers can be used to detect activities, such as postures, ambulation, exercise, and even household activities (e.g., [14-16]). A variety of methods and models have been used for feature generation and classification. Our focus in this work is not on any *particular* activity recognition algorithm, per se, but instead on the issues one encounters when trying to apply pattern recognition to the problem of monitoring stereotypical motor movements in constrained and naturalistic settings.

Most prior work in accelerometer-based activity recognition uses supervised learning strategies. Activities are performed by people wearing wired or wireless accelerometers on one or more body locations. Annotators (usually the researchers) then use video or audio to label the start and end points of each behavior of interest. Algorithms are then tested on the datasets using cross-fold validation. We use this same approach, but describe the challenges we have encountered in the stereotypical motor movement domain.

DATA COLLECTION

The current investigation consisted of a series of six single case studies, with direct replication across participants. For each participant, the study included repeated observations of body rocking, hand flapping, and/or simultaneous body rocking and hand flapping while children wore sensors in laboratory (Study 1) and classroom (Study 2) settings.

Participants

Six participants were recruited from The Groden Center, RI, a school for children and young adults with ASD. The study was approved by a human subjects review board and parental consent was obtained for each participant. Children included in the study: (1) Had a documented DSM-IV-TR diagnosis of ASD made by a licensed psychologist familiar with the child; (2) Were between the ages of 12-20 yrs.; (3) Had a clinically significant score on the Stereotyped

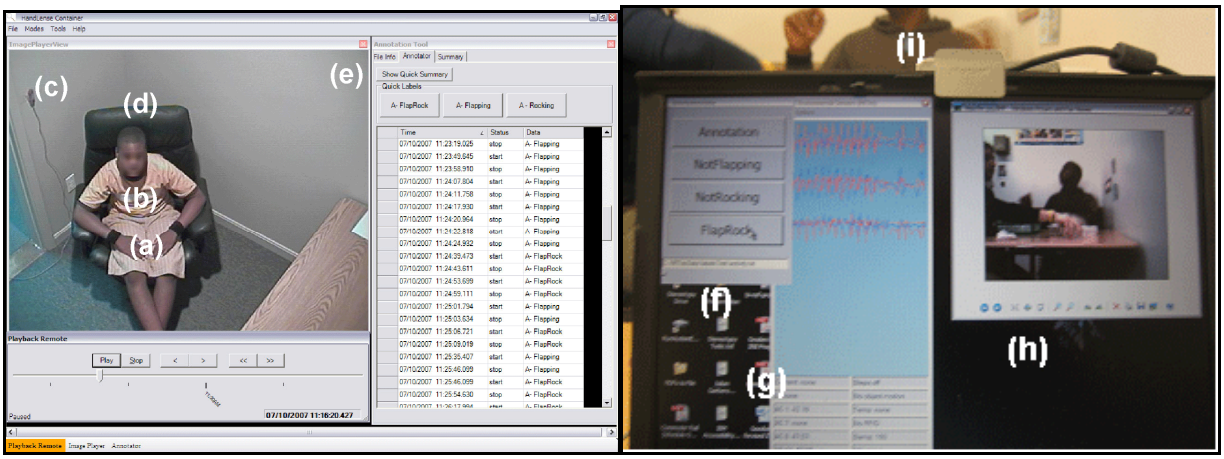

Figure 1. (a) A wireless accelerometer placed on each wrist. (b) A wireless accelerometer placed on the chest. (c) Receiver for sensor data. (d) An image of a child in the laboratory setting. (e) The video coding software that allows frame-accurate annotation. (f) The real-time activity annotator. (g) The acceleration data window plotting data streams in real-time. (h) The video window with images being captured. (i) USB camera clipped on to the top of the laptop.

Behavior subscale of the Repetitive Behavior Scale-Revised (RBS-R; [17]) for body rocking and/or hand flapping; (4) Tolerated the wireless sensors; and (5) Exhibited, on average, at least 10 hand flapping or body rocking incidents per hour.

Sensors

Each participant wore three wireless accelerometers [18] placed simultaneously on the left wrist and right wrist using wristbands, and on the torso using a thin strip of comfortable fabric tied around the chest (see Figure 1(a-b)). The wrists and torso were chosen because stereotypical hand flapping and body rocking are associated with movements in these areas. The sensors were small enough to be worn on these locations comfortably and without restricting movement. All participants tolerated wearing the sensors for the duration of each observation. Also, visual inspection of each participant's real-time acceleration data prior to analysis confirmed that there were no equipment failures or other problems occurring (improper attachment, weak signal strength, unusual amount of signal loss, etc.).

In the configuration used in the experiments, the devices were set to transmit 3-axis +/- 2g motion data at 60Hz to a nearby receiver (Figure 1(c)). The body can block the 2.4GHz range low-power radio signal, so there is occasional signal loss experienced that the pattern recognition algorithms must compensate for. The receiver was plugged into a standard computer (desktop in the laboratory and laptop in the classroom), where data from the three sensors were synchronized and saved to disk. Simultaneously, a video camera was used to capture video of the scene that could be synchronized with the accelerometer streams and used for annotation of activity.

Setting and Procedure

We undertook data collection in both laboratory and classroom settings to determine the accuracy of recognition performance across both constrained and real-world environments. Participants were seated during all observations in both environments.

Laboratory (Study 1)

Observations were undertaken in a laboratory setting at The Groden Center where there were limited stimulus materials, one-to-one monitoring by a familiar teacher, and no other students present. The lab is divided into three areas: (1) A soundproof room equipped with a discrete, ceiling mounted camera and microphone to record observations; (2) An observation area behind the glass; and (3) An adjacent office containing a computer and video monitor.

While wearing the sensors, participants were observed in the lab while sitting in a comfortable chair with a familiar teacher (see Figure 1(d)). There were no structured activities involved in these observational sessions. However, teachers familiar with the participants were invited to bring objects (e.g., headphones, books, toys) that participants typically interacted with when engaging in stereotypical motor movements.

Classroom (Study 2)

Observations were also undertaken in a classroom setting at The Groden Center, which included a diverse set of stimuli, demands for shared attention, and other students present. While wearing the sensors, each participant was observed on two separate occasions in class while seated at a desk. These observations included typical classroom activities (e.g., eating lunch, spelling program, sorting), with participants working both on their own and with a familiar teacher.

Over a period of 12 months and during regular school hours (9:00-3:00), we recorded one 10-30 minute session per week per participant[1]. This data collection effort resulted in

[1] When we began this work, the primary focus was to collect large numbers of examples of stereotypical movements quickly. Sessions that

ID	Total Duration (min:sec)	% Engaged	M (SD) (sec)	Num Stereo| Total Stereo	Consistency
6	47:42	28%	7 (6)	3 | 372	Mild
7	18:18	17.5%	3 (2)	2 | 345	Very
8	10:00	8.5%	4 (2)	3 | 149	Very
9	36:53	45%	9 (12)	2 | 240	Very
10	32:14	48%	7 (7)	3 | 253	Mild
11	67:28	71%	21 (23)	2 | 199	Very

Table 1. Summary of participant stereotypical movements

6.5 hours of data for Study 1 and 4.75 hours of data for Study 2. These data included at least 2 sessions per participant per study.

Stereotypical Motor Movements

One of the first challenges we encountered was simply the diversity of stereotypical motor movements observed in our participants, and the difficulty associated with annotating those movements. Table 1 summarizes quantitative and qualitative stereotypical motor movement characteristics of the participants averaged across observation sessions in the lab and classroom.

The Total Duration is the total time spent engaged in stereotypical motor movements across lab and classroom sessions. The % Engaged is the percentage of time participants engaged in stereotypical motor movements during the data collection sessions. M and SD are the mean duration and standard deviation of each participant's episodes of stereotypical movements. Num Stereo is the number of different types of stereotypical motor movements observed during all sessions and Total Stereo is the total number of episodes of those movements for each participant. This includes hand flapping, body rocking, and simultaneous hand flapping and body rocking – dubbed "flaprock." Finally, consistency is a subjective grade (none, mild, or very) assigned by a trained behavioral scientist indicating how consistent each participant's stereotypical motor movement appeared to be.

Annotation

Each session involved two observational coding procedures. The first, *real-time coding*, was undertaken during the sessions to see how well start time, end time, and type of stereotypical motor movement could be documented in real-time (i.e., live) by a trained observer. The second, *offline coding*, was undertaken after the sessions using video records and computerized annotation software.

had infrequent episodes of stereotypical movements (< 10 per hour) were aborted to reduce stress on participants, and the data were discarded. In hindsight, given the difficulty of acquiring examples, data from these sessions should have been coded and used in the analysis.

Real-time Coding

Start time, end time, and type of stereotypical motor movement were coded in real-time using custom annotation software (see Figure 1(f)). The activity annotator included three buttons that corresponded to the stereotypical motor movements under observation (i.e., hand flapping, body rocking, flaprock). Pressing a button once marked the start of the corresponding stereotypical motor movement. Pressing a button a second time marked the end of the corresponding stereotypical motor movement.

Offline Coding (video records)

A digital camera (mounted in the ceiling of the laboratory; attached to the front of the laptop in the classroom (Figure 1(i)) was used to record each session. The camera was connected to a computer that synchronized the saved video with the accelerometer data streams. Start time, end time, and type of stereotypical motor movement were coded offline by *two independent raters* using a custom video coding software application (Figure 1(e)).

RECOGNITION EVALUATION AND EXPERIENCES

In this section, we describe in detail our experience applying physical activity pattern recognition to the stereotypical motor movement recognition domain.

Algorithm

Prior work [18] demonstrates that decision tree classifiers can be used to effectively recognize a variety of physical activities. We are ultimately interested in creating a real-time recognition tool, and decision trees have a desirable combination of properties: They have performed well in prior experiments reported in the literature on posture and ambulatory recognition, and they are fast to run once trained.

We use five time and frequency domain features computed for each acceleration stream. These are: (1) The distances between the means of the axes of each accelerometer to capture sensor orientation for posture; (2) Variance to capture the variability in different directions; (3) Correlation coefficients to capture the simultaneous motion in each axis direction; (4) Entropy to capture the type of stereotypical motor movement; and (5) FFT peaks and frequencies to capture differentiation between different intensities of the stereotypical motor movements. The features are computed for a window of data, assembled into a vector, and used as input to the C4.5 classifier in the WEKA toolkit [19]. WEKA is then used to evaluate classification performance using 10-fold cross validation.

Stereotypical motor movements were labeled as flapping, rocking, or flaprock (i.e., simultaneous flapping and rocking). Non-stereotypical motor movements were labeled as unknown segments. Including an unknown class resulted in highly skewed class distributions, such that the frequencies of stereotypical motor movements were substantially lower than the examples of the unknown class when stereotypical motor movements were not occurring. To reduce skewness in the present data, all classifiers used balanced data for training and natural imbalanced data for testing. Balancing the data

Exp	Description	Goals
#1	Trained using participant-dependent data and offline annotations from the laboratory. Tested using cross-validation.	1. Measure the performance of the classifier in a constrained setting. 2. Measure the agreement between 2 offline annotators. 3. Measure performance on agreement and disagreement segments.
#2	Trained using participant-dependent data and offline annotations from the classroom. Tested using cross-validation.	1. Measure the performance of the classifier in a naturalistic setting. 2. Measure the agreement between 2 offline annotators. 3. Measure performance on agreement and disagreement segments.
#3	Trained using participant-dependent data from the classroom environment and tested it on the lab data, and vice versa.	1. Measure the impact of inter-session variability. 2. Understand the impact of the setup on the quality of the training data and the classifier. 3. Compare the performance in the classroom and the lab.
#4	Trained and compared 3 different methods: (1) One-annotator training that uses offline annotations; (2) One-annotator training that uses real-time annotations; and (3) Two-annotator training that uses agreement segments from 2 offline annotators for training.	1. Understand the impact of the annotation (e.g. offline, real-time, multiple annotators) on the performance of the classifier. 2. Measure the agreement between offline and real-time annotations. 3. Determine and compare where errors occur in real-time and offline annotations scenarios.
#5	Trained the classifier with data from all the participants but one and tested the performance on the left out participant.	1. Measure the performance of the classifier using participant-independent data. 2. Determine if some stereotypical motor movements are more consistent across participants and therefore detectable using participant-independent training.

Table 2. Summary of Experiments

was done by randomly under-sampling the majority class (i.e. unknown) and re-sampling minority classes (i.e. stereotypical motor movements).

Nine acceleration streams (x, y and z from three accelerometers) were broken into 50% overlapping sliding windows of length 1 second. Our choice of 1 second was based on pilot work where we changed the window length from 200 ms to 5 seconds and measured the performance of the C4.5 classifier over pilot datasets. A window of 1 second obtained good overall accuracy while minimizing the classification delay.

Cubic spline interpolation was used to fill in missing data points (e.g. due to wireless signal loss). Windows that lost more than 50% of their expected data points were excluded from the analysis. This amounted to less than 1% of the data.

We conducted five types of analyses that are summarized in Table 2. To measure the performance of the activity classifier, we computed recognition accuracy, true positive rate (TP), false positive rate (FP), precision, and recall.

In what we will call *one-annotator training*, we perform 10-fold cross-validation over each participant's data and present averaged results across different sessions in the classroom and the lab. In *two-annotator training*, we train on only agreement segments between two annotators and test on the complete data including both agreement and disagreement segments. For the *agreement* portion of the data, we perform 10-fold cross-validation. For the *disagreement* portion, we train on the agreement data and test on the disagreement data. Results are then combined and averaged across sessions for each participant. Finally, we report on the percentages of agreement between two offline annotations and real-time-offline annotations using Cohen's Kappa inter-rater reliability statistic.

Experiment 1: Performance in a Laboratory

Table 3 shows the overall performance results of the algorithm averaged over multiple sessions for each participant in the lab. *Accuracy* is the average accuracy of the classifier across all sessions. *Accuracy (Agree)* is the accuracy of the classifier on examples where both offline annotators agreed. *Accuracy (Disagree)* is the accuracy of the classifier on examples where both annotators disagreed. *TP* and *FP* are the true and false positive rates, respectively. These are followed by precision and recall [19]. Finally, *K* is Cohen's Kappa, a statistic representing inter-rater reliability between two offline annotators.

The performance of the algorithm appears to be directly dependent on at least three factors: (1) The duration of each episode of stereotypical motor movement; (2) The percentage of time participants engaged in stereotypical motor movements; and (3) The consistency with which participants performed these movements.

Participants 6, 7, and 8 had the shortest mean duration for an episode of stereotypical motor movement (7, 3, and 4 seconds, respectively) and spent the least amount of time engaged in these movements (28%, 17.5%, and 8.5%, respectively). As expected, the recognition performance fpr these participants with respect to Precision and Recall is significantly lower than participants 9, 10, and 11 who engaged in stereotypical motor movements more often and for longer periods of time.

Participant 6 exhibited the most inconsistent stereotypical motor movements, displaying a range (i.e., topography,

Participant ID	Accuracy	Accuracy (Agree)	Accuracy (Disagree)	TP	FP	Precision	Recall	K
6	79.0%	82.5%	47.4%	0.75	0.08	0.60	0.75	0.82
7	92.2%	92.9%	51.1%	0.86	0.12	0.63	0.86	0.85
8	90.0%	91.0%	51.5%	0.76	0.07	0.49	0.76	0.89
9	93.3%	93.7%	56.6%	0.90	0.05	0.72	0.90	0.95
10	87.6%	90.0%	52.3%	0.87	0.09	0.81	0.87	0.86
11	95.0%	96.4%	68.3%	0.91	0.03	0.78	0.92	0.88
Mean	89.5%	91.1%	54.5%	0.84	0.07	0.67	0.84	0.87

Table 3. Performance of the classifier on 6 participants in laboratory (offline one-annotator training)

Participant ID	Accuracy	Accuracy (Agree)	Accuracy (Disagree)	TP	FP	Precision	Recall	K
6	84.6%	85.9%	64.9%	0.82	0.07	0.58	0.82	0.72
7	87.5%	88.9%	62.8%	0.89	0.12	0.75	0.88	0.83
8	89.7%	90.6%	47.6%	0.74	0.07	0.55	0.74	0.86
9	91.2%	93.7%	39.8%	0.90	0.05	0.88	0.90	0.92
10	88.0%	90.2%	43.0%	0.85	0.08	0.72	0.83	0.83
11	90.7%	93.0%	58.5%	0.90	0.07	0.86	0.90	0.88
Mean	88.6%	90.4%	52.8%	0.85	0.08	0.72	0.84	0.84

Table 4. Performance of the classifier on 6 participants in a classroom (offline one-annotator training)

intensity, duration) of different flapping and rocking movements. This resulted in both the lowest performance accuracy (79.0%) and the lowest TP rate (0.75). Conversely, participants 9 and 11 consistently engaged in the same stereotypical motor movements for 45% and 71% of the duration of the data, respectively and had the longest episodes with an average of 9 and 21 seconds, respectively. This resulted in the highest performance accuracies (93.3% and 95.0%, respectively) and the highest TP rates (0.90 and 0.91, respectively).

A major concern is the high false positive rates averaging 0.07 across all participants. For intervention applications that target specific stereotypical motor movements, the system would incorrectly deliver the intervention 7% of the time when the participant is not engaged in the behavior. A closer look at the distribution of FP errors across the different activities reveals that more than 75% of the FP errors are for the *unknown* class and less than 25% of the errors are shared between specific stereotypical motor movements. This brings the average false positive rate for specific stereotypical motor movements down to approximately 0.03, which is a more desirable FP rate when an intervention is to be delivered *only* when a stereotypical motor movement is occurring. Standard smoothing techniques may further reduce these errors.

Finally, the highest agreement between offline annotators is for participant 9 (kappa=0.95) and the lowest agreement is for participant 6 (kappa=0.82). To determine whether the majority of classification errors occurred on the subset of data where annotators were not in agreement, we evaluated the classifier on agreement and disagreement segments independently. Not surprisingly, the algorithm performed poorly on segments where there was disagreement between the annotators. However, this modestly impacted the overall performance of the classifier because the frequency of disagreement in offline annotation was relatively low, averaging less than 3% of the collected data for each participant.

Experiment 2: Performance in Classroom

Table 4 describes results from the classroom experiments. Similar to the lab setting, participants 9 and 11 showed the highest frequency and the longest duration of stereotypical motor movements and therefore performed the best with respect to accuracy (91.2% and 90.7%, respectively), TP rate (0.90 and 0.90, respectively), precision (0.88 and 0.86, respectively) and recall (0.90 and 0.90). The worst performance with respect to accuracy came from participant 6 who also performed the worst in the lab. Participant 6 was particularly difficult to annotate in the classroom because he frequently transitioned between different types of stereotypical motor movements of relatively short duration. It appears, however, the more constrained lab environment made it easier for the annotators to monitor and label the participant's stereotypical motor movements. This is reflected in the agreement between annotators with kappa for participant 6 being 0.72 in the classroom and 0.82 in the lab.

Similar to the lab setup, participants 6 and 8 had the worst performance in the classroom with respect to precision (0.58 and 0.55, respectively) and recall (0.82 and 0.74,

Participant ID	Train Classroom, Test Lab	Train Lab, Test Classroom
6	64.9%	64.8%
7	90.8%	86.8%
8	90.7%	91.9%
9	91.8%	83.9%
10	75.0%	73.4%
11	75.8%	87.7%
Mean	83.7%	81.4%

Table 5. Performance of the classifier using independent training and testing data from the classroom and the lab

respectively). However, for participant 7, the precision improved in the classroom relative to the lab (lab: 0.63; classroom: 0.75), likely because the number of stereotypical motor movements recorded in the classroom were more than double those recorded in the lab. This provided more examples for the classifier to learn the movements.

Experiment 3: Comparing Performance in Classroom and in Laboratory

To measure the impact of inter-session variability, we trained the classifier using data from the classroom environment and tested it on the lab data and vice versa.

From Table 5, we can observe that the average performance using classroom data for training (83.7%) is similar to using lab data for training (81.4%). Although the classroom is a less constrained setup where participants engaged in a wider range of movements and positions, it does not seem to have impacted the performance of the classifier.

Notably, however, the performance on participants 7 and 9 was significantly better when the classifier used the classroom data for training. This appears to be related to the number of training examples provided to the classifier. Specifically, we recorded an average of 108 episodes for participant 7 in the classroom per session versus 42 episodes in the lab and we recorded 48 episodes for participant 9 in the classroom per session versus 13 episodes in the lab. For participants 7 and 9, each episode resulted in an average of 7 and 16 training examples respectively. Having more training examples appears to result in higher classification accuracy for the classroom classifier.

Experiment 4: Comparing Performance using Real-time and Offline Annotations

Table 6 compares the impact of real-time and offline annotation on the performance of the classifier in both the lab and the classroom settings. The column labeled *Offline* reiterates the overall accuracy reported in previous sections with one-annotator training. *Real-time* describes the results using real-time annotations for training and offline annotations for testing. *Agreement* describes the results from training the algorithm on segments where both offline annotators agreed. Finally, *K* is Cohen's Kappa inter-rater reliability statistic that measures the agreement between the real-time and off-line annotations.

Our first observation is that a strong association exists between duration of stereotypical motor movements and performance of the real-time annotator. For example, participants 6, 7, and 8 have the shortest mean durations (7, 3, and 4 seconds, respectively) and the least percentage of engagement in stereotypical motor movements (28%, 17.5%, and 8.5%, respectively). The kappa values between offline and real-time annotations for these participants are also lowest in both the classroom and the lab. The real-time annotations and offline annotations differ in at least two ways. First, the real-time annotator frequently misses short episodes of stereotypical motor movements. For participant 8, this constituted approximately 33% of the episodes that were labeled offline. Second, when the real-time annotations overlap with corresponding offline annotations, the real-time onsets and offsets are shifted in time but biased slightly towards errors in onset. These two factors appear to reduce the performance of the real-time classifier relative to the offline one-annotator classifier, particularly when stereotypical motor movements are of short duration. Further, the impact of short duration movements seems more evident in the classroom environment where it is more likely that an annotator will miss subtle movements due to increased general activity.

Participant ID	Classroom				Lab			
	Offline	Real-time	Agreement	K	Offline	Real-time	Agreement	K
6	86.5%	75.8%	84.5%	0.42	79.0%	77.5%	77.7%	0.55
7	86.8%	80.4%	89.2%	0.32	96.5%	96.4%	92.8%	0.37
8	95.0%	91.1%	91.9%	0.54	95.8%	95.1%	91.7%	0.33
9	83.7%	82.2%	92.1%	0.76	86.0%	91.8%	93.6%	0.69
10	81.9%	85.9%	91.4%	0.71	77.5%	82.0%	87.5%	0.59
11	84.0%	82.6%	92.3%	0.68	83.2%	93.5%	95.3%	0.81
Mean	86.3%	83.0%	90.2%	0.57	86.3%	89.4%	89.8%	0.56

Table 6. Performance of the classifier using real-time and offline annotations

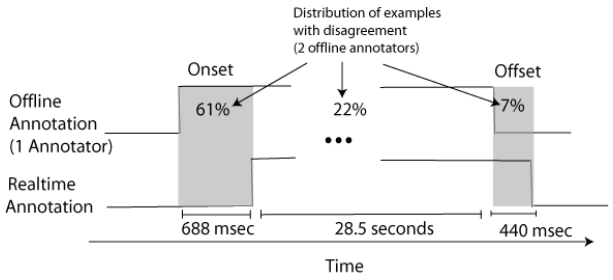

Figure 2. Distribution of disagreements between offline annotators with respect to the onset and offset of an activity for participant 11.

Our second observation is that when both the duration of stereotypical motor movements is long (7 seconds or more) and the percentage of engagement is high (40% or more), there is little difference between real-time and offline one-annotator classifiers. We would expect the offline annotations to be of higher quality, with more accurate boundaries than the real-time labels, because real-time labeling is challenging given the differing speeds, frequency, and consistency of stereotypical motor movements. A surprising result is that the performance of the classifier using real-time annotations in the lab for participants 9, 10, and 11 outperformed the offline one-annotator classifier. Considering the data of participant 11 (see Figure 2), the highest frequency of offline disagreements occurred around the boundaries of the stereotypical motor movement, particularly in the area that separates the real-time and offline onsets (61% for participant 11). The lowest frequency of offline disagreements occurred in the area that separates the real-time and offline offsets (7% for participant 11). The rest of the disagreements were scattered within an episode (22%) or occurred in isolation (10%). As a result, the real-time training data included approximately 29% of the examples where the offline annotators disagreed, whereas the offline annotation included all the examples with disagreement. This may partially explain why the real-time classifier outperformed the offline one-annotator classifier on participants with longer episodes.

Our final observation is that using agreement data from two annotators is only useful when the duration of stereotypical motor movement is long. Table 6 shows that for all participants with episodes of long duration (9, 10, and 11), training the classifier on agreement data results in the best performance in both the classroom and the lab. The longer duration appears to allow for higher agreement between annotators and better quality training examples.

Participant ID	Accuracy	TP	FP	Precision	Recall
6	74.3%	0.52	0.13	0.48	0.52
7	77.1%	0.53	0.23	0.61	0.53
8	72.9%	0.48	0.15	0.58	0.48
9	82.3%	0.60	0.19	0.62	0.60
10	73.0%	0.45	0.14	0.45	0.45
11	83.1%	0.67	0.19	0.61	0.67
Mean	77.1%	0.54	0.17	0.56	0.54

Table 7. Performance across different participants

It also appears that offline annotation facilitates labeling subtle and transitive variations on stereotypical motor movements, and that with no way to model the uncertainty of the annotator, the algorithm overemphasizes examples that are not particularly good for training. These transitive examples are likely to be missed in real-time annotation and thus are not included in training. For stereotypical motor movements of long duration, missing noisy transitive examples in real-time annotation might improve accuracy. For stereotypical motor movements of short duration, real-time annotation seemingly misses both noisy transitive examples and good examples of short duration (e.g., 33% of the episodes were missed for participant 8). In this case, the increase in performance due to loss of noisy transitive examples did not offset the reduction in performance due to loss of good but short examples.

Experiment 5: Performance across Different Participants

In this experiment, we trained the classifier with data from all the participants but one and tested the performance on the left-out participant. This procedure was repeated across all participants and results were averaged across activities.

The overall performance is relatively low with an average TP rate of 0.54 and an average FP rate of 0.17. The FP rate is dominated by errors associated with the unknown class. There was considerable variability across participants with respect to topography, duration, frequency, and consistency of the movements. This results in overall low performance using participant-independent training. For example, the duration of stereotypical flapping episodes varied from 1 second to several minutes and involved different hand postures and movements across participants.

We also found that some stereotypical motor movements were more consistent across some participants than others. For example, we observed that body rocking is more consistent than hand flapping. Table 7 shows that the best performance on participant-independent training is for participants 9 (Precision 0.62 and Recall 0.60) and 11 (Precision 0.61 and Recall 0.67). Unlike other participants, both engaged primarily in body rocking (82% and 95% of the time they were observed having stereotypical motor movements, respectively). Because body rocking is more consistent across participants (i.e., less variability in how body rocking is performed), the results for these two participants were higher than other cases.

DISCUSSION

To the best of our knowledge, this is the only study on real data from children with stereotypical motor movements from multiple settings and with varying degrees of complexity. Enabling detailed and precise information on the occurrence,

type of movement, and duration of stereotypical motor movements using a system children can easily wear in everyday settings, and that can reliably and automatically recognize these behaviors, could be used for new behavioral and medical research to: (1) Clarify what setting events are associated with stereotypical motor movements; (2) Determine the functional significance of stereotypical motor movements, not only to shed light on the mechanisms that maintain it, but also to determine appropriate treatments; (3) Assess the effects of behavioral and pharmacological interventions intended to decrease the incidence or severity of stereotypical motor movements; and (4) Facilitate more precise intervention efforts before stereotypical motor movements are entrenched in an individual's repertoire.

The problem of accurately recognizing stereotypical motor movements in children with ASD and creating a real-time monitoring tool is more challenging than it may appear at first due to the complexity of the domain. First, there was considerable variability in the topography, duration, frequency, and consistency with which participants performed their stereotypical motor movements. Each child had very specific stereotypical motor movements that required participant-dependent data to train the classifier. Second, both real-time and offline annotations were difficult to generate, even by trained experts. The annotators had more difficulty and disagreement in documenting stereotypical motor movements in real-time than offline. In real-time annotation, the annotators often missed the start and stop times of the activity and sometimes missed the whole activity altogether. In offline annotation, the annotation tool did not account for the uncertainty of the annotator but rather provided discrete markers for the beginning and the end of the stereotypical motor movement. This resulted in noise around the boundaries of each stereotypical motor movement that appears to especially impact recognition of shorter movement segments. Third, during regular school hours, it can be difficult to collect enough training data from participants who engage in infrequent stereotypical motor movements. For some of our participants, the data were sparse, with less than 10% of the data representing specific stereotypical motor movements. Gathering naturally evoked, longer-term data outside of school settings (home, community, etc.) with a mobile system would provide more training examples that might improve recognition.

A key challenge in this domain versus other activity recognition domains is the problem of acquiring adequate participant-dependent training data in real-life situations without an undue burden on the child, a researcher, or caregiver. Training a classifier on agreement data from two annotators produced the best results when the duration of the stereotypical motor movements were long, but acquiring such annotation is unrealistic for all but highly controlled research settings. In this case, most of the examples of transitive behavior were eliminated from the training data by virtue of the disagreement between the annotators. A more realistic deployment scenario would involve the caregiver utilizing a real-time annotation tool on a mobile device that records naturally evoked stereotypical motor movements and captures uncertainty in the annotations.

An encouraging and somewhat surprising result is that classifiers trained on real-time annotations performed slightly better than classifiers trained using offline annotations from one annotator for participants with stereotypical motor movements of long duration (7 seconds or more) who engaged in the behavior frequently. This suggests that transitive and subtle examples typically missed in real-time annotation are not particularly good for training. However, real-time annotation of stereotypical motor movements of short duration misses a significant number of valid episodes that results in worse performance than a single offline annotator. For children with episodes of long duration, our results suggest that deploying a real-time system that acquires annotations in real-time may be feasible. This might be accomplished by providing a teacher/caregiver with a mobile device that facilitates real-time annotation and training. However, to deploy such a system for children with episodes of short duration, research efforts are still needed to: (1) Improve the accuracy of real-time annotation, for example using auto segmenting techniques; and (2) Capture the uncertainty in the annotations particularly on the boundaries of stereotypical motor movements of short duration.

The average performance of the classifier on the stereotypical motor movements in a naturalistic classroom was (Accuracy: 88.6%; TP: 0.85; FP: 0.08) and (Accuracy: 89.5%; TP: 0.84; FP: 0.07) in the lab. Since the classroom is a less-constrained environment than the lab, enabling participants to engage in a wider range of movements (i.e., they have more opportunities to interact with objects in the environment), we expected more movement variability in the classroom and thus better training data, but that turned out not to be true. This result is encouraging in two respects. It indicates that recording data in more constrained environments may still capture important characteristics of the stereotypical motor movements with insignificant impact on the performance of the classifier. This also facilitates the annotation task since annotators found it easier to annotate data in constrained setups such as the lab where it was easier to observe the participants.

Although these results were compiled offline, the entire system has recently been implemented on a mobile device, and activity detection runs in real-time (Windows Mobile, 528 MHz ARM CPU). It should be possible to improve the results presented here using temporal filtering techniques to further lower the FP rates, which would be beneficial for intervention efforts. There are other supervised algorithms based on graphical models, and unsupervised learning algorithms, that could be explored as well.

Finally, it is important to emphasize that automatic, real-time detection of stereotypical motor movements in children with ASD using comfortable, miniature wireless sensors could both advance autism research and enable new intervention

tools that help children and their caregivers monitor and cope with these behaviors. For research, our system has the potential to overcome many of the problems Sprague and Newell [10] associate with direct observational methods mentioned in the introduction. Specifically, using acceleration data, pattern recognition algorithms can accurately document high-speed motor sequences; indicate when a sequence has started and ended; and handle concomitantly occurring stereotypical motor movements. Automating stereotypical motor movement detection in this way could free up a human observer to concentrate on and note environmental antecedents and consequences necessary to determine what functional relations exist for this perplexing class of behavior. For intervention, mobile classifiers could be integrated into a real-time intervention system where real-time training data are provided by caregivers and feedback is provided to participants when stereotypical motor movements are detected. Our present results suggest that such a system would require child-specific training data, and that the training data could be acquired from a non-laboratory setting. It also appears that real-time annotation is possible, especially for children who engage in stereotypical motor movements of long duration. For children who engage in stereotypical motor movements of short duration, it might be possible to enhance the performance by deemphasizing boundary examples and by accounting for uncertainty in annotations. Such a system could facilitate efficacy studies of behavioral and pharmacological interventions intended to decrease the incidence or severity of stereotypical motor movements.

ACKNOWLEDGMENTS

The authors thank Autism Speaks and the Nancy Lurie Marks Family Foundation for funding this work. We also thank the children and parents who graciously agreed to participate in this research. The sensors used in this work were developed with funding from NSF grant #0313065.

REFERENCES

[1] U.S. CDC, "Prevalence of Autism Spectrum Disorders -- Autism and Developmental Disabilities Monitoring Network," 2002.

[2] A. P. A., *Diagnostic and Statistical Manual on Mental Disorders, 4th Ed.,* vol. IV-TR. Washington, DC: Amer. Psychiatric Publishing, 2000.

[3] A. Baumeister and R. Forehand, "Stereotyped acts," in *Int'l. Rev. of Res. in Mental Retardation: VI.*, Ed. New York, 1973, pp. 55-96.

[4] M. H. Lewis and J. W. Bodfish, "Repetitive behavior disorders in autism," *Mental Retardation and Devel. Disabilities Res. Rev.,* vol. 4, pp. 80-89, 1998.

[5] S. J. LaGrow and A. C. Repp, "Stereotypic responding: A review of intervention research," *Amer. J. of Mental Deficiency,* vol. 88, pp. 595-609, 1984.

[6] G. Berkson and R. K. Davenport Jr, "Stereotyped movements of mental defectives. I. Initial survey," *Amer. J. of Mental Deficiency,* vol. 66, pp. 849-52, 1962.

[7] O. I. Lovaas, A. Litrownik, and R. Mann, "Response latencies to auditory stimuli in autistic children engaged in self-stimulatory behavior," *Behavior Res. and Therapy,* vol. 9, pp. 39-49, 1971.

[8] R. S. Jones, D. Wint, and N. C. Ellis, "The social effects of stereotyped behaviour," *J. of Mental Deficiency Res.,* vol. 34, pp. 261-8, 1990.

[9] C. H. Kennedy, "Evolution of stereotypy into self-injury," in *Self-Injurious Behavior: Gene-Brain-Behavior Relationships.*, Washington, DC: Amer. Psych. Assoc., 2002, pp. 133-143.

[10] R. L. Sprague and K. M. E. Newell, *Stereotyped Movements: Brain and Behavior Relationships*. Washington, DC: Amer. Psych. Assoc., 1996.

[11] D. A. Pyles, M. M. Riordan, and J. S. Bailey, "The stereotypy analysis: An instrument for examining environmental variables associated with differential rates of stereotypic behavior," *Res. in Developmental Disabilities,* vol. 18, pp. 11-38, 1997.

[12] J. A. Kientz, G. R. Hayes, T. L. Westeyn, T. Starner, and G. D. Abowd, "Pervasive computing and autism: Assisting caregivers of children with special needs," *Pervasive Computing,* vol. Jan-Mar, pp. 28-35, 2007.

[13] T. Westeyn, K. Vadas, X. Bian, T. Starner, and G. D. Abowd, "Recognizing mimicked autistic self–stimulatory behaviors using HMMs," in *Proc. of ISWC* , 2005, pp. 164-169.

[14] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proc. of PERVASIVE*, 2004, pp. 1-17.

[15] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A hybrid discriminative/generative approach for modeling human activities," in *Proc. of IJCAI*, 2005, pp. 766 - 722.

[16] P. Lukowicz, J. A. Ward, H. Junker, M. Stager, G. Troster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphone and accelerometers," in *Proc.of PerCom*, 2004, pp. 18-32.

[17] J. W. Bodfish, F. J. Symons, D. E. Parker, and M. H. Lewis, "Varieties of repetitive behavior in autism: Comparisons to mental retardation," *J. of Autism and Developmental Disorders,* vol. 30, pp. 237-243, 2000.

[18] E. Munguia Tapia, S. S. Intille, L. Lopez, and K. Larson, "The design of a portable kit of wireless sensors for naturalistic data collection," in *Proc. of PERVASIVE*, 2006, pp. 117-134.

[19] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann, 1999.