Automatically characterizing driving activities onboard smart wheelchairs from accelerometer data

HiuKim Yuen[1], Joelle Pineau[2] and Philippe Archambault[3]

***Abstract*— Wheelchairs play an important role for people living with locomotor impairments. However, power wheelchair users frequently report both minor and major accidents. The goal of this paper is to advocate for the use of robotic technology, in particular sensor-based detection and automatic classification of activities, to track and characterize activities onboard smart wheelchairs. Experiments were conducted in a clinical setting, in which experienced wheelchair users were asked to conduct a set of typical wheelchair activities. This paper presents an end-to-end pipeline for accurately classifying these activities from accelerometer data using signal processing and machine learning methods. Our classifier achieved an overall accuracy of around 50% in a more than 25 classes classification problem, compared to less than 4% with a random classifier. We also explored the possibility of discovering hidden patterns of activities using unsupervised topic modeling methods. We demonstrated the power of the inferred patterns with two use cases, namely story telling and hazard discovery. Altogether, this work provides new tools for characterizing the usage of smart wheelchairs with typical users.**

Index Terms — **assistive technologies, wheelchairs, event detection, accelerometers, rehabilitation robotics**

I. INTRODUCTION

Mobility plays an important role in social participation and quality of life. For individuals who live with locomotor impairments, mobility can be facilitated by the optimal use of assistive devices such as power wheelchairs (PW) [1]. However, PW users frequently report both minor accidents, such as colliding with people, furniture and walls, and major accidents such as tips and falls, which can lead to serious injuries [2]. In order to provide better assistance to this population, the design of intelligent powered wheelchairs using robotics and intelligent system technologies, has received significant attention from the robotics community in recent years [3].

During the last decade, significant research on intelligent wheelchairs has focused on the design and control aspects, including but not limited to human-machine interfaces and autonomous navigation [4], [5], [6], [7], [8], [9]. However, due to the fact that wheelchair-related accidents are not uncommon [10], we believe that monitoring is an equally important aspect in the development of intelligent wheelchairs, or assistive robots in general. In fact, monitoring plays a very important role in the users' training process of powered wheelchairs. Given limited number of training sessions between clinicians and patients before a decision is made regarding patient suitability for controlling the PW on their own, it is important for the clinicians to receive as much useful information as possible. In this regard, an automatic system to characterize driving activities would be helpful because it provides an objective and comprehensive summaries on the patients' driving experiences.

With the goal of developing a full-fledged monitoring system that can characterize wheelchair activities and evaluate safety performances during use of intelligent wheelchairs, this paper presents an end-to-end pipeline, from capturing sensor data to automatic activity recognition, together with empirical validations. More specifically, in term of activity recognition, we have two branches, namely event classification and pattern discovery. In event classification, we derive an efficient and robust activity classifier that can effectively identify previously observed and labelled activities. In pattern discovery, we explore unsupervised learning algorithms to discover high level patterns without the use of any manual annotation. Together, these offer substantial tools for evaluating safety performance and detecting hazardous zones during the use of intelligent wheelchair, as well as characterizing diverse driving activities.

As far as we know, this is the first attempt to apply topic modeling in wheelchair activities. This is also one of the first efforts to train an activity classifier as well as to give a thorough evaluation of the performance with participations from real users in a clinical settings.

II. METHODOLOGY

Figure 1 shows the end-to-end pipeline for our proposed multi-layered model, from capturing raw accelerometer data to event classification and pattern discovery. Each component of this pipeline is presented in detail in this section.

A. Data Logging

For the purposes of this study, a data-logging platform, called the Wireless Inertial Measurement Unit with GPS (WIMU-GPS) (Figure 2), was developed and installed on the smart wheelchair [11]. In this paper, we use only the 3D accelerometer data, which captures the acceleration magnitude in x, y and z directions at a rate of 250 Hz. Figure 3 is a sample of 3D accelerometer signals captured from our sensor. The rationale behind using only 3D accelerometer is to limit the number of sensor inputs in order to explore the power of the proposed methods.

[1]HiuKim Yuen is with School of Computer Science, McGill University, Montreal, Canada `hyuen@cs.mcgill.ca`

[2]Joelle Pineau is with School of Computer Science, McGill University, Montreal, Canada `jpineau@cs.mcgill.ca`

[3]Philippe Archambault is with School of Physical and Occupational Therapy, McGill University, Montreal, Canada `philippe.archambault@mcgill.ca`

Fig. 1: End-to-end pipeline

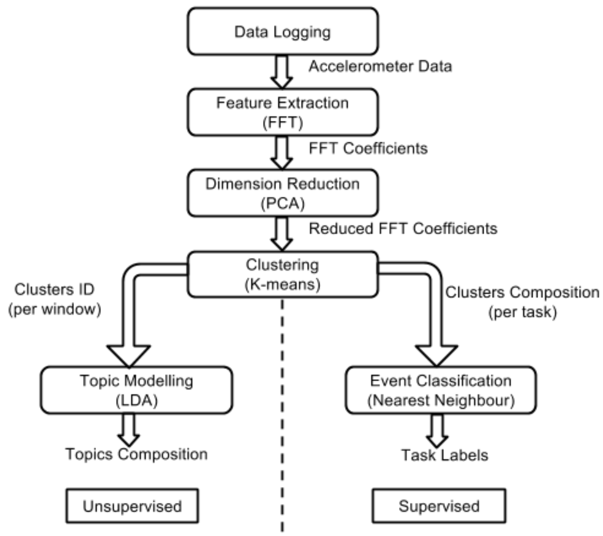
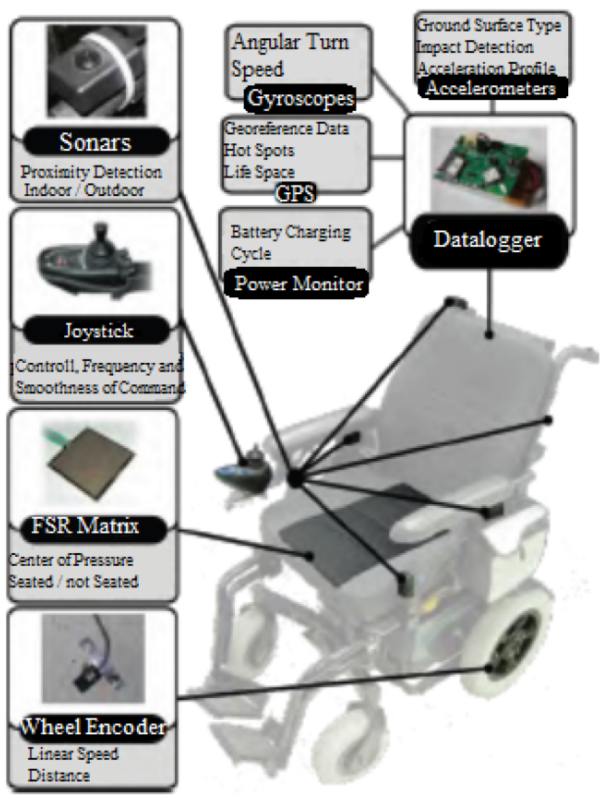


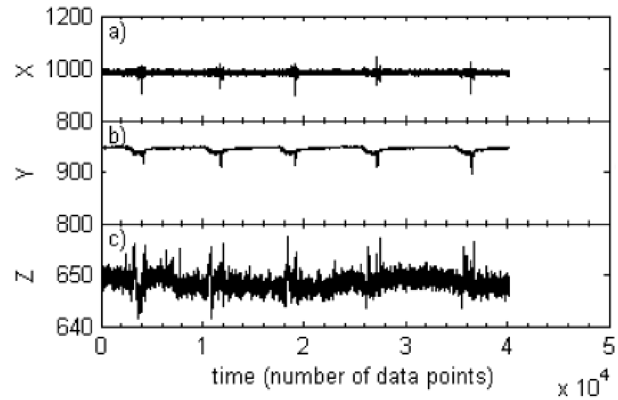
Fig. 2: Overview of datalogging platform (from [11])



B. Feature Extraction and Dimension Reduction

To convert the recorded time-series data into a discrete set of feature vectors, we split the data stream into regular intervals, called windows, and extract representative properties from each window. A sliding window (i.e. having overlap between windows) of size ranging from one to a few seconds has been shown to produce good results in activity recognition [12], [13], [14], [15]. Previous work [12] also

Fig. 3: Sample 3D accelerometer data



considered a detailed comparison of the classification performance on wheelchair activities using four different properties of time series, namely time-domain features, frequency-domain features, wavelet transform features and time-delay embedded features. Among these, frequency-domain features had the strongest predictive performance. Therefore in this paper, we consider only frequency-domain features, using a window size of 2 seconds (and 0.2 seconds sliding overlap).

For each window, we apply a Fast Fourier Transform [16] on each acceleration direction and extract the amplitudes of frequencies ranging from 1Hz to 50Hz (we drop frequencies greater than 50 because the signal strength of those are comparatively very weak). Altogether in 3 directions, we obtain 150 features, i.e. $\{F^x_1, F^x_2, ..., F^x_{50}, F^y_1, F^y_2, ..., F^y_{50}, F^z_1, F^z_2, ..., F^z_{50}\}$. We then apply Principal Component Analysis [17] to reduce the number of features to a small dimension, d, i.e. $\{F_1, F_2, ..., F_d\}$.

As a minor point, for the purposes of testing our approach for event classification, we eventually divide the recorded data into separate training and testing sets. The best PCA transform is selected using only the training data. We can then apply the same transformation matrix on the testing data.

C. Clustering

At this stage, the output of the feature extraction and dimension reduction could be used directly for output classification, as is common in the machine learning literature. However a significant limitation of this approach is that the classification step (especially the training phase) can be computationally expensive because of the large amount of windows. To overcome this, we further reduce the data using clustering methods to find representative samples of the training data.

We apply K-means clustering on all the windows [18]. As a result of this procedure, each window is assigned a cluster ID. The cluster IDs can then be used directly as input features in Topic Modeling (unsupervised branch of the pipeline in Fig. 1).

Alternatively, for the purposes of event classification (supervised branch of the pipeline), we can also compute a cluster composition for each task. Define N_i as the number of windows for task i, and $w_{i,j}$ as the $j-th$ window of task i and $c_{i,j} \in \{1, 2, ...K\}$ as the assigned cluster of $w_{i,j}$ after clustering. Cluster composition of task i, i.e. $\mathbf{CC_i}$, is then defined as a vector, in which each element corresponds to the percentage of which a particular cluster appeared in task i:

$$\mathbf{CC_i} = < CC_i^1, CC_i^2, ..., CC_i^K >, \text{ where}$$
$$CC_i^k = \sum_{j=1}^{N_i} I\{c_{i,j} = k\}/N_i, \text{ and}$$
$$I\{eq\} = \begin{cases} 1 & \text{if } eq \text{ is true} \\ 0 & \text{if } eq \text{ is false} \end{cases}$$

The cluster composition vector can be used directly as an input to the event classification module. In this case, each sample corresponds to a task, in contrast to the unsupervised case where each sample corresponds to a window with an associated cluster ID.

Similarly to PCA, the K-means clustering selects the K centroids using only the training data. Cluster membership of the datapoints in the testing data is assessed using the clusters selected with the training data.

D. Event Classification

The purpose of the event classification module is to take the cluster composition vector and using supervised learning methods to produce an output corresponding to an activity label.

We considered a variety of methods, including the usual top performing Support Vector Machine [20], and the very simple Nearest Neighbour algorithm [21]. In general, we found that Nearest Neighbour worked faster and achieved better (or equally good) performance (We do not intend to argue that Nearest Neighbour will always do better than other classifiers though, and comparisons between different classifiers are outside the scope of this paper). This is consistent with related work on activity recognition from time series data [19]. So all the results reported below use this approach. In short, given a training set D, denote $y_{\mathbf{z}}$ as the label of training sample $\mathbf{z} \in \mathbb{R}^N$, the predicted label $\hat{y}$ on testing sample $\mathbf{x} \in \mathbb{R}^N$ would be:

$$\hat{y}(\mathbf{x}) = y_{\mathbf{z}*} \text{ where } \mathbf{z}* = \underset{\mathbf{z} \in D}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{z}\|_2, \tag{1}$$

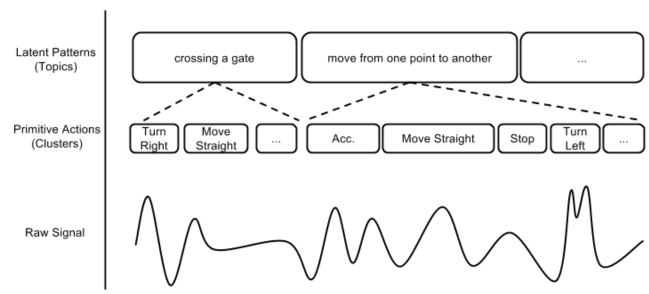
where $\|\mathbf{x} - \mathbf{z}\|_2$ is the euclidean distance between cluster composition vectors.

E. Topic Modeling

Given enough labelled training samples, event classification can effectively recognize activities with reasonable accuracies. However, in real life conditions, this poses significant limitations in terms of (1) dealing with scarcity of labeled data, (2) handling activities that change over time, and (3) discovering new activities. As such we propose to use methods from topic modeling to characterize activities on the smart wheelchair using only unlabeled sensor data.

The basic hypothesis of this approach is that activity patterns should possess a hierarchical structure, as illustrated in Figure 4. The lowest level contains the raw input signal, in our case 3D accelerometer data. On top of this are some primitive action patterns that generate the underlying signals. Primitive actions, as we defined, would be short lasting, roughly 2-3 seconds. These are exactly what we try to capture with the clustering step. Ideally, each cluster would correspond to one type of primitive action. Moving up the hierarchy, these lower level primitive actions are assumed to be generated during the course of some higher level activity patterns, which are unlabeled. The goal of this section is to propose the use of topic modeling methods to infer the high level activity patterns from unabelled data.

Fig. 4: Activity pattern hierarchy



To learn the latent semantic, we use a probabilistic topic model - a type of statistical model used originally in natural language processing for discovering latent topics from text documents [22]. Here is a brief description of topic models: given a text document about a particular topic, say "Machine Learning", we would expect some words to appear more frequently, says "classification", "performance" or "algorithm", etc. Moreover, a document could be composed of multiple topics each with different proportions. The only observable variables are the words of the documents. Usually the "bags-of-words" methodology is applied, only the counts of the words matter, not the ordering. On the other hand, topics (represented as probability distributions), are the unobserved variables that we want to learn. Specifically, in this paper, we consider Latent Dirichlet Allocation (LDA), which is an instance of topic modeling [23]. One particular property of LDA that sets it apart from others is its generative nature. Each word is modeled as if it were generated from the underlying latent topics probabilistically.

To apply LDA to learn the hidden structure, we first have to define what is a document and what a document constitutes in the context of smart wheelchair activities. Our approach is to pull together a fixed number of consecutive windows and consider them as a single document. If we define the document length as L, then the number of documents we get from task i would be $\lfloor N_i/L \rfloor$. By then putting all the documents from all tasks together, we have a total of M documents, where

$$M = \sum_i \lfloor N_i/L \rfloor.$$

Extending our previous notation, if we define $w'_{i,j}$ as the $j-th$ window of document i, and $c'_{i,j}$ as the assigned cluster of window $w'_{i,j}$, the word vector of document i is defined as

$$\mathbf{d_i} =< c'_{i,1}, c'_{i,2}, ..., c'_{i,L} > .$$

Now that we have the representation of documents, another parameter we need to fix in LDA is the number of topics T in our model. We will discuss the choice of T in the "Experiments and Results" section. One output of LDA that we are interested in is the probability distribution of the T topics for each of the M documents denoted by $\theta \in \mathbb{R}^{M x T}$, where $\theta_{i,j}$ is the probability that a given word in document i is generated from topic j.

III. DATASET

A. Data Collections

In a clinical setting, under the monitoring of therapists, 7 real patients (which we will refer to as participants for the rest of the text) were asked to perform a list of driving tasks, extracted from the Wheelchair Skills Test (WST) [24]. The WST provides a training and testing protocol developed to help clinicians assess and train wheelchair users. As such, it represents a rich and diverse set of wheelchair driving activities characteristic of everyday use.

Table I summarizes the types of tasks, together with the number of trials, carried out by each participant. There are a total of 743 tasks, and the average duration of a single trial is around 18 seconds.

B. Training and Testing Sets

Two different sets of experiments are presented. In the first set of experiments, we treat each participant individually, and build a personalized event classifier. In this case, for each participant, we use the first trial of each type of tasks as testing data, and use the remainder as training data. This case is denoted "Individual-Set" in results below. Final classification results are calculated by taking an average over the personalized classifiers. In the second set of experiments, we build a classifier over all subjects and evaluate its ability to generalize to new subjects. As such, we use the first participant's performance as testing data, and use the performance of the other 6 as training data. We use the term "Group-Set" to refer to this classifier in the rest of the paper.

There are two major differences between the two sets of experiments. First, we obviously have more samples in "Group-Set". Second, samples from the "Group-Set" will have higher variances because they are coming from different participants. We can imagine that over multiple trials of the same task, the variations coming from different people would be greater than the variations coming from the same person. It is also worth emphasizing that in "Group-Set", we are trying to test on an unseen participant.

TABLE I: Dataset Summary

Task Code	Description	Average Duration (secs)	P1	P2	P3	P4	P5	P6	P7	Total
T1	Rolls forward 10m	16.99	5	5	5	5	5	5	5	35
T2	Rolls backward 5m	22.39	5	5	5	5	5	5	5	35
T3	Descends 5deg incline	17.71	5	5	5	5	5	5	5	35
T4	Descends 5deg incline	18.46	5	5	5	5	5	5	5	35
T5	Ascends 5cm level change	10.80	5	0	6	5	8	4	5	33
T6	Gets over 15cm pot-hole	9.88	2	0	5	3	0	3	3	16
T7	Descends 5cm level change	7.55	4	0	5	4	5	4	6	28
T8	Gets through hinged door in push direction	49.78	2	1	1	2	1	1	1	9
T9	Gets through hinged door in pull direction	27.17	2	1	1	2	1	1	1	9
T10	Gets over 2cm threshold	12.25	0	4	5	6	0	5	0	20
T11	Rolls 2m on soft surface	12.27	3	2	3	3	0	2	2	15
T12	Turns 90deg left while moving forward	13.77	5	5	5	5	5	5	5	35
T13	Turns 90deg left while moving backward	22.10	5	5	5	5	5	5	5	35
T14	Turns 90deg right while moving forward	12.67	5	5	5	5	3	5	5	33
T15	Turns 90deg right while moving backward	17.24	5	5	5	5	3	5	5	33
T16	Turns 180deg in place clockwise	8.39	4	5	3	5	0	5	0	22
T17	Turns 180deg in place counterclockwise	8.55	5	5	3	5	0	5	10	33
T18	Maneuvers sideways right	33.55	3	3	3	3	3	3	3	21
T19	Maneuvers sideways left	36.74	2	3	3	3	3	3	4	21
T20	Frontal collision	10.91	5	5	5	5	5	5	5	35
T21	Lateral collision right	13.03	5	5	5	5	5	5	5	35
T22	Lateral collision left	10.60	5	5	5	5	5	5	5	35
T23	Collision on moving object	8.59	5	5	5	5	8	5	5	38
T24	Avoids moving object - left	13.68	5	5	5	5	0	5	5	30
T25	Avoids moving object - right	13.36	5	4	5	5	3	5	5	32
T26	Rolls 2m across 5deg side-slop (right-side down)	10.20	0	5	5	0	0	0	0	10
T27	Rolls 2m across 5deg side-slop (left-side down)	10.70	0	5	5	0	0	0	0	10
T28	Rolls 100m to local gym	66.75	1	0	1	2	0	0	0	4
T29	Gets through swing door	14.68	0	5	6	0	0	0	0	11
Total			103	108	125	113	83	106	105	743

P1 to P7 indicate the participants, numbered from 1 to 7. The numbers under columns P1 to P7 represent the number of trials of a particular tasks for a particular participant.

IV. EXPERIMENTS AND RESULTS

A. Event Classification

1) Parameter Fitting: There are a few parameters to select for the proposed method, in particular the dimensionality of the PCA projection (d) and the number of clusters (K) for K-means clustering. For the PCA projection, preliminary results show that the 10 most significant components are sufficient to account for over 98% of the variance.

A common way to determine the number of clusters is by analyzing the intra-cluster variances and inter-cluster variances [25]. However, in our case, optimizing cluster quality does not necessarily align with the goal of optimizing classification performance. Instead, we performed a grid search over cluster sizes from 10 to 100, at an interval of 10 using cross-validations within the training set to select the best number of clusters. In general, we found that the performance usually levels off at around 30 to 50 clusters. More clusters sometimes lead to slightly better results, but not significantly. We set the final number to $K = 40$.

2) Results: We compare classification accuracy for both the Individual-Set and Group-Set setting in Table II. Note that in addition to the method advocated above, we also present results for the case where we use the reduced FFT output directly as a feature (first column), alongside the output of the clustering step (second column). On the other hand, Table III shows the confusion matrix of the "Group-

Set" experiment, where we can observe the breakdowns of which tasks are being mis-classified, and as what other tasks.

Overall, our average accuracy is slightly less than 50%. We would say that our classifier is doing reasonably well considering that there is more than 25 classes, where a random classifier would give an accuracy of just less than 4%. More importantly, by looking at the confusion matrix, we found that most of the mis-classifications came from similar items, e.g. "T20, 21 and 22 - frontal and left/ right lateral collisions", in which they were confused with one another. Even though the classifier failed to distinguish between frontal collision and lateral collisions, it is still doing a good job differentiating collisions as a whole from other tasks, which is mostly enough for practical use. We also believe that this kind of confusion can be resolved easily by incorporating additional sensor input.

Other than the overall accuracy, there are also two important aspects to observe here. First, the better performance of cluster composition on "Group-Set" suggested that this method is more robust to variances, as is the characteristic of the "Group-Set" data. Second, classification on cluster composition is much cheaper in term of computational cost, and thus useful in real-time operation.

TABLE II: Classification Accuracies

Experimental Sets	per-window FFT	per-task Cluster Composition
Individual-Set	49.37%	48.28%
Group-Set	35.92%	46.60%

B. *Pattern Discovery via Topic Modeling*

It is notoriously difficult to evaluate the performance of pattern discovery methods, thus we use a mix of results, including qualitative inspections of topic compositions, and some more quantitative measures in our evaluation.

1) Parameter Fitting: The main parameter to select for LDA is the number of topics. One commonly used metric to evaluate LDA is perplexity [23]. The idea is to set aside some testing data, and infer their likelihood using the trained model. We did cross validation on the training data and found that perplexities stabilized at around 15 topics in most of our experimental settings, and so used 15 topics for the rest of the experiments. In general, there may not be a "correct" number of topics; different numbers of topics can potentially model different complexities of activity patterns.

2) Topic Composition of Documents: As mentioned above, one output of LDA is the probability distribution of topics for each document. We define the topic composition of document i as

$$TC_i =< \theta_{i,1}, \theta_{i,2}, ..., \theta_{i,T} >$$

which is essentially the same as the probability distribution of topics of document i. Intuitively, we consider the composition as an expected realization of the probability distribution. For a given document i of length L, the expected number of words generated from topic t, is simply $\theta_{i,t} \cdot L$. Normalizing it with the total number of words coming from all topics, i.e. $\sum_{t=1}^{T} \theta_{i,t} \cdot L = L$, gives exactly $\theta_{i,t}$.

For demonstration purposes, we show the results for the first participant (similar results are observed for other participants). Also due to space limits, showing the full graph of 301 documents is not possible. Therefore three regions are selected to illustrate some observations, as shown in the upper section of Figure 5. Each column represents one document, which is composed of multiple vertical bars, which sum to 1. Each of the 15 bars represents the composition of a particular topic. For comparison, we also computed the cluster composition (40 clusters) of each document, and plotted them in the lower section of the same figure. The three selected regions are i) "T1: Rolls forward 10m", ii) "T8-T9: Gets through hinged door" and iii) "T28: Rolls 100m to local gym". Tasks in each region possess a certain kind of characteristic: Region i) contains the most clearly defined activities whereas Region ii) contains the most chaotic ones. If you refer to Table III, they correspond to the parts where classification accuracies are 100% and 0% respectively. Region iii), on the other hand, is a good example to demonstrate complex activities. Complex activity is defined as an activity that is composed of numerous interrelated subroutines. As we can imagine, "Rolls 100m to local gym" potentially involves numerous sub-activities, in which "Rolls forward" features prominently.

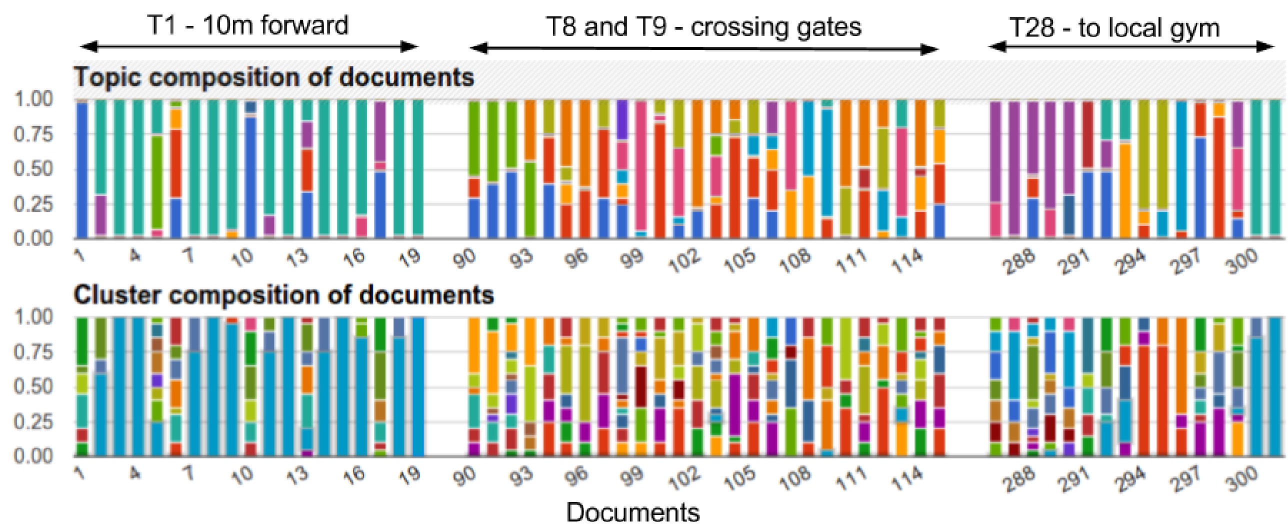
3) Story Telling: We begin with a few qualitative observations, together with a potential use of the computed topic composition called story telling. First, it is readily seen that cluster composition is much more noisy than topic composition, especially for Region ii). Most documents contain more than 4 or 5 major clusters, whereas in the topic composition, documents are mostly dominated by 1 to 2 major topics. Taking a closer look at Region iii) of topic composition, we are able to tell a brief story of what happened during the "Rolls 100m to local gym" period. The dominant topics in the first 5 documents correspond to the dominant topics in the region of "T24, T25: Avoids moving objects" (which are not shown in the figure). We then have 2 to 3 not-so-obvious documents, followed by two documents showing backward-moving patterns, which correspond to the dominant topic in the region of "T2: Rolls Backward 5m" (which is also not shown in the figure). After another 3 to 4 not-so-obvious documents, the activity ends with 2 forward-moving patterns, which correspond to the dominant topic in the region of "T1: Rolls forward 10m" (same dominated color as the first region). The important thing to note here is that if we are to look at the cluster composition instead, we will not be able to tell any of these. Therefore, by uncovering the higher level patterns, it does help us better understand the driving activities.

4) Task Composition of Topics: Another interesting aspect to consider is what constitutes a topic in reverse, in terms of the underlying labels. Ideally, if we have perfectly labelled

TABLE III: Confusion Matrix on Group-Set

	Predicted Tasks																								
Task	T1	T2	T3	T4	T5	T6	T7	T8	T9	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T28
T1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T2	0	0.4	0	0	0	0	0	0	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0	0	0	0
T3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T4	0	0	0.2	0.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T5	0	0	0	0	0.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0.4
T6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T7	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T9	0	0	0	0	0	0	0	0.5	0	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0
T11	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T12	0	0	0	0	0	0	0	0	0	0	0.6	0.2	0	0	0	0.2	0	0	0	0	0	0	0	0	0
T13	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.8	0	0	0	0	0	0	0	0	0	0	0
T14	0	0	0	0	0	0	0	0	0	0	0.2	0	0.6	0	0	0	0	0	0	0.2	0	0	0	0	0
T15	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0.6	0	0	0	0.2	0	0	0	0	0	0	0
T16	0	0	0	0	0	0	0	0	0	0	0	0	0	0.25	0.25	0.25	0	0	0	0.25	0	0	0	0	0
T17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.8	0.2	0	0	0	0	0	0	0	0	0
T18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
T19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0	0	0	0
T20	0	0	0	0	0	0	0	0	0.2	0	0.2	0	0	0	0	0	0	0	0.2	0.2	0	0.2	0	0	0
T21	0	0	0	0	0	0	0	0	0	0	0.2	0	0.4	0	0	0	0	0	0	0.2	0.2	0	0	0	0
T22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.2	0	0.6	0.2	0	0	0
T23	0	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0	0	0	0	0	0.2	0.4	0	0
T24	0	0	0	0	0	0	0	0	0	0	0.2	0	0.2	0	0	0	0	0	0	0	0	0.4	0.2	0	0
T25	0	0	0	0	0	0	0	0	0	0	0.2	0	0	0	0	0	0	0	0	0	0	0.4	0.4	0	0
T28	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

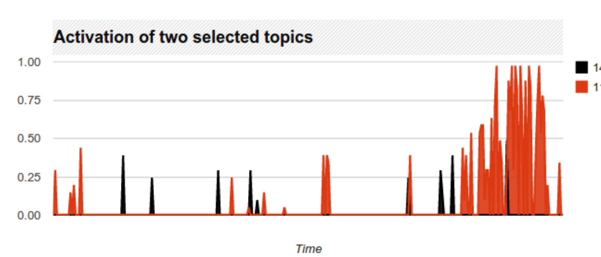
Fig. 5: Topic/Cluster compositions of documents



windows, we could find the composition of primitive actions for each topic. That might not be practical though, since giving labels in a per-window basis involves a tremendous amount of work. As a secondary measure, we use per-task labels (which is also the labelled task code in our dataset) to approximate per-window labels. This means that for all windows coming from a particular task, we simply label them with the task code, and use this to calculate the task composition for each topic. The results are shown in Table IV (we only include tasks that account for at least 10% of total.)

5) Hazard Discovery: Another potential use of the inferred topics in real life is what we call hazard discovery. If we consider a particular topic, and plot its composition across documents, then we can observe the activation of that topic across time (documents are aligned with time). We selected two topics, 11 and 14, to show the idea. The result is shown in Figure 6. Referring to Table IV, topics 11 and 14 constitute mostly "Avoid objects" and "Collisions". Suppose that from prior knowledge, we know that these types of tasks are dangerous, then by analysing their activations across time, we could identify some hazardous zones during the use of the wheelchair. By correlating this with the smart wheelchair's localization in the environment, it may be possible to identify problematic areas, in addition to difficult activities.

Fig. 6: Topic activations for two selected topics



6) Quantitative Measures: In a broader sense, both topics and clusters define a grouping of data points, with the goal

TABLE IV: Task Compositions of Topics

Topics	Dominant Tasks				
1	T4 (14%)	T5 (13%)	T3 (12%)	T1 (10%)	T21 (10%)
2	T8 (19%)	T24 (10%)	T18 (10%)		
3	T18 (26%)	T19 (24%)			
4	T5 (23%)	T4 (21%)	T22 (14%)	T6 (11%)	T23 (10%)
5	T3 (44%)	T4 (31%)	T11 (13%)		
6	T15 (30%)	T12 (16%)	T13 (10%)		
7	T18 (17%)	T19 (12%)			
8	T12 (13%)	T7 (11%)	T21 (10%)		
9	T24 (14%)	T25 (13%)	T18 (12%)		
10	T4 (30%)	T3 (18%)	T22 (10%)		
11	T25 (31%)	T24 (22%)	T28 (17%)		
12	T1 (71%)	T28 (14%)			
13	T2 (51%)	T13 (20%)			
14	T24 (16%)	T20 (14%)	T3 (13%)	T21 (13%)	T8 (10%)
15	T8 (11%)	T13 (10%)			

of putting similar items in the same group and putting different items in different groups. Purity, precision and recall offer quantitative measures to evaluate this kind of grouping quality. They were originally used to analyze clusters, but can be extended easily to analyze topics. In this subsection, we compare these metrics between cluster composition and topic composition. More detailed explanation on these metrics can be found in the Information Retrieval literature (see Chapter 16.3 of [26]), but we will give a short description here. Continuing with our previous notations, purity, precision and recall of cluster composition are defined as:

$$Purity_C = \frac{1}{\sum_i N_i} \sum_k \max_i (CC_i^k \cdot N_i) \tag{2}$$

$$Precision_C = \frac{TP_C}{TP_C + FP_C} \tag{3}$$

$$Recall_C = \frac{TP_C}{TP_C + FN_C} \tag{4}$$

$$\begin{aligned} TP_C &= \sum_k \sum_i \binom{CC_i^k \cdot N_i}{2} \\ &= \sum_k \sum_i \frac{(CC_i^k \cdot N_i) \cdot (CC_i^k \cdot N_i - 1)}{2} \end{aligned}$$

$$FP_C = \sum_k \sum_{i_1} \sum_{i_2 > i_1} (CC_{i_1}^k \cdot N_{i_1}) \cdot (CC_{i_2}^k \cdot N_{i_2})$$

$$FN_C = \sum_{k_1} \sum_{k_2 > k_1} \sum_i (CC_i^{k_1} \cdot N_i) \cdot (CC_i^{k_2} \cdot N_i)$$

Purity, precision and recall for topic composition are defined similarly by replacing CC_i^k with TC_i^t, $\sum_k$ with $\sum_t$ and setting $N_i = L$. For example:

$$Purity_T = \frac{1}{\sum_i L} \sum_t \max_i (TC_i^t \cdot L) \tag{5}$$

$$= \frac{1}{M} \sum_t \max_i (TC_i^t) \tag{6}$$

Intuitively, purity measures the dominance of the most frequent class within the groups, whereas precision and recall measure the correctness of grouping similar items. TP, FP, FN stand for True Positive (number of pairs of windows with the same task labels put in the same group), False Positive (number of pairs of windows with different task labels put in the same group) and False Negative (number of pairs of windows with same task labels put in different groups) respectively.

Table V shows that topic composition performs much better than cluster composition in terms of these metrics. Column 1 and 3 show the results with the best cross-validation selected parameters (40 clusters and 15 topics respectively). For the purpose of comparing the same number of groupings between clusters and topics, we also include results for cluster composition using 15 clusters in Column 2.

TABLE V: Purity, Precision and Recall on Individual-Set

	Cluster Composition (40 clusters)	Cluster Composition (15 clusters)	Topic Composition (15 topics)
Purity	35.74%	25.64%	52.25%
Precision	24.01%	13.33%	36.45%
Recall	18.48%	34.40%	65.99%

C. SUMMARY

To summarize the experimental results, we have achieved around 50% accuracy in Event Classification. We would argue that the classifier is doing reasonably well considering that it is a multi-class classification problem, where a random classifier would give an accuracy of just less than 4%. Moreover, the misclassified items are usually confused with similar items, which is mostly tolerable for practical use. Note also that we have limited the sensor inputs to accelerometer data in this study, and we believe that the classifier can be improved by introducing additional inputs. Again, the rationale behind the choice of limited inputs is to explore the power of the proposed pipeline and methodology instead of constructing an optimized classifier that is ready to use in real life.

As a second contribution, we demonstrate the usefulness of Pattern Discovery with topic modeling in terms of story telling and hazard discovery. We present quantitative results showing that topic composition is a better grouping than cluster composition in terms of purity, precision and recall. Given that unsupervised learning is hard to evaluate, our mixed qualitative and quantitative approach show that topic modeling does provide valuable insights understanding the latent semantics of driving activities. The most promising outcome is that the topic modeling side of the pipeline can be done in a totally unsupervised manner, meaning that no manual annotation is required.

V. DISCUSSION

This paper presents several machine learning approaches to characterize and discover activities during the use of intelligent powered wheelchairs. As a long term vision, our work contributes to the development of a full-fledged monitoring system on smart wheelchairs, as well as other assistive and rehabilitation robots. It is worth noting that the analysis presented here examined the case where the smart wheelchair was under manual control of the human participant. Future work will extend the investigation to the case where the smart wheelchair is under the control of the AI system. We expect the methods to transfer readily, though empirical results may show that the smart controller yields topics composed of significantly different patterns of low-level activities (tasks) when achieving complex behaviors. This may provide useful insights on how to improve automated control strategies for smart wheelchairs. The methodology presented could eventually also be applied to other assistive robots and devices where the collection of accelerometer data is feasible.

It is worth noting that the work presented here is closely related to research on the broader question of automatic recognition of human activities, in which machine learning technologies have been widely used [13], [14]. The pipeline we describe shares some similarities with recent work in this area. In general, much of the focus in that field has been on questions of automatic classification. The pattern discovery question, which is one of the important aspects of our work, has received relatively less attention, though the use of the topic models has been advocated as a good approach for tackling this difficult problem [15].

ACKNOWLEDGEMENTS

The authors would like to thank several colleagues who collaborated on earlier versions of this work: Patrick Boissy, Francois Michaud, Athena Kardehi Moghaddam, and Francois Routhier. This work was financially support by the Natural Sciences and Engineering Research Council Canada (NSERC), the Fonds de recherche du Quebec – Nature et technologies (FRQNT) through the Regroupement strategique INTER, and the CanWheel CIHR Emerging Team.

REFERENCES

[1] I. Pettersson, G. Ahlström, and K. Törnquist. "The value of an outdoor powered wheelchair with regard to the quality of life of persons with stroke: A follow-up study." Assistive technology 19.3 2007: 143-153.

[2] R. C. Simpson, E. F. LoPresti, and R. A. Cooper. "How many people would benefit from a smart wheelchair?." Journal of rehabilitation research and development 45.1, 2008.

[3] IROS 2012 Workshop on Progress, Challenges and Future Perspectives in Navigation and Manipulation Assistance for Robotic Wheelchairs. http://www.radhar.eu/events/IROS2012-robotic-wheelchairs.

[4] P. Boucher, A. Atrash, S. Kelouwani, W. Honoré, H. Nguyen, J. Villemure, F. Routhier, P. Cohen, L. Demers, R. Forget and J. Pineau. "Design and validation of an intelligent wheelchair towards a clinically-functional outcome." Journal of neuroengineering and rehabilitation 10.1, 2013.

[5] C. Gao, T. Miller, J.R. Spletzer, I. Hoffman, T. Panzarella. "Autonomous docking of a smart wheelchair for the automated transport and retrieval system (ATRS)". J Field Robot 2008.

[6] O. Horn O, M. Kreutner. "Smart wheelchair perception using odometry, ultrasound sensors, and camera". Robotica, 2009.

[7] R.C. Simpson. "Smart wheelchairs: A literature review". Journal of Rehabilitation Research & Development. 2005, 42:4. 423-438.

[8] L. Montesano, M. Diaz, S. Bhaskar, J. Minguez. "Towards an intelligent wheelchair system for users with cerebral palsy". IEEE Trans Neural Syst Rehabil Eng 2010.

[9] S. Gulati, B. Kuipers. "High performance control for graceful motion of an intelligent wheelchair." In Proceedings - IEEE International Conference on Robotics and Automation. 2008.

[10] R.L. Kirby, S.A. Ackroyd-Stolarz, M.G. Brown, S.A. Kirkland, D.A. MacLeod. "Wheelchair-related accidents caused by tips and falls among noninstitutionalized users of manually propelled wheelchairs in Nova Scotia." American Journal of Physical Medicine & Rehabilitation 73.5 1994: 319-330.

[11] P. Boissy, S. Brière, M. Hamel, M. Jog, M. Speechley, A. Karelis, J. Frank, C. Vincent, R. Edwards, C. Duval and the EMAP group. "Wireless inertial measurement unit with GPS (WIMU-GPS)Wearable monitoring platform for ecological assessment of lifespace and mobility in aging and disease." Annual International Conference of the IEEE. 2011.

[12] J. Pineau, A.K. Moghaddam, H.K. Yuen, P. Archambault, F. Routhier, F. Michaud, P. Boissy. "Automatic Detection and Classification of Unsafe Events during Power Wheelchair Use." Unpublished.

[13] N. Ravi, N. Dandekar, P. Mysore, M.L. Littman. "Activity recognition from accelerometer data." AAAI. 2005.

[14] L. Bao, S.S. Intille. "Activity recognition from user-annotated acceleration data." Pervasive Computing. Springer Berlin Heidelberg, 2004. 1-17.

[15] T. Huynh, M. Fritz, and B. Schiele. "Discovery of activity patterns using topic models." Proceedings of the 10th international conference on Ubiquitous computing. ACM, 2008.

[16] J.S. Walker. "Fast fourier transforms". Vol. 24. CRC press, 1996.

[17] I. Jolliffe. "Principal component analysis". John Wiley & Sons, Ltd, 2005.

[18] J.B.MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations". 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. 1967: 281-297.

[19] J. Frank, S. Mannor, J. Pineau, D. Precup. "Time Series Analysis Using Geometric Template Matching". Transactions on Pattern Matching and Machine Intelligence (PAMI). 35(3). 2013: 740-754.

[20] C.C. Chang, and C.J. Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 2011: Article. 27.

[21] T. Cover, P. Hart. "Nearest neighbor pattern classification." Information Theory, IEEE Transactions on 13.1 1967: 21-27.

[22] M. Steyvers, and T. Griffiths. "Probabilistic topic models." Handbook of latent semantic analysis 427.7 2007: 424-440.

[23] D.M. Blei, A.Y. Ng, and M.I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research Vol. 3 2003: 993 - 1022

[24] Dalhousie University. Wheelchair Skills Test, version 4.1. Available at: `http://www.wheelchairskillsprogram.ca/eng/4.1/WST\_Manual\_Version4.1.51.pdf`. Accessed March 5, 2010.

[25] S. Ray, and R.H. Turi. "Determination of number of clusters in k-means clustering and application in colour image segmentation." Proceedings of the 4th international conference on advances in pattern recognition and digital techniques. 1999.

[26] C.D. Manning, P. Raghavan, and H. Schtze. Introduction to information retrieval. Vol. 1. Cambridge: Cambridge University Press, 2008: 356 - 359.