# Automatic Assessment of Problem Behavior in Individuals with Developmental Disabilities

**Thomas Plötz**[1], **Nils Y. Hammerla**[1], **Agata Rozga**[2], **Andrea Reavis**[3], **Nathan Call**[3,4], **Gregory D. Abowd**[2]

[1] Culture Lab
School of Computing Science
Newcastle University, UK

[2] School of Interactive Computing
Georgia Institute of Technology
Atlanta, Georgia, USA

[3] Marcus Autism Center
Atlanta, Georgia, USA

[4] School of Medicine, Emory University
Atlanta, Georgia, USA

## ABSTRACT

Severe behavior problems of children with developmental disabilities often require intervention by specialists. These specialists rely on direct observation of the behavior, usually in a controlled clinical environment. In this paper, we present a technique for using on-body accelerometers to assist in automated classification of problem behavior during such direct observation. Using simulated data of episodes of severe behavior acted out by trained specialists, we demonstrate how machine learning techniques can be used to segment relevant behavioral episodes from a continuous sensor stream and to classify them into distinct categories of severe behavior (aggression, disruption, and self-injury). We further validate our approach by demonstrating it produces no false positives when applied to a publicly accessible dataset of activities of daily living. Finally, we show promising classification results when our sensing and analysis system is applied to data from a real assessment session conducted with a child exhibiting problem behaviors.

### Author Keywords

problem behavior assessment, developmental disabilities, autism, mobile sensing, activity recognition

### ACM Classification Keywords

H.1.2 User/Machine Systems I.5 Pattern Recognition: J.4 Social and Behavioral Sciences

### General Terms

Algorithms, Design, Experimentation, Measurement

## INTRODUCTION

Many individuals with developmental disabilities, including those on the autism spectrum, engage in problem behaviors [8, 14, 19]. Behavior problems, such as temper tantrums, destructive behaviors, aggression toward others, and self-injury, are part of the clinical description of autism [4, 31]. Beyond the potential for harm or injury to the individual or those nearby, negative consequences of these behaviors extend to many aspects of the individual's life. They disrupt family functioning and increase caregiver stress and anxiety [15, 19], interfere with learning and socialization [16], and negatively impact long-term prognosis [17]. Thus, many treatment programs have been developed to reduce the frequency and severity of problem behaviors in children with developmental disabilities [20, 21]. While treatments themselves differ in approach, all require the collection of accurate data on the frequency and severity of problem behaviors to understand why and when they occur and to determine if there is a change in the behavior as a result of treatment.

The two main methods for measuring problem behaviors include standardized, validated parent- or teacher-report checklists [1, 3, 27], and direct observations [10, 13]. The former provide quick and cost-effective means of gathering data and are widely used in research settings. However, they do not capture precise frequencies of occurrence of the behavior. Thus, the standard procedure for measuring problem behavior in clinical settings consists of having an observer track and record the frequency of the behavior based on precise pre-determined definitions. While such observations yield rich data regarding frequency and context of problem behavior, there are drawbacks. Definitions of problem behaviors can be subjective and somewhat arbitrary, requiring extensive training and reliability assessments. In addition, certain behavior types can be especially difficult to recognize based on what they look like, while others are difficult to track accurately and objectively. There is no way to objectively assess the intensity of a behavior by human observation alone, even though this is the very characteristic of behavior that may improve with treatment. Finally, direct observation is time intensive and expensive to conduct, and thus, can only be employed to gather small samples of behavior.

Accurate assessment of problem behavior is both key to successful treatment planning and evaluation and the main drawback of current methods of manual observation and track-


*UbiComp '12*, Sep 5-Sep 8, 2012, Pittsburgh, USA.

ing. Therefore, our goal is to explore how technology and computational analysis, i.e., activity recognition using body-worn sensors, can support the gathering of objective, accurate measures of the frequency of problem behaviors. Direct sensing and assessment has the potential for enhancing current clinical practice by providing analysis that is more objective and consistent, and less expensive and time intensive than manual assessments. The complexity of problem behavior and the large variance in its manifestations implies non-trivial challenges to sensor data analysis. The same holds for the design of a safe, robust, and reliable sensing system for a vulnerable population. This paper describes the first system of its kind, which replicates experts' assessments of problem behavior in clinical settings. As such it represents the first milestone towards our ultimate goal of developing a sensing and analysis system for continuous unsupervised behavior assessment in everyday life situations.

We observed current practice at a treatment clinic where behavior is assessed for the frequency of aggression (directed at others), disruption (directed at the environment), and self-injury (directed at self). Based on these observations we designed and developed a sensing system based on tri-axial accelerometers worn on the individual's limbs. Computational analysis is based on unsupervised segmentation of sensor data streams into behavior episodes that are then classified using an activity recognition system based on a novel, problem-specific feature representation capturing energy characteristics and sensor orientations, and statistical classifiers.

We rigorously tested the developed system in three sets of practical experiments. First, we evaluated its sensitivity by analyzing a large dataset of simulated assessment sessions where experienced staff members at the clinic engaged in typical problem behaviors while wearing the sensing system. The automatic analysis detected severe behavior episodes with a precision of $> 95\%$ (recall: $41.5\%$) and an average accuracy of approximately $80\%$ for differentiating among aggression, disruption, self-injury, and movements unrelated to problem behavior. Second, we evaluated the system on a standard activity recognition dataset (OPPORTUNITY challenge [26]), which contains data recorded using a comparable sensing system and covering activities of daily living (ADL) that —by definition— do not include problem behavior episodes. Our system achieved a negligible number of false positive predictions. Third, we evaluated our system in a real clinical assessment session with an autistic child who occasionally engages in problem behavior. Our automatic analysis largely replicates the results of expert assessment.

## CLINICAL ASSESSMENT OF PROBLEM BEHAVIOR – CURRENT PRACTICE IN BEHAVIOR CLINICS

The work presented in this paper was conducted in close collaboration with a local behavior treatment clinic. In the following section we describe the clinic's behavior assessment practices, which are representative of typical procedures in such facilities and thus form the foundation for our research.

When individuals with developmental disabilities engage in problem behaviors, caregivers typically seek professional help to address these behaviors. The first step is to objectively assess the frequency and severity of the problem behavior, its topology (characteristics), and to understand its causes and functions so that an appropriate, targeted treatment plan can be devised. Once treatment commences, there is a need to gather data to determine whether the child is responding to the treatment. Upon treatment completion it is common practice to follow-up with the family to ensure that treatment gains are being maintained and generalize to the child's everyday life. The key variable underlying this entire process consists of expert assessments of frequency and topology of the target behavior, rooted in the gold-standard practice of direct observation [10].

### Functional Behavioral Assessment

The key variable in matching interventions to individuals and their particular problem behavior is the *function* of that problem behavior [13, 24, 29]. Function refers to the antecedent variables —both internal to the individual and external in the environment— that evoke and the consequences that maintain the behavior. Common functions for problem behavior include desire for caregiver attention, access to preferred items, or escape from/avoidance of demands to engage in non-preferred activities.

At the outset of treatment, identifying the function of an individual's problem behavior is often accomplished using so-called *functional behavioral assessment* (FBA [12, 18]). During this procedure, the individual is observed in test conditions in which potential antecedents of problem behavior are introduced (e.g., attention is withheld). Rates of problem behavior that occur during these conditions are compared to control conditions that don't contain variables that might evoke problem behavior (i.e., child is provided with attention). The function of problem behavior is determined by identifying those test conditions in which the rate of problem behavior is elevated relative to the control condition.

An FBA is usually conducted within specialized clinic facilities and with highly trained staff, both of which are necessary to collect the requisite observational data and ensure the safety of all involved. Current practice consists of sessions conducted within treatment rooms equipped with one-way mirrors, microphones, and cameras to allow unobtrusive data collection. One staff member remains in the room with the child in order to administer the various test conditions, while another observes from an adjacent room through a one-way mirror and flags occurrences of target behaviors. The latter are operationally defined to allow for consistent scoring (Table 1). A second observer annotates $\geq 20\%$ of sessions for inter-observer agreement calculation.

### Tracking Treatment Progress and Outcome

Once functions of targeted problem behaviors are identified and treatment begins, there is a need for ongoing data collection to monitor the child's progress and, as needed, to make the necessary adjustments to the treatment plan. Tracking of the occurrence of problem behaviors typically takes place during therapy sessions, following the aforementioned assessment procedure. Once a child has completed treatment, follow-up services with families are conducted to determine whether treatment gains are being maintained. These follow-

| Behavior | Operational Definition |
|---|---|
| Aggression (AGG) | **Biting:** top and bottom teeth come into contact with any part of a person's body<br>**Grabbing:** squeezes/pinches/grabs person's body part/clothing with one/both hands<br>**Hair Pulling:** grabbing another person's hair with one or both hands resulting in moving the person's head from its original position<br>**Hitting:** hands/forearms contact any part of a person's body from distance of $\geq$ 6"<br>**Object AGG:** throwing object within 2 feet of a person from distance of $\geq$ 6"<br>**Kicking:** foot/leg contacts any part of a person's body from distance of $\geq$ 6"<br>**Pushing:** forcefully moving a person from their original location using one/both hands |
| Self-Injurious Behavior (self-injury) | **Self-Biting:** jaw opens and teeth come into contact with any part of body<br>**Body Slapping/hitting:** hits/slaps any part of his/her body with an open palm or closed fist from a distance of $\geq$ 6"<br>**Face slapping:** slaps face with and open palm from a distance $\geq$ 6"<br>**Head banging:** head forcefully comes into contact with the ground or any other hard surface from a distance of $\geq$ 6"<br>**Head Hitting:** hits head with open/closed fist or with object from distance of $\geq$ 6"<br>**Self Kicking:** foot contacts another part of body from distance of $\geq$ 3 " |
| Disruption | **Body Slamming:** runs into objects from 6 " or greater<br>**Furniture:** tipping furniture 45 degrees from its original position<br>**General:** hands/feet/body come into contact with floor/wall/object from $\geq$ 6 "<br>**Object Disruptions:** pushing or swiping objects from surfaces or throwing an object not within 2 feet of a person<br>**Property Destruction:** rips or tears an object |

**Table 1. Operational definitions used for assessment of problem behavior.**

up services are provided in the families' homes and communities post discharge. During these visits, a therapist observes and records data on caregiver implementation of the treatment components and on problem behavior using paper and pencil methods. If needed, additional training is provided in the form of didactic instruction, modeling, rehearsal, and performance feedback.

## AUTOMATIC ASSESSMENT OF PROBLEM BEHAVIOR

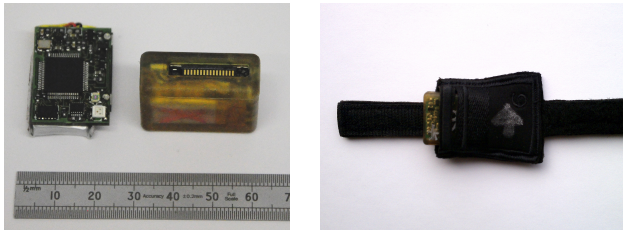
Logging and evaluating frequency of occurrence of specific problem behaviors is central to assessing whether treatment strategies are effective, and whether treatment gains generalize outside of the clinic. We have identified key challenges with current data collection and analysis methods that we believe Ubicomp systems are uniquely positioned to address:

1. Relying on direct observation to gather data during treatment sessions places a strain on staffing.
2. The need for a high level of agreement between observers necessitates precise definitions of problem behavior that can be subjective and somewhat arbitrary, such as the need to define distance metrics to help observers agree on what constitutes sufficient movement for a hit or kick.
3. Precise measurement is problematic for behaviors that occur at a very high rate (e.g., 1Hz) or at a very low rate (e.g., 1 per week), or for covertly occurring behaviors.
4. Relying on parent-reports may present an inaccurate picture of the extent to which treatment gains generalize to the child's home and school.

We developed an assessment system consisting of on-body sensing (Figure 1) and automatic analysis, which has the potential to replicate and augment current clinical assessment practices to yield more accurate, objective, and reliable measurement of frequency and typology of problem behaviors.

### Wearable Sensing System

Problem behavior (Table 1) is typically linked to intensive and characteristic physical movements by the individual engaging in the behavior. Thus, our methodology is based on direct recordings of movements using wearable sensors.



(a) data logger (size in mm) (courtesy of axivity.com)

(b) sensor strap with attached data logger

**Figure 1. Sensing system consisting of tri-axial accelerometers (left: data loggers), and straps for sensor placement on limbs (right)**

The application scenario of recording potentially aggressive and disruptive behavior of vulnerable individuals places specific constraints on a wearable sensing system. Robustness and durability obviously represent major constraints. Furthermore, the system should be designed in a way that maximizes the likelihood of being tolerated by the potential wearer, not to mention safety issues that require effective elimination of potential injuries. In order to capture as much detail on behavior as possible, the use of a single sensing system is inappropriate. Even when optimizing on-body placement of a single data logger [5], chances are high that certain types of problem behavior will be missed. Finally, continuous operation over multiple days needs to be ensured for integration into everyday routine with sporadic clinic consultation only.

*Sensor Straps* Our sensing system is based on four small data loggers that continuously record tri-axial acceleration signals (see below). In order to attach the devices, we designed straps for wrists and ankles that effectively keep the sensors in place even during rough treatment. The straps are made of hypoallergenic and robust fabric with attached Velcro® locks designed to obstruct one-handed removal. After fastening the straps, the sensors, which are housed in a small pocket in the strap, are secured and kept in place with fixed orientation. All borders of the straps are finished with a seam made of extra-strong yarn to ensure durability. The straps are black and very thin (less than one-inch wide).

*Data Loggers* The sensing system used for capturing behavior data is based on Axivity AX3 data loggers, each consisting of a 16bit micro-controller, a micro-electro mechani-

cal systems tri-axial accelerometer and a large block, single layer chip NAND flash [6]. Also included are ambient light and temperature sensors (not used in this work), and a real time clock, which is stabilized by a 20ppm oscillator. The device is powered by a rechargeable Lithium-Polymer battery. It is hermetically encapsulated in a tough macromelt polymer, which is shock-proof, food safe, wipe clear and sterilizable using alcohol. Following a full charge the device can log continuous data from all sensors at a rate of 100Hz for a period of 15 days (approx. one week for 200Hz). We chose AX3 data loggers especially due to their robustness and durability, which is necessary for the potentially rough treatment of the devices if assessed individuals actually engage in problem behavior.

**Computational Behavior Assessment: System Overview**
Figure 2 gives an overview of the analysis system for problem behavior assessment. The sensors attached to the limbs continuously record tri-axial acceleration data and store it to on-board memory (*recording*). To allow for discrimination between different kinds of complex behaviors and for identifying the exact moment of impact, we sampled with a rather high sampling rate of approx. 200Hz within a range of $\pm 16$g. Manufacturing tolerances of the data loggers result in differences in the absolute sampling rates of the particular sensors involved. Furthermore, over time, inevitable sensor drifts have to be compensated. Such drifts, caused, for example, by temperature or humidity differences, slightly change the effective sampling rate of the data loggers. To ensure constant and identical sampling rates for all sensors used over the analyzed recording period, all data are resampled to a fixed rate of 100Hz using cubic interpolation.

Recorded sensor streams are then analyzed for behavior episodes (*segmentation*; behavior episodes are underlined in red in Figure 2). Feature representations of these automatically extracted segments are fed into a statistical classification system, which discriminates among behavior episodes of aggression, disruption, self-injury, and other (*classification*).

**Detection of Behavior Episodes – Segmentation**
The assessment system will be used for the analysis of large amounts of sensor data recorded in sessions of considerable length. In order to effectively process these streams of sensor readings we employ an explicit, lightweight segmentation procedure that identifies behavior episodes before classifying them regarding their type. Behavior episodes represent human activities that are defined by continuous movements resulting in sufficiently large sensor displacements and orientation changes, and include both the problem behaviors we are interested in measuring as well as "regular" activities like a stride or a hand wave. The goal of the segmentation step is to highlight these building blocks of human behavior by filtering the input data and reducing it to segments that can then be analyzed in more detail in the next step of the recognition pipeline. The key idea in our segmentation is to first identify certain characteristic points within the continuous sensor data streams. Based on these *seed points* we identify the boundaries of the surrounding behavior episode in order to capture not only the impact but also to include characteristic motions before and after the specific behavior.

As most problem behaviors involve high amplitude movements (e.g., punch or kick), a main criterion for segmentation are peaks in the short-term signal energy. However, some disruptive events, like tipping over furniture, are mainly composed of characteristic changes in limb inclination that may be missed if energy was the only criteria used. Therefore, our segmentation procedure additionally considers limb inclination changes in terms of relative sensor orientation changes. By abstracting from absolute values it becomes robust with respect to factors such as sensor displacement. Based on the spherical representation of the acceleration signals $\mathbf{x}^{\mathcal{S}} \in \mathbf{R}^3_{\mathcal{S}(\sigma,\phi,\mu)}$ —$\sigma, \phi, \mu$ denote radial distance, inclination, and azimuth— we calculate short term energy $E_1$ and magnitude of orientation change $E_2$ (with $\Delta \sin\{\sigma, \phi\}$ as first derivatives of spherical angles $\sigma$ and $\phi$) using a sliding window procedure. The weighted sum $\mathcal{E}$ of both components serves as a 1D representation, covering signal energy and sensor orientation changes in a compact way:

$$E_1 = 1/N \sum_{i=1}^{N} (x_i^s)^2 \quad (1)$$

$$E_2 = 1/N \sum_{i=1}^{N} \gamma_i \quad (2)$$

$$\text{with} \quad \gamma_i = \sqrt{(\Delta \sin \sigma_i)^2 + (\Delta \sin \phi_i)^2} \quad (3)$$

$$\mathcal{E} = \alpha E_1 + \beta E_2 \quad (4)$$

Weights can be derived in cross-validation experiments or explicitly set to incorporate prior knowledge to personalize the procedure (e.g., for slim vs. more corpulent individuals). For our experiments we set $\alpha = 1.5$, $\beta = 1$, and $N = 32$ as the frame-length for the sliding window procedure.

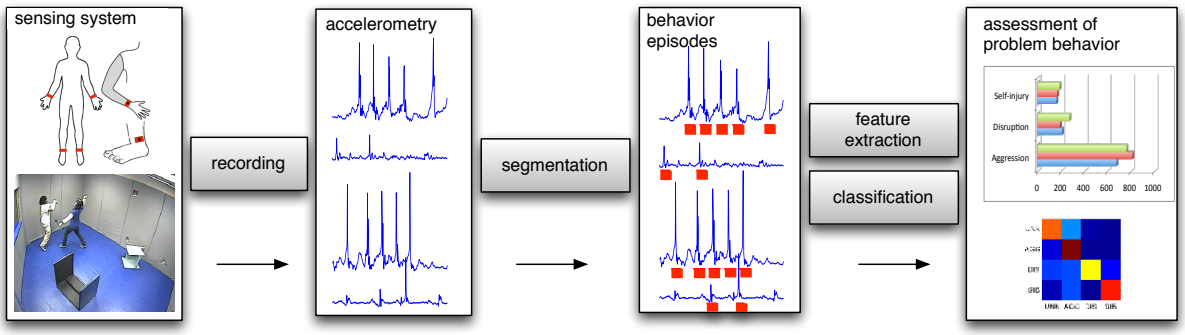



**Figure 2. Analysis of problem behavior based on tri-axial acceleration data – system overview (see text for description; best viewed in color)**

Local maxima in the $\mathcal{E}$-representation of the input data are used as seed points. For peak detection we utilize a hysteresis approach with data-driven threshold estimation. Starting from a particular seed point the surrounding segment is extracted by aggregating adjacent samples until the lower cut-off point is breached. Since the $\mathcal{E}$-representation encodes both energy and orientation change information in a combined signal, this aggregation is very effective. The $\mathcal{E}$-magnitude of energy maxima typically exceeds those of orientation changes by far. Thus, seed points usually correspond to energy maxima, e.g., the moment of impact during a kick. These events are surrounded by orientation changes, i.e., foot approaching before the actual kick and moving back afterwards. Consequently, local minima in the vicinity of seed points represent the boundaries of behavior episodes. Imperfect peak detection combined with this aggregation often results in the generation of segment duplicates, which are reduced to a unique set using straightforward post-processing.

### Feature Extraction

The main criteria for the design of a feature representation of the acceleration input data as it is fed into the subsequent statistical classifier are: *(i)* independence of the resulting representation on the length of the analyzed signals (since we are avoiding explicit sequence models; see below); and *(ii)* the need to capture *characteristic* differences between the activity classes of interest. Especially in the case of the latter and in the light of the target application domain, it is worth reconsidering what kind of differences an automatic analysis system would need to deal with. For example, with *aggression*, the majority of activities correspond to the person hitting someone else. The "target" of the aggressive act typically reacts and may deflect or block the hitting limb. Kicking is usually accompanied by orientation changes for the sensors attached to the feet. In contrast to the rather soft target of aggressive behavior (i.e., human body), *disruption* is directed towards more rigid objects like furniture. The recorded signals show characteristic shapes and/or oscillations after impact. In the case of *self-injurious behavior*, the actor and target are the same individual, and this typically results in more forceful impact as the "attacker" deliberately does not deflect or move back. Consequently, these events show the highest absolute energy of all problem behaviors along with unique changes in limb inclination (e.g. hitting the head). Figure 3 shows examples of all three classes of problem behavior and the corresponding raw acceleration data (magnitude) recorded by the body-worn sensors.

Based on these constraints and observations, we calculate features that cover: *i)* spectral characteristics of acceleration signals; *ii)* orientation change statistics; and *iii)* explicitly integrate signal energy that is normalized regarding segment length (Algorithm 1). Features are calculated for every detected segment, i.e., behavior episode, and separately for each sensor. First, $f = 33$ Fourier descriptors $\{\mathcal{F}_i^c(s)|i = 1 \dots f, c = \{x, y, z\}\}$ are calculated for every segment $s$ and per channel $c$ of the acceleration signal $\mathbf{x} \in \mathbf{R}^3$. The actual choice of $f$ has been determined in cross-validation experiments (results not shown). The second set of features consists of a probabilistic representation of the orientation changes within the analysis window. We calculate the empirical cumulative density function (ECDF, see [25] for ECDF-based representations in activity recognition) of the $E_2$ signal (Equation 2) and integrate the first 20 coefficients, i.e., a compact yet meaningful approximation of the ECDF, into our feature representation. Finally, the segment's energy, normalized by its duration, is added resulting in $D = 120$-dimensional feature vectors per segment and sensor.

**Algorithm 1** Feature extraction (segment-wise)

**Input:** accelerations $\mathbf{x} \in \mathbf{R}^{3\times l}$ for segment $s$ ($l$ = segment length), and orientation change signal $\boldsymbol{\gamma}$ (Equation 3); $f$ = #Fourier coeff.; $n$ = #ECDF coeff.
**Output:** features $\mathbf{c} \in \mathbf{R}^D$ for $s$; $D = f \times 3 + n + 1$
$\{\mathcal{F}_i^c(s)|i = 1 \dots f, c = \{x, y, z\}\} = \texttt{calcFTDesc}(\mathbf{x})$
$\mathbf{O}$ = $\texttt{calcECDF}(\boldsymbol{\gamma}, n)$ {calculate first $n$ coefficients of ECDF representation of orientation changes in segment}
$NRJ(\mathbf{x}) = 1/l \sum_{i=1}^{l} \sum_{j=\{x,y,z\}} x_{i,j}^2$
$\mathbf{c} = (\{\mathcal{F}^c|c = \{x, y, z\}\} \quad \mathbf{O} \quad NRJ)^T$
**return** $\mathbf{c}$

The discrimination of problem behavior is based on statistical classifiers. In order to allow for robust parameter estimation of the classification system, our feature extraction process is finalized by means of PCA-based dimensionality and de-correlation. Based on the analysis of the Eigenvalue spectrum of a cross-validation dataset, we project the $D = 120$-dimensional feature vectors onto a lower-dimensional subspace, which captures 95% of the feature space variance.

### Fine-Grained Classification of Problem Behavior

Behavior episodes extracted from recorded sensor signals represent potential candidates for problem behaviors of interest. Two key aspects are of interest for the envisioned applications. First, how many instances of a problem behavior occur over the time of system deployment? Second, what types of different problem behavior occur? The classification step of our recognition pipeline helps answering these questions for those types of problem behavior that can be captured by our sensing system.

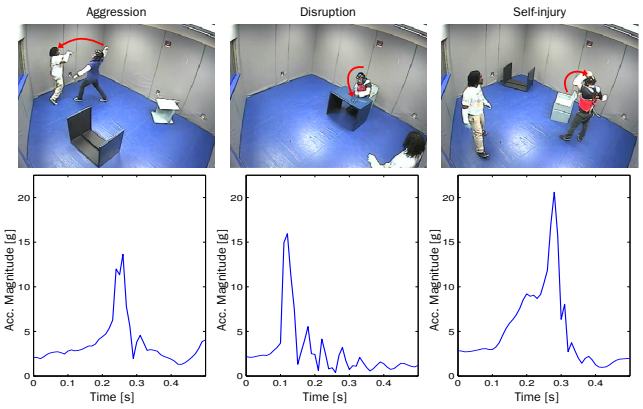



**Figure 3. Examples of problem behavior (top) and their manifestation in raw sensor data (lower row: magnitudes of 3D acceleration signals). Recordings from simulation sessions with staff-members engaging in typical problem behavior (wearing protective gear).**

Feature extraction produces a compact and meaningful representation of behavior episodes with fixed dimensionality. As the characteristic differences between different types of behavior can be small, plain distance-based classification approaches such as KNN are likely to fail. Consequently, we apply more complex statistical modeling methods for the recognition of problem behavior. We explore the effectiveness of the three main types of statistical classifiers that each focus on different aspects in the statistical modeling process [7]: *i)* Naive Bayes (NB) classifier, a rather simple probabilistic example of generative modeling; *ii)* C4.5 decision tree classifier, the standard implementation of predictive modeling; and *iii)* Support Vector Machine (SVM) classifier, the most prominent example of discriminative modeling, which has proven very successful for a number of classification problems especially if only little sample data is available for training. We deliberately did not include explicit sequence models into the evaluation (e.g., hidden Markov models) since they are prone to overfitting if data are analyzed that exhibit high *intra*-class but low *inter*-class variance as is the case for the analyzed behavioral data [9].

## EXPERIMENTAL EVALUATION

The experimental evaluation of any kind of technology designed to assess the behavior of a vulnerable population is challenging. Automatic predictions need to be rigorously validated following established protocols, and based on a solid statistical basis, i.e., a representative and significantly large dataset. However, the collection of such a dataset is hard if solely focusing on recordings of actual clients. Ethical and safety issues are two major obstacles. It is hard to predict if/when an individual will engage in problem behavior, which complicates the recording of such a dataset. We address these challenges utilizing a three-stage experimental evaluation. With this procedure we are in the position to extensively evaluate and validate the developed system.

*Stage 1 (*SIMPROB*)* For system development and validation we recorded a dataset where experienced members of staff of the collaborating behavior clinic simulated assessment sessions in the clinic's facilities. They were asked to engage in a variety of problem behaviors as they experienced them in their clinical practice. This gives us a rich dataset of realistic behaviors that is used for systematic evaluation of our system's segmentation performance and classification accuracy.

*Stage 2 (*ADL*)* Arguably, the high frequency of occurrence of problem behaviors in the SIMPROB dataset is not representative for actual clinical assessment sessions. In order to evaluate the system's precision more realistically we conducted a second set of experiments. We evaluated our system on a standard activity recognition database that covers activities of daily living but —by definition— does not contain any problem behaviors. The success of our automatic assessment system is measured by the false alarm rate, i.e., by the number of falsely predicted problem behaviors.

*Stage 3 (*KID*)* In the third experiment we used the system for a real assessment session in the behavior clinic with an autistic child who engages in problem behavior. We evaluated the system's capabilities for replicating human expert assessments according to current clinical practice in terms of segmentation's recall and overall classification accuracy.

| | left wrist | right wrist | left ankle | right ankle | total |
|---|---|---|---|---|---|
| SIMPROB dataset | | | | | |
| aggression | 311 | 377 | 22 | 50 | 760 |
| disruption | 91 | 132 | 12 | 34 | 269 |
| self-injury | 70 | 114 | 0 | 1 | 185 |
| total | 472 | 623 | 34 | 85 | 1,214 |
| KID dataset | | | | | |
| aggression | 14 | 17 | n/a – child did not | | 31 |
| disruption | 95 | 73 | tolerate sensors | | 168 |
| self-injury | 40 | 86 | on ankles | | 126 |
| total | 149 | 176 | n/a | | 325 |

**Table 2. Summary of datasets recorded for system evaluation (GT).**

### Data Collection and Ground Truth Annotation

SIMPROB We recruited five members of the clinic's therapy staff (2 females, 3 males; all right-handed) to help us run 11 simulated assessment sessions within the clinic's facilities. Participants were asked to take on one of three roles: *i)* the individual who engages in problem behaviors; *ii)* the therapist who is typically in the room with the child and is the target of the child's aggressive behaviors; *iii)* the data collector who watches the assessment through a one-way mirror and records the frequency of behavior according to the operational definitions (Table 1). Actors changed roles to increase the variability of expression of the various problem behaviors. On average, two minutes of sensor data were collected per session for the actor simulating the child. SIMPROB contains a total of $1,214$ problem behaviors (Table 2).

To prevent injuries, actors wore protective gear, including a padded vest, a helmet, and limb-protectors. This equipment is routinely used at the clinic when assessments are conducted with very aggressive individuals. Live-annotation from behind a one-way mirror represents the "best-practice" in data collection at the clinic. The annotator watches the session and notes each time a target problem behavior of interest occurs (time-stamp, type). We cannot assume that this live annotation is accurate as some events might be missed by the annotator and the time-stamp is likely to be inaccurate due to human reaction times. In addition, the live-annotation does not contain information regarding the specific limb involved, which is needed for model training. In order to obtain ground truth (GT) annotation for model training and validation, a trained researcher re-annotated the sessions based on video-footage. She noted the exact moment of impact for each instance of problem behavior, and then categorized its type and the limb involved. Annotation is based on detecting and labeling problem behavior *events*, neglecting their duration, which is standard practice in this clinical assessment.

ADL Arguably SIMPROB contains an artificially high number of problem behaviors. Thus, experiments based on it are ideal for evaluating the precision of our analysis system but recall assessment would be overly optimistic. For a more realistic picture the assessment system also has to be evaluated on "regular," i.e., non problem behavior data. We discarded the idea of extending SIMPROB by letting the actors wear the sensing system outside the simulation sessions. GT annotation was difficult to integrate into clinic routine, and

impossible to obtain outside the clinic for privacy reasons. Instead, we used an alternative dataset for recall evaluation.

Within the OPPORTUNITY project, a major activity recognition dataset was recorded with a focus on activities of daily living – ADL [26]. A total of 72 sensors of 10 modalities, embedded into objects or body-worn, were employed for recording people's morning routine, resulting in a "particularly large number of atomic activities (more than $27,000$), collected in a very rich sensor environment."[26] By definition, the recorded activities do not contain any kind of problem behavior but the complete variety of domestic activities.

We used the OPPORTUNITY challenge task B2 (Multimodal activity recognition: Gestures – test set) for evaluation on $> 1$ hour of "regular", i.e., non problem behavior data. The annotated activities comprise opening and closing kitchen furniture and appliances, cleaning the table, moving objects, and NULL. Since the recorded morning routine has been conducted with no further constraints in a kitchen environment, the NULL class also contains a large variety of "other" activities, including walking, sitting down, standing up, unspecified hand gestures etc. It is imperative that our analysis system not confuses these regular activities with severe problem behavior. We evaluated the recall of our system by applying it "as is" to the ADL dataset. For compatibility with our sensing system we used the acceleration data recorded by the limb-worn inertial measurement units, which represents an identical sensor placement as in our other experiments. We upsampled the ADL data from 30Hz to 100Hz using cubic interpolation. Sensor orientations at every limb were manually transformed to match those of our sensing systems. Absolute accelerations were measured in earth's gravity $g$ in both sensing systems. By means of this mapping procedure we ensured that both signal types are comparable.

KID In the third experiment we used the assessment system for the analysis of a real functional assessment session in the clinic. During this session (length: $> 50$ min.) the child (male, aged 11, weight 63 lbs, right-handed) engaged in $325$ problem behavior episodes (168 disruption, 31 aggression, 126 self injury). Manual annotation was obtained using the same procedure as for the SIMPROB dataset. This experiment directly corresponds to the envisioned clinical application case. It also reflects the practical challenges faced by wearable assessment systems such as ours, as the child only tolerated the sensing system on his upper limbs. Consequently, the evaluation is based on problem behaviors observed for the arms only. This child exhibits problem behavior according to a specific pattern. Over the course of the session he engaged in a variety of behaviors that involved playing with toys and high energy activity such as jumping and running around. The therapist that accompanied the child over the course of the session did not disrupt any severe behavior unless there was imminent danger, such as falling off a chair. Often the severe behavior occurred in batches where multiple events followed closely on each other.

### Results

We report segmentation and classification results separately, and for all three stages of analysis (Table 3). For the SIMPROB and KID datasets we employed 10-fold cross-validation procedure for classifier training and system optimization. The derived system is then used "as is" for the analysis of the ADL dataset, which does not contain problem behavior, and we report absolute numbers of false positive predictions, i.e., erroneous detections of problem behavior episodes. For these false predictions we also provide classification results with respect to the problem behavior classes of interest.

SEGMENTATION The accuracy of segmentation are reported in the upper half of Table 3. For SIMPROB, behavior episodes were detected with an average precision of $41.5\%$ and average recall of $95.4\%$ across all limbs. While high precision values were achieved for the detection of behavior episodes on wrists ($51.3$ and $63.7\%$), the segmentation lacks precision for behavior episodes involving the ankles ($7.9$ and $19.8\%$), which corresponds to over-segmentation. The over-segmentation for episodes involving the ankles stems largely from the abundance of high energy episodes during walking, as each step, particularly when running, produces sharp peaks in the signal energy. Note, however, that such false positives will be addressed in the next step of the analysis, since our classification algorithms will classify these episodes as "unknown" (i.e., not problem behavior related). The human-annotated events that are missed by our segmentation step (false negatives) typically involve low amplitude motions that do not produce a sufficient displacement of the sensors to be detected in the current sensor configuration.

A total of $677$ false positives were produced during segmentation of the ADL dataset. Since this dataset does not contain any actual problem behavior episodes we report absolute numbers. Again the lower limbs were more affected by over-segmentation (no erroneous prediction on arms). For the KID dataset, precision of detecting behavior episodes (involving arms only) is largely comparable to the SIMPROB dataset, though recall drops about $14\%$ to $81.2\%$.

CLASSIFICATION The accuracy of classification of the behavior episodes extracted in the segmentation step are reported in the lower half of Table 3. We evaluated the effectiveness of three types of statistical classifiers: Naive Bayes (NB), Decision Trees (C4.5), and Support Vector Machines (SVM; with RBF kernel). For the latter we optimized the slack $C$ and the kernel parameter $\gamma$ in a grid-search procedure as it is standard for SVM-based applications [28]. SVM-based classification consistently outperformed the other two modeling technique throughout all three tasks.

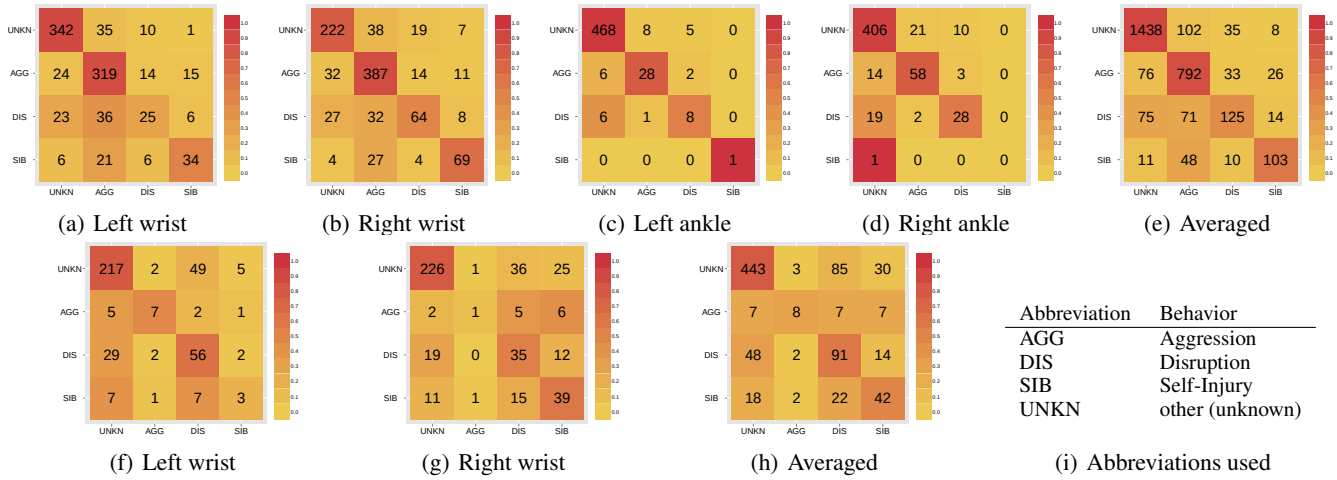
Overall classification accuracy for differentiating among the relevant classes of behavior episodes was, on average, $80.3\%$ for SIMPROB, $99.6\%$ for ADL, and $69.7\%$ for KID. The confusion matrices in Figure 4 (upper row: limb-based and averaged results for SVM-based classification on SIMPROB; lower row: same for KID) indicate that our classification procedure effectively compensated for the over-segmentation effect seen in the first step of the analysis procedure, which resulted in low precisions for detection of behavior episodes.

The averaged confusion matrix for the KID task (Figure 4(h)) shows that we can successfully reject unknown instances and differentiate between disruption and self-injury. The successful modeling of disruptive behavior for this specific child is reasoned in the reduced complexity compared to the

| | SIMPROB | | | ADL | | KID | | |
|---|---|---|---|---|---|---|---|---|
| *Segmentation of behavior episodes (BE)* | | | | | | | | |
| | Precision [%] | Recall [%] | #BE (GT) | #False Positives | | Precision [%] | Recall [%] | #BE (GT) |
| left wrist | 51.3 | 95.1 | 472 | 0 | No problem BE in ADL set | 29.6 | 80.5 | 149 |
| right wrist | 63.7 | 96.3 | 623 | 0 | (by definition); overall | 32.5 | 81.8 | 176 |
| left ankle | 7.9 | 94.1 | 34 | 327 | duration: >1 hour; average | n/a – child did not tolerate | | |
| right ankle | 19.8 | 98.8 | 85 | 350 | length FP: $1.2s(\pm 0.3)$ | sensors on lower limbs | | |
| average | 41.5 | 95.4 | 303.5 | 169.3 | | 31.1 | 81.2 | 162.5 |
| total | | | 1,214 | 677 | | | | 325 |

| *Classification* | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | accuracy [%] | | | | accuracy [%] | | | | accuracy [%] | | | |
| | NB | C4.5 | SVM | #BE (pred.) | NB | C4.5 | SVM | #BE (pred.) | NB | C4.5 | SVM | #BE (pred) |
| left wrist | 68.9 | 65.4 | 78.5 | 917 | n/a – no (false) prediction | | | 0 | 69.6 | 63.5 | 71.6 | 395 |
| right wrist | 63.1 | 56.9 | 77.6 | 956 | of any BE | | | 0 | 63.8 | 57.8 | 69.4 | 434 |
| left ankle | 87.8 | 89.7 | 94.7 | 532 | 29.4 | 98.3 | 99.4 | 350 | n/a – child did not tolerate | | | |
| right ankle | 77.8 | 74.9 | 87.6 | 562 | 60.6 | 59.6 | 99.7 | 327 | sensors on lower limbs | | | |
| average | 69.8 | 66.6 | 80.3 | 741.8 | 44.5 | 76.3 | 99.6 | 169.3 | 65.9 | 59.9 | 69.7 | 414.5 |
| total | | | | 2,967 | | | | 677 | | | | 829 |

**Table 3. Evaluation results for all three tasks. Classification accuracy regarding aggression, disruption, self-injury, and unknown based on automatically extracted segments (as reported in upper half – segmentation).**




(a) Left wrist (b) Right wrist (c) Left ankle (d) Right ankle (e) Averaged

(f) Left wrist (g) Right wrist (h) Averaged (i) Abbreviations used

**Figure 4. Confusion matrices for SVM-based classification of extracted behavior episodes (top: SIMPROB task; bottom: KID task; all matrices row-wise normalized). Absolute numbers may differ from ground truth totals (Table 2) due to false negative predictions in segmentation stage.**

SIMPROB task, as just a few characteristic motions occur (mainly hitting furniture, walls). Aggression on the other hand cannot be identified as reliably, which indicates large variations for this specific category.

## RELATED WORK

The de-facto standard procedure for assessing problem behavior is based either on standardized parent- or teacher-reports [1, 3, 27], or on direct human observation in clinical settings. Although these procedures are widely employed, and represent current best practice in problem behavior assessment, they result in data that is far from optimal. Reasoned by potentially subjective and arbitrary definitions of problem behavior (not to mention their severity), and difficulties in observing and tracking certain kinds of behaviors, an objective and accurate assessment is often difficult to achieve. These drawbacks served as the motivation for the development of the approach presented in this paper.

Few publications exist that address the use of automatic analysis techniques to assess behavior related to developmental disabilities and autism. Most of these papers focus on specific behavioral phenomena rather than assessing a broader range of behaviors. For example, Goodwin and colleagues developed a system for recognizing stereotypical movements (not problem behavior) in individuals with autism [2, 11]. Similar to our work they used wrist-worn accelerometers for recording data on movements of the limbs. By means of a decision tree classifier a frame-wise recognition of two types of stereotypical movements —hand flapping and body rocking— was performed with satisfying accuracy in two different environments (classroom and laboratory). Westeyn et al. described the classification of a range of self-stimulatory behaviors typically observed in individuals with autism using body-worn tri-axial accelerometers and an HMM-based analysis approach [30]. Interestingly, they also let an ac-

tor mimic the behavior that was the target of the analysis. However, the dataset that was collected is very small and the overall procedure of rather exploratory nature. Finally, Min and coworkers also focused on detecting self-stimulatory behavior in individuals on the autism spectrum using on-body sensing [22, 23]. The focus of their work is on exploring the effectiveness of various signal processing techniques.

## DISCUSSION

SUMMARY The goal of this paper was to explore how technology and computational analysis can support the clinical practice of problem behavior assessment in individuals with developmental disabilities. We developed a body-worn sensing system and activity recognition techniques that effectively gather objective measures of the frequency of problem behaviors. Using our system we were able to replicate current manual assessments, i.e., clinical best practice, with high accuracy (Figure 5). This is very promising, especially in light of an extremely challenging application domain. Children with developmental disabilities pose substantial challenges for the wearable sensing system (e.g., tolerance by the wearer, durability) and the analysis algorithms (e.g., substantial variability in manifestations of problem behaviors and their similarity to day-to-day activities).

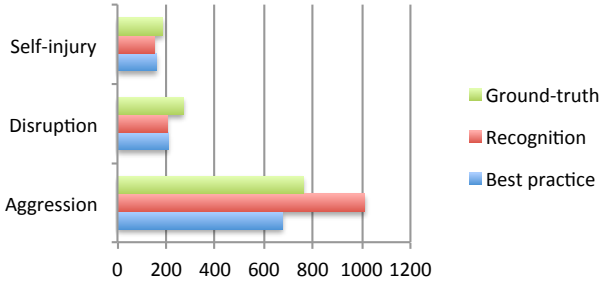



**Figure 5. Comparison of summary reports (count of behavior occurrence) for GT annotation, automated recognition and BP (SIMPROB).**

Our main validation experiment included simulated data performed by trained clinic staff. We focused on simulated data for system development and reliability evaluation because it afforded us the opportunity to maintain strict control of the experiments, which is important at this stage of our exploratory research. The staff were instructed to generate a reasonable number of problem behavior episodes across the three classes of problem behavior, which is an obvious advantage over uncontrolled data collection with actual patients. Given the staff's extensive training and experience in working with the target population, the simulated sessions are realistic in the sense that the participants exhibited problem behaviors typically observed and treated by the clinic. Demographic variance in the actors' themselves (e.g., gender, height, weight) further increased variability in the expression of the observed behavior data. Because of these factors, we can hypothesize that these data are a very reasonable proxy for problem behaviors of the target population.

The results of the case study, where we applied our sensing and analysis system to a real behavior assessment session with an autistic child, confirm this hypothesis. We were able to replicate the promising recognition results from the validation experiment with a moderate drop in recognition accuracy. Further evidence of the effectiveness of our assessment system was given by its evaluation on non-problem, i.e., "regular" behavior data. Our system produced almost no false positives on a major activity recognition dataset that contains a broad range of domestic activities of daily living.

LESSONS LEARNED AND FUTURE WORK The case study also unveiled further challenges that we need to face in future work. The first concerns the number and placement of sensors on the participant. For example, the child did not tolerate the sensors on his legs. Consequently, we were unable to assess problem behaviors linked to activities of his lower limbs using our current setup. Moreover, this child engaged in a number of problem behaviors that wrist- and ankle-mounted accelerometers would not capture, including biting and head butting. Thus, one key future direction for our work is to investigate how to further minimize the number of sensors and adjust their on-body positions for robust and reliable sensing of a larger range of behaviors.

Furthermore, the need for adaptation techniques became apparent as the child engaged in problem behavior in a very idiosyncratic way. For example, his aggressive behaviors showed a large variability in expression, but occurred only sporadically over the course of the session. On the other hand, his disruptive and self-injurious behaviors occurred with much higher frequency and took on very characteristic forms. Such intra-individual variability in expression of problem behavior is clinically meaningful, yet is not being captured using current practices that focus solely on recording frequencies of occurrence. There is much potential for automated analysis systems to quantify such variability.

While the focus of the current analysis was on problem behaviors, we acknowledge that body-worn accelerometers are appropriate for detecting other types of clinically meaningful behaviors that involve body movements, particularly repetitive and stereotyped behaviors (e.g., hand flapping, body rocking) often exhibited by individuals with autism. In the current analysis, these behaviors would have been classified as non-problem behavior related, and as such, placed in the unknown class. However, our analysis can be extended to differentiate these clinically meaningful behaviors from incidental movements and activities also classified as unknown.

Comparing our additional ground truth annotation to current clinical practice, which involves a human annotator flagging the occurrence of problem behaviors as they happen, reveals several sources of inaccuracy in this practice. First, the observed behaviors can occur at a high frequency, at times faster than a human can manually track. Second, the observer may actually be occluded from seeing the behavior. Both of these sources of error can be improved upon by our automated classification. On the other hand, the automated technique can be inaccurate in cases where the problem behavior is very similar to other (non-severe) behavior. We saw this for the classification of disruption behaviors by foot-mounted sensors (e.g., kicking), which are very similar to ordinary walking behaviors. Furthermore, disruption represents the most challenging class of behavior, likely because it involves a more diverse set of activities than, for example, self-injury (e.g., it includes both hitting the furniture/wall but

also swiping objects, tipping furniture, throwing furniture). One way we hope to address both of these shortcomings is to analyze the sensor data across multiple limbs.

The purpose of automated techniques as they were presented in this paper is to enable clinical researchers to explore new areas of inquiry into behavior analysis, beyond simple frequency counting. For example, our colleagues hypothesize that automatically reinforced behaviors (i.e., the sensation or stimulation provided by the behavior is in itself reinforcing) are going to be more consistent than the same behaviors expressed for a different function, such as for attention. We saw evidence supporting this hypothesis in the analysis of the KID dataset where behavior episodes linked to aggression showed large variance in their sensor data manifestations. These hypotheses, and others like them, can now be formulated and tested in terms of the classification capabilities that our computational approach encourages. Finally, in addition to refining our procedure for clinical assessments, we are working towards our goal of an automatic assessment system for settings outside the clinic. Such a system would enable clinicians to gather data on the occurrence of problem behavior in natural environments, which would allow them to track whether treatment gains observed in a clinical setting generalize to the child's day-to-day life.


### Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1029679. Parts of this work have been funded by the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE), and the German Research Foundation (DFG, Grant No. PL554/2-1).


## REFERENCES


1. T. M. Achenbach and L. A. Rescorla. *Manual for the ASEBA Preschool Forms & Profiles.* Burlington, VT: University of Vermont, Research Center for Children, Youth, & Families., 2000.
2. F. Albinali, M. S. Goodwin, and S. S. Intille. Recognizing stereotypical motor movements in the laboratory and classroom: a case study with children on the autism spectrum. *Proc. Int. Conf. Ubiquitous Computing*, 2009.
3. M. Aman, N. Singh, A. Stewart, and C. Field. Psychometric characteristics of the aberrant behavior checklist. *American J. of Mental Deficiency*, 89:492–502, 1985.
4. *Diagnostic and statistical manual of mental disorders*. Number 4. American Psychiatric Association, 1994.
5. L. Atallah, B. Lo, R. King, and G. Yang. Sensor Positioning for Activity Recognition Using Wearable Accelerometers. *Trans. on Biomedical Circuits and Systems*, 5(4):320–329, 2011.
6. `www.axivity.com`. last visited: June 12th, 2012.
7. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley-Interscience, 2nd edition, 2001.
8. A. Eisenhower, B. Baker, and J. Blacher. Preschool children with intellectual disability; syndrome specificity, behaviour problems, and maternal well-being. *J. of Intellectual Disability Research*, 49:657–671, 2005.
9. G. A. Fink. *Markov Models for Pattern Recognition – From Theory to Applications*. Springer, 2008.
10. S. L. Foster and J. D. Cone. Design and use of direct observation. In A.R. Ciminero, K. Calhoun, and H. E. Adams, editors, *Handbook of behavioral assessment*, pages 253–354. Wiley, New York, 1986.
11. M. S. Goodwin, S. S. Intille, F. Albinali, and W. F. Velicer. Automated Detection of Stereotypical Motor Movements. *J. of Autism and Developmental Disorders*, 41(6):770–782, 2010.
12. F. M. Gresham, T. Watson, and C. Skinner. Functional Behavioral Assessment: Principles, procedures and future directions. *School of Psychology Review*, 30:156–172, 2001.
13. G. P. Hanley, B. A. Iwata, and B. E. McCord. Functional analysis of problem behavior: A review. *J. of Applied Behavior Analysis*, 36(2):147–185, 2003.
14. S. Hartley, D. Sikora, and R. McCoy. Prevalence and risk factors of maladaptive behaviour in young children with autistic disorder. *J. of Intellectual Disability Research*, 52(10):819–829, 2008.
15. S. Herring, L. Gray, J. Taffe, G. Tonge, D. Sweeney, and S. Einfield. Behaviour and emotional problems in toddlers with pervasive developmental disorders and developmental delay: association with parental mental health and family functioning. *J. of Intellectual Disability Research*, 50:874–882, 2006.
16. R. Horner, E. Carr, P. Strain, A. Todd, and H. Reed. Problem behavior interventions for young children with autism: A research synthesis. *J. of Autism and Developmental Disorders*, 32(5):423–446, 2002.
17. P. Howlin, S. Goode, J. Hutton, and M. Rutter. Adult outcome for children with autism. *J. of Child Psychology and Psychiatry*, (45):212—229, 2004.
18. B. A. Iwata and A. S. Worsdell. Implications of Functional Analysis Methodology for the Design of Intervention Programs. *Exceptionality*, 13(1):25–34, 2005.
19. L. Lecavalier, S. Leone, and J. Wiltz. The impact of behaviour problems on caregiver stress in young people with autism spectrum disorders. *J. of Intellectual Disability Research*, (50):172–183, 2006.
20. W. Machalicek, M. O'Reilly, N. Beretvas, J. Sigafoos, and G. Lancioni. A review of interventions to reduce challenging behavior in school settings for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, 1(3):229–246, 2007.
21. J. Matson and S. Lovullo. A review of behavioral treatments for self-injurious behaviors of persons with autism spectrum disorders. *Behavior Modification*, 32(1):61–76, 2008.
22. C.-H. Min and A. H. Tewfik. Automatic characterization and detection of behavioral patterns using linear predictive coding of accelerometer sensor data. In *Proc. Int. Conf. Engineering in Medicine and Biology*, 2010.
23. C.-H. Min and A. H. Tewfik. Novel pattern detection in children with Autism Spectrum Disorder using Iterative Subspace Identification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, 2010.
24. M. R. Patel, J. E. Carr, C. Kim, A. Robles, and D. Eastridge. Functional analysis of aberrant behavior maintained by automatic reinforcement: Assessments of specific sensory reinforcers. *Research in Developmental Disabilities*, 21(5):393–407, 2000.
25. T. Plötz, N. Hammerla, and P. Olivier. Feature Learning for Activity Recognition in Ubiquitous Computing. In *Proc. Int. Joint Conf. on Art. Intelligence*, 2011.
26. D. Roggen et al. Collecting complex activity data sets in highly rich networked sensor environments. In *Proc. Int. Conf. Networked Sensing Systems*, 2010.
27. J. Rojahn, J. Matson, D. Lott, A. Esbensen, and Y. Smalls. The behavior problems inventory: An instrument for the assessment of self-injury, stereotyped behavior, and aggression / destruction in individuals with developmental disabilities. *J. of Autism and Developmental Disorders*, 31(6):577–588, 2001.
28. B. Schölkopf and A. Smola. Learning with kernels: Support vector machines, regularization, optimization, and beyond., 2002.
29. R. Smith and B. Iwata. Antecedent influences on behavior disorders. *J. of Applied Behavior Analysis*, 30:343–375, 1997.
30. T. Westeyn, K. Vadas, T. Starner, and G. Abowd. Recognizing Mimicked Autistic Self-Stimulatory Behaviors Using HMMs. *Proc. Int. Symp. Wearable Computing*, 2005.
31. International statistical classification of diseases and related health problems (icd-10). World Health Organization, 1992.