Interaction Detection in Egocentric Video: Toward a Novel Outcome Measure for Upper Extremity Function

Article *in* IEEE Journal of Biomedical and Health Informatics · December 2016

DOI: 10.1109/JBHI.2016.2636748

CITATIONS
9

READS
77

2 authors:

Jirapat Likitlersuang
Harvard University
14 PUBLICATIONS **65** CITATIONS
SEE PROFILE

José Zariffa
Toronto Rehabilitation Institute - University Health Network
78 PUBLICATIONS **710** CITATIONS
SEE PROFILE

Interaction Detection in Egocentric Video: Towards a Novel Outcome Measure for Upper Extremity Function

Jirapat Likitlersuang, *Graduate Student Member, IEEE* and José Zariffa, *Member, IEEE*

***Abstract*—In order to develop effective interventions for restoring upper extremity function after cervical spinal cord injury, tools are needed to accurately measure hand function throughout the rehabilitation process. However, there is currently no suitable method to collect information about hand function in the community, when patients are not under direct observation of a clinician. We propose a wearable system that can monitor functional hand use using computer vision techniques applied to egocentric camera videos. To this end, in this study we demonstrate the feasibility of detecting interactions of the hand with objects in the environment from egocentric video. The system consists of a pre-processing step where the hand is segmented out from the background. The algorithm then extracts features associated with hand-object interactions. This includes comparing motion cues in the region near the hand (i.e. where the object is most likely to be located) to the motion of the hand itself, as well as to the motion of the background. Features representing hand shape are also extracted. The features serve as inputs to a random forest classifier, which was tested with a dataset of 14 activities of daily living as well as non-interactive tasks in 5 environments (total video duration of 44.16 minutes). The average F-score for the classifier was 0.85 for leave-one-activity out in our dataset set and 0.91 for a publicly available set (1.72 minutes) when filtered with a moving average. These results suggest that using egocentric video to monitor functional hand use at home is feasible.**

***Index Terms*—Computer vision, Egocentric, Outcome measures, Spinal cord injury, Upper limb rehabilitation.**

I. Introduction

Cervical spinal cord injuries (SCI) can result in paralysis in the upper extremities (UE) and severely limit independence in activities of daily living (ADLs). As a result, the recovery of arm and hand function is the top priority for individuals with cervical SCI [1]. Multiple treatments have been proposed for hand function recovery, ranging from conventional occupational therapy to functional electrical stimulation [2], but further improvements are sorely needed. However, in order to develop new and effective interventions, as well as adapt current treatments to best suit each individual's needs, it is important to have tools to accurately measure hand function throughout the rehabilitation process.

Most existing outcome measures rely on direct observation by a trained clinician in standardized environment (for example the Graded Redefined Assessment of Strength, Sensibility and Prehension [3], the Toronto Rehabilitation Institute Hand Function Test [4], or the Capabilities of Upper Extremity Test [5]). Other measures use questionnaires to gauge independence in ADLs (e.g. the Spinal Cord Independence Measure [6]), but suffer from a reliance on self-report. An important gap left by these tools is that there is currently no viable method to quantitatively collect information about UE function and use once a patient has returned to their home and community. The importance of collecting UE function at home is highlighted in a study by Van Den Berg-Emons et al., which found that rehabilitation physicians underestimated the amount of inactivity in individuals with chronic physical conditions, including SCI [7].

A wearable sensor that collects such data would give a better reflection of a patient's level of independence at home, and help to measure the true impact of an intervention aiming to restore function. Methods based on wearable cameras that record the user's point of view (egocentric video) have the highest potential in this regard because they provide rich data, in contrast to simpler wearable devices (e.g. accelerometers) that are unlikely to capture the complexity of human hand function.

Here we propose a novel system aimed at monitoring hand use at home. In particular, we focus on the problem of detecting interactions of the hand with objects in the environment using an egocentric camera. We propose a hand-object interaction detection system, which we define as providing a binary decision about whether or not the hand is manipulating an object for a functional purpose, irrespective of the specific activity. We postulate that interaction detection will form the basis for a flexible and robust system that will provide valuable information about the amount of functional hand use and level of independence in the community. In this work, we demonstrate the feasibility of hand-object interaction detection from egocentric video, which is a novel computer vision (CV) problem that will form the basis for the proposed wearable system.

Manuscript received May 29, 2016; revised December 1, 2016; accepted December 2, 2016. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2014-05498) and the Rick Hansen Institute (G2015-30).

J. Likitlersuang (jirapat.likitlersuang@mail.utoronto.ca) and J. Zariffa (jose.zariffa@utoronto.ca) are with the Toronto Rehabilitation Institute - University Health Network, Toronto, ON, M5G 2A2, Canada and the Institute of Biomaterials and Biomedical Engineering (IBBME), University of Toronto, Toronto, ON, M5S 3G9, Canada.

II. Related Work

Egocentric vision-based methods have high potential for rehabilitation applications because they are wearable, user specific, and able to follow ADLs in detail. However, little has been done to date to translate advances in egocentric vision to rehabilitation applications. The following describes studies on wearable technologies in healthcare and egocentric CV research that are related to the work described in this paper.

A. Wearable and home sensors for healthcare purposes

Applications for wearable sensors in healthcare include monitoring of health and wellness, rehabilitation assessment, home-based rehabilitation interventions and safety monitoring such as fall detection [8]. When the objective is to monitor UE function, options based on wearable sensors are currently limited. A few studies have explored the use of accelerometers or inertial measurement units (IMUs) to monitor arm movements and attempt to quantify reaching function. For example, in the work by Patel et al. [9], machine learning techniques were applied to accelerometer data in order to successfully estimate the full Wolf Functional Ability Scale (FAS) in stroke survivors. Similarly, Cruz et al. used IMUs to predict FAS scores after stroke [10]. Beyond predicting scores in specific clinical scales, several groups have demonstrated the use of accelerometers to quantify the amount of UE use in the community [11], [12]. The accelerometers are able to provide information about the ratio of hand use between the impaired and unimpaired arm, which is useful after stroke but may be less applicable in conditions that result in more bilateral impairments, such as SCI.

While accelerometers have been used to capture information about reaching and overall UE use, it remains challenging for them to capture information about hand function specifically. The versatility of human hand postures combined with limitations in sensor placement have made it difficult for sensor systems to completely capture the complexity of the hand function [13]. A recently proposed sensor based on magnetometry can capture information about hand movements [14], though the relationship between amount of movement and functional abilities is complex and requires further investigation [15]. Egocentric vision is a recent and promising approach for wearable monitoring of hand function.

B. Egocentric computer vision research for UE monitoring

Although the use of CV for analyzing hand gestures in video from fixed cameras has been the subject of a large body of work [13], [16] egocentric-based CV methods have only recently become an active topic of research [17], [18]. Most relevant to our objective here are studies that have sought to analyze UE function from egocentric video [19]. One group of studies has considered the problem of hand detection and segmentation in egocentric video, which is a crucial pre-processing step in the extraction of any information about the hand. Hand region segmentation in general remains a challenging task for egocentric camera systems, as hands and the background are dynamic and change rapidly from frame to frame, making it hard to separate each region. Recent work on hand segmentation has shown that the most robust and reliable performance for practical applications is often achieved using a flexible colour model of the skin. For example, Li and Kitani proposed a system in which a collection of classifiers were trained based on colour and texture cues, and the most appropriate classifier selected for each given frame under test, based on global appearance of the frame [20]. Several modifications and refinements to this method have been proposed, for example in [21]-[23]. In an alternative approach to deal with widely varying illuminations and scenes, work from Zariffa and Popovic proposed a method based on colour histograms that are adaptive at every frame rather than based on a priori colour model [24]. That study, as well as recent work by Betancourt et al. [25] and Bambach et al. [26] proposed the use of a hand detection step preceding the hand segmentation.

Beyond hand detection and segmentation, there have also been attempts to use egocentric videos for hand activity recognition in ADLs. Some of these studies have employed explicit hand segmentation followed by activity recognition [27]-[30] whereas others have performed the activity recognition task without explicit hand segmentation [31]-[33]. However, all of these activity recognition methods are limited to a predefined set of activities, which may be too restrictive to capture unconstrained hand use in the home or community.

To the best of our knowledge, no previous study has explored the problem of detecting manipulations between the hand and objects (i.e. "hand-object interaction detection"), which is a separate problem from recognizing specific activities.

III. Methods

This paper focuses on the development of a system that can detect interactions of the hand with objects. To facilitate development, we began the validation of this algorithm with a dataset from able-bodied participants. Application to participants with SCI will be addressed in future work.

A. Dataset

We created our own dataset, the Adaptive Neurorehabilitation Systems Laboratory dataset of able-bodied participants ("ANS Able-Bodied"), where a user is wearing a pink glove. The use of the glove was intended to allow easy segmentation in order to focus our testing on the interaction detection problem. The final system will not require the user to wear a glove. The ANS Able-Bodied dataset consists of egocentric video recordings reflective of ADLs obtained using a commercially available egocentric camera (Looxcie 2™) worn by the participant over the ear. The video was recorded in .mp4 format at 480p resolution at 30 frames per second. The data collection was performed at the Intelligent Design for Adaptation, Participation and Technology (iDAPT) Home Lab, a home simulation laboratory in the Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada. For this study, 4 healthy participants (25 ± 6 years of age) were recruited to perform 14 different common ADL tasks (Fig. 1), which are designed to represent everyday

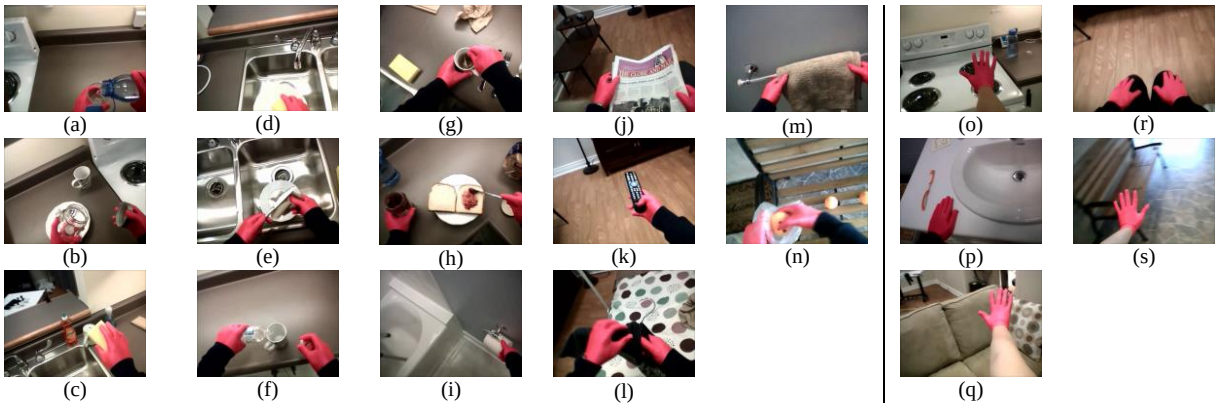

Fig. 1. Example frames of the dataset collected for each of the 14 ADLs at the Home Lab: (a) pouring a water bottle into a coffee cup, (b) opening a jar, (c) picking up a sponge, (d) washing dishes, (e) drying dishes, (f) pouring water from a disposable water bottle, (g) making tea, (h) making a sandwich, (i) changing tissue paper, (j) reading a newspaper, (k) pressing the TV remote, (l) hanging a T-shirt, (m) folding a towel, and (n) picking up a tennis ball, as well as the negative data of no interaction in each room: (o) kitchen, (p) washroom, (q) living room, (r) bedroom, and (s) in front of the house.

activities defined by the American Occupational Therapy Association (AOTA) as important (such as personal care, eating, and social/leisure participation) [34]. The dataset also includes non-interactive tasks from 4 participants (23 ± 1 years of age), 2 of whom also participated in the ADL set. The non-interactive tasks consist of resting the hands statically and moving them in the air without any object interactions, in each of the environments where the 14 interaction tasks were conducted (living room, bedroom, kitchen, washroom and in front of the house). Each of the ADL and non-interactive tasks range from 629 frames (20.97 seconds) to 11,244 frames (6.25 minutes) in duration, resulting in 44.16 minutes for the total dataset. Note that the ADL recordings also include periods when the hand is not interacting with objects (e.g. before and after the task, or in between steps of a more complex activity). The study participants provided written consent prior to participation in the study, which was approved by the Research Ethics Board of the institution (Research Ethics Board, University Health Network: 13-6950-DE). The dataset is available for academic purposes upon request.

B. Hand Segmentation

For the ANS Able-Bodied dataset, 16 frames of the hand wearing the glove were selected from different lighting conditions and environments (i.e. different rooms), where a rectangular area within the region of the glove was selected as a region of interest (ROI). A 3D colour histogram of this ROI was generated in the HSV colour space. Hand segmentation was then performed with the algorithm described in [24], with the exception that the new glove histogram was used instead of the Jones and Regh [35] colour model for the hand detection step (Fig. 2a, b).

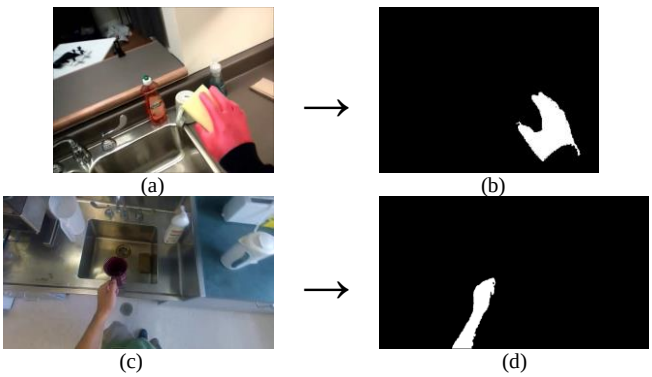

Fig. 2. Example of hand segmentation (a) Image from the ANS Able-Bodied dataset (b) Binary image of the segmentation of the glove, (c) Image from first-person POV video of EDSHK, and (d) Binary image of the segmentation of the skin [20].

Beyond capturing a new dataset of our own, we also sought to evaluate the effectiveness of our approach using a publicly available dataset. We used a subset of the CMU EDSH dataset, EDSHK [20], corresponding to 3,104 frames (1 minute and 43 seconds at 30 frames per second) of making tea in the kitchen. Since this publicly available dataset is of a user's bare hand, a skin detection method was used. Because the work by Li and Kitani [20] has shown high performance on the EDSHK dataset, we used the method described in that study for hand segmentation (Fig. 2c, d).

While both of the segmentation methods described above support the segmentation of multiple hands, for simplicity in this study only one segmented hand region was selected, namely the largest connected component obtained from the hand pixel detection process. Note that when one hand crosses or touches the other, the algorithm will consider both hands together as one segmented region.

Only frames that contain a hand, based on the hand segmentation results, were used in the feature extraction and classification.

C. Feature Extraction

Designing image features that can robustly differentiate between an inactive hand and a hand manipulating an object is a research challenge that, to the best of our knowledge, has never been addressed before. It is expected that two categories of features will play an important role in the interaction detection: motion cues and hand shape. These are detailed in the following sections.

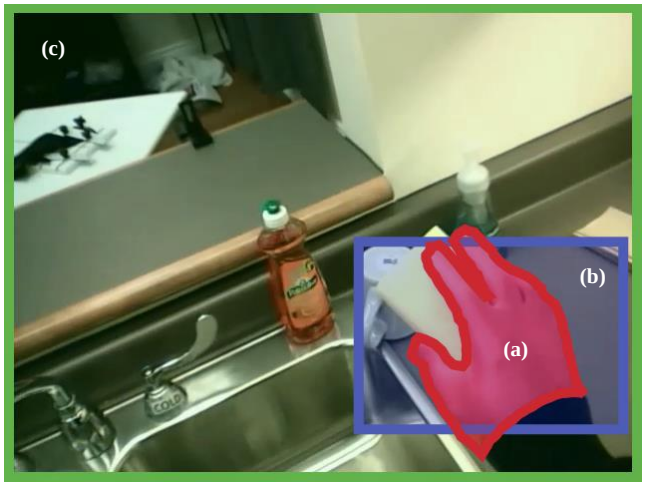

Fig. 3. Example of regions for analysis of optical flow of (a) the hand, (b) the boxed neighborhood of the hand, and (c) background region further from the hand and outside the hand neighborhood box.

1) Motion cues:

Motion cues were obtained using dense optical flow [36]. It is expected that when an object is held in the hand, the object will be moving with a similar optical flow to the hand. Conversely, an item in the frame that is not being interacted with will be more likely to have motion similar to that of the background, as a result of the motion of the head. It can be assumed that the potential object that is being interacted with by the hand must be located near the hand, and for this reason a bounding box was created around the hand. The box was centered on the centroid of the hand segmentation, and its dimensions were 10% of the frame height in each vertical direction and 15% of the frame width in each horizontal direction. Three regions are thus defined: the segmented hand (Fig. 3a), the bounding box around the hand (Fig. 3b), and the background (Fig. 3c). The dense optical flow was first computed for the entire frame and then separated into each of these three regions as described above. The dense optical flow from each region was summarized into respective histograms of magnitude and direction, each with 15 bins. The bins were normalized such that the result is the value of the probability density function at the bin, i.e. the integral over the full range of bins is 1. This allowed the histograms to be compared between the three regions, despite their different dimensions. The final feature consists of two vectors: the subtraction of the histograms of the bounding box near the hand (Fig. 3b) from those of the hand (Fig. 3a), and the subtraction of the histograms of bounding box near the hand (Fig. 3b) from those of the background (Fig. 3a). A lower value after subtraction denotes a closer similarity in motion. Therefore, these features quantify whether the content of the bounding box has a motion more similar to that of the hand or to that of the background.

2) Hand shape:

Hand shape may reflect grip type, which would in turn be a powerful indicator of hand activity, and was represented using histograms of gradients (HOG). The design of our HOG features is identical to the work by Cai et al. [37] in grasp classification from egocentric cameras. A cell size of 8 × 8 pixels and block size of 16 × 16 pixels were used. The HOG features were extracted from the same bounding box used for the hand region and its surrounding in the motion feature analysis described earlier (i.e., Fig. 3a and 3b). Since this bounding box dimension could change due to the bounding box exceeding the dimension of the image or as a function of the image resolution, the bounding box image was resized to 100 by 100. Principal Component Analysis (PCA) was then applied to the HOG feature vector in order to reduce its dimensionality from 960 to 60 (identical dimensions as the features extracted from optical flow).

D. Classifier

Given that our goal is to determine if the hand is interacting with an object, the nature of the classification is a binary classification task. For all of the datasets, we used a random forest classifier [38] where the number of trees in the forest was 150. We tested a number of trees ranging from 50 to 200 and determined that 150 trees maximized accuracy.

The classifier was trained using manually labelled data where each frame is either classified as interaction or no interaction. An interaction between an object and the hand is only considered to happen when the hand manipulates the object for a functional purpose, e.g. resting a hand on the object would not constitute an interaction.

IV. Experiments

To explore the effectiveness of our system in detecting interactions as a distinct problem from activity recognition, we designed testing such that the system never learned the activities beforehand. We also performed testing on a separate dataset from a different recording environment and camera system.

A. Leave-one-activity out

In order to test the robustness of the system in different activities and environments, we applied our system to our ANS Able-Bodied dataset using a leave-one-activity out method. The goal of this evaluation method is to test the system on an activity that has never been trained. In other words, in ADLs, we left one activity out for testing while training on the other 13 activities and the full 5 non-interaction tasks. Similarly, for non-interaction task testing, we left one non-interactive task out for testing, while training on the other 4 non-interaction tasks and a full set of 14 ADL tasks. On average, depending on the activity being left out, the training set consists of 38,264±2,367 frames (1,275.47±78.90 seconds) of interaction (51%) and 37,047±2,469 frames (1,234.90±82.30 seconds) of no interaction (49%). One ADL task used in the test set consists on average of 2,885±2,323 frames (96.17±77.43 seconds) of interaction (81 %) and 667±574 frames (22.23±19.13 seconds) of no interaction (19%), while one non-interaction task used for testing consists on average of 5,955±778 frame (198.50±25.93 seconds, i.e. 100% no-interaction). The classification was compared with manually labeled data. In order to capture the performance of only the hand-object interaction detector, any frame with poor hand segmentation was manually eliminated and not included in the dataset. The accuracy and the F-score for each left-out activity, as well as the overall average accuracy and F-score

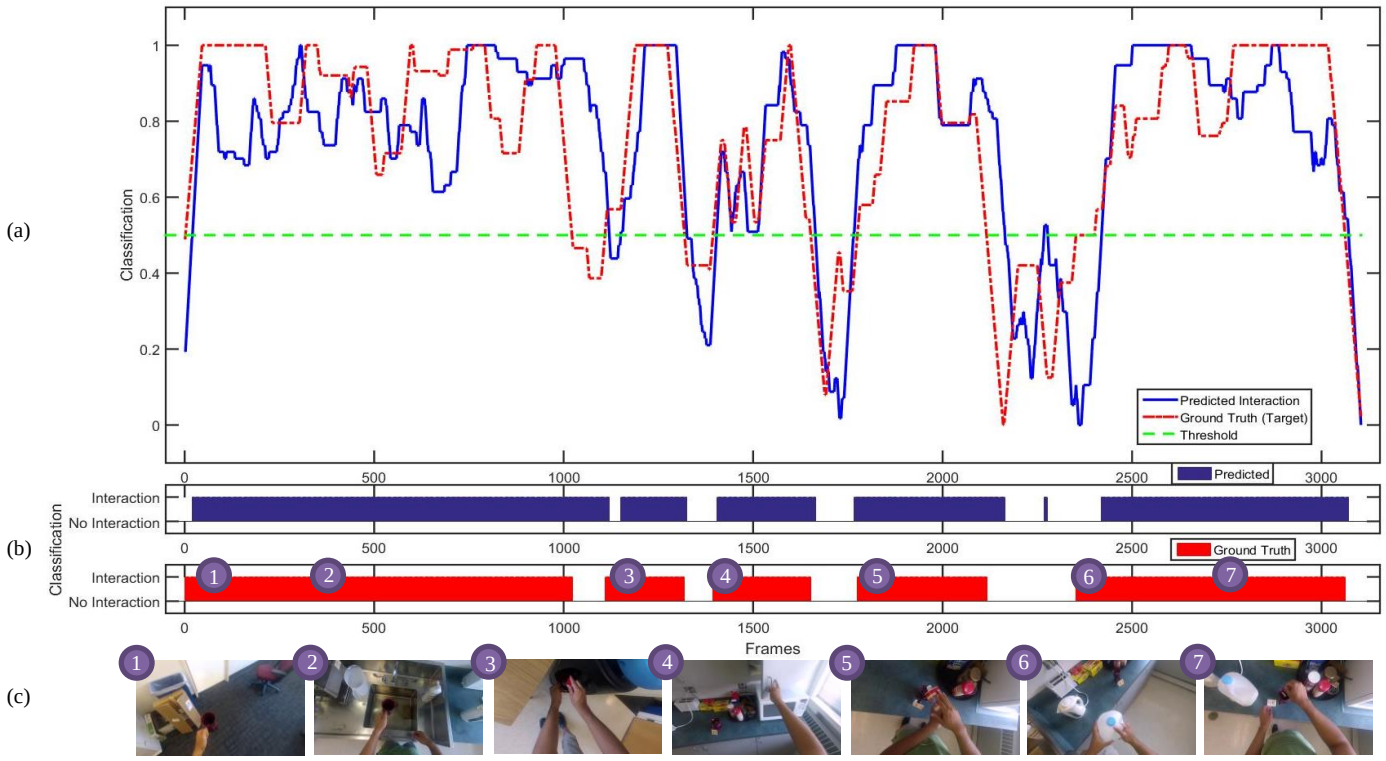

Fig. 4. (a) Plot of moving average of predicted interaction from the classifier (solid blue line) and manually labelled target (dashed red line) in the EDHSK dataset, where 1 is an interaction and 0 is no interaction with the object, as well as the threshold chosen at 0.5. (b) Binary graph showing interaction or no interaction decisions, based on threshold crossings in (a). (c) Example frames of the activities in the associated duration. Note that the change from interaction to no interaction or vice versa indicates a change in activities.

are shown in Table I. The F-score is not provided for testing on the non-interaction tasks because it is not defined when there are no positive entries in the test set.

B. *Testing on a public dataset*

We further evaluated the system using the publicly available EDSHK Dataset. Here, the system is trained using our full ANS Able-Bodied dataset, which includes 40,390 frames of interaction (1,346.33 seconds, 51%) and 39,105 of non-interaction (1,303.50 seconds, 49%). The EDHSK consists of 2,363 frame interaction (78.77 seconds, 76 %) and 741 frames of no interaction (24.70 seconds, 24 %). The classification is compared with data manually labeled for interaction, where all frames are included in the dataset, regardless of hand detection or segmentation quality. This is to provide a realistic assessment of the performance of the system, including cases of poor segmentation of the hand from the background. The accuracy and the F-score on the test EDSHK dataset are shown in Table I.

Whether the hand is interacting with an object, at rest, or moving, the activity will last for a certain duration. We therefore applied a moving average filter to the binary output of the interaction classifier as well as to the manual labels, in order to promote temporal smoothness in the output. We constructed the coefficients of our filter so that the last 90 frames (corresponding to 3 seconds) were equally weighted and averaged (Eq. 1).

$$s_i = \frac{1}{90} \sum_{j=i}^{i+90-1} a_j \qquad (1)$$

Where $\{ a_j \}_{i=1}^{N}$ is the binary hand-object interaction data and $\{ s_i \}_{i=1}^{N-90+1}$ is the new sequence after moving average.

The moving average smoothing method and the associated window duration were chosen empirically on the basis of their ability to meaningfully summarize the number and duration of underlying activities. The output of the moving average was then normalized by subtracting the minimum value over the entire video and dividing by the difference of maximum and minimum values. This was then thresholded such that any frame that was > 0.5 was considered to be an interaction (note: 1 denotes interaction and 0 no interaction). The F-score and accuracy after the application of the moving average are also shown in Table I.

The interaction prediction and the true target (ground truth) from manual labeling after the moving average are also summarized as time series (Fig. 4a), from which meaningful metrics of hand use can be extracted, including the number and duration of interactions (Fig. 4b). Furthermore, different interaction activities and the transitions between them can be distinguished (Fig. 4c). Note that Fig. 4 consists of all frames in EDHSK including those where the hand is not detected. In this example, the interaction detection identified 6 interactions with a total duration of 43.28 seconds, while the manual labeling identified 5 interactions with a total duration of 42.40 seconds.

C. *Features Analysis*

We analyzed each of the features by including only one type of feature at a time (optical flow or HOG) in the classifier. This evaluation was conducted using our dataset for each activity as well as by taking the average of the accuracy

TABLE I
F-SCORE AND ACCURACY OF LEAVE-ONE-ACTIVITY-OUT AND EDHSK

Activity left out	F-score	Accuracy
Pouring a water bottle	0.86	0.76
Opening a jar	0.91	0.84
Picking up a sponge	0.84	0.73
Washing dishes	0.78	0.66
Dry dishes	0.90	0.82
Pouring a disposable water bottle	0.86	0.75
Making tea	0.87	0.76
Making a sandwich	0.86	0.76
Changing tissue paper	0.82	0.71
Reading a newspaper	0.89	0.80
Pressing the TV remote	0.88	0.80
Hanging a T-shirt	0.87	0.78
Folding a towel	0.86	0.77
Picking up a tennis ball	0.75	0.61
Negative - Kitchen	-	0.77
Negative – Washroom	-	0.89
Negative - Living room	-	0.80
Negative – Bedroom	-	0.84
Negative – Front of the house	-	0.87
Mean (S.D.)	0.85 (± 0.04)	0.77 (± 0.07)
EDSHK	0.85	0.76
EDSHK w/ moving average	0.91	0.85

Standard deviation provided in brackets.

TABLE II
F-SCORE AND ACCURACY OF EACH FEATURE

Activity left out	Optical flow only		HOG only	
	F-score	Accuracy	F-score	Accuracy
Pouring a water bottle	0.75	0.62	0.85	0.74
Opening a jar	0.90	0.83	0.90	0.82
Picking up a sponge	0.73	0.58	0.83	0.72
Washing dishes	0.79	0.68	0.75	0.62
Dry dishes	0.80	0.68	0.91	0.84
Pouring a disposable water bottle	0.72	0.60	0.86	0.76
Making tea	0.81	0.70	0.87	0.77
Making a sandwich	0.75	0.64	0.86	0.76
Changing tissue paper	0.72	0.59	0.83	0.72
Reading a newspaper	0.58	0.44	0.91	0.84
Pressing the TV remote	0.63	0.50	0.90	0.82
Hanging a T-shirt	0.76	0.64	0.88	0.79
Folding a towel	0.78	0.66	0.87	0.78
Picking up a tennis ball	0.69	0.59	0.76	0.63
Negative - Kitchen	-	0.70	-	0.74
Negative – Washroom	-	0.67	-	0.89
Negative - Living room	-	0.74	-	0.78
Negative - Bedroom	-	0.74	-	0.84
Negative – Front of the house	-	0.79	-	0.84
Mean (S.D.)	0.74 (± 0.08)	0.65 (± 0.09)	0.86 (± 0.05)	0.78 (± 0.06)
EDHSK	0.74	0.62	0.87	0.78

Standard deviation provided in brackets

and the F-score over all activities, and finally by testing the individual features on the EDHSK dataset (Table II).

D. Performance Comparison

To the best of our knowledge, no previous algorithms have been proposed for the interaction detection problem. In order to provide a comparison point for our results, we implemented a simpler approach based on the assumption that in egocentric videos the hand will be visible mainly when the user is interacting with objects. Thus, the comparison method ("Visible Hand") classifies any frame where the hand is present as containing an interaction (Table III). For the ANS Able-Bodied dataset, Table III considered the same frames as those from Tables I and II, plus frames with no hands. For the EDHSK dataset, all frames are included for all analyses.

TABLE III
ACCURACY OF VISIBLE HAND AND INTERACTION DETECTION METHOD

Dataset	Visible Hand	Interaction Detection
ANS Able-Bodied - ADL tasks (S.D.)	0.86 (± 0.04)	0.81 (± 0.04)
ANS Able-Bodied - Non-interactive (S.D.)	0.33 (± 0.06)	0.89 (±0.04)
Mean (S.D.)	0.72 (± 0.24)	0.83 (± 0.06)
EDSHK	0.80	0.76

Note that this table considers all frames, including those with no hands (frames with poor segmentation are still excluded from the ANS dataset). Standard deviation provided in brackets

V. DISCUSSION

Outcome assessment of the upper limb is crucial in the evaluation of interventions to help restore function after cervical SCI. Despite the importance of having appropriate outcome measures, there is currently no suitable method to collect information about hand function in the community. In this study, we demonstrate the feasibility of interaction detection from egocentric video, which can reflect functional use of the hand. This approach can be used as the basis for a wearable system that has the potential to automatically collect data about hand use in the home and community environments and provide clinicians and researchers with valuable new outcome measures. An individual who is able to independently perform ADLs involving the UE is likely to have more frequent interactions with objects than a more impaired individual.

The interaction classifier designed in this study is robust in multiple activities seen in ADLs, as shown by the F-score, which averages 0.85±0.04 for the tasks in our dataset (Table I). The classification is also robust against non-interactive data, as shown by the average accuracy of 0.83±0.05 (row 15 to 19 of Table I).

The classifier was also shown to be perform well on a publicly available dataset (F-score of 0.85 on EDSHK), showing it to be robust to variations in environments and camera systems.

The use of a moving average smoothed out short-term fluctuations and highlighted longer-term trends. Consider Fig. 4, where the moving average and binary interaction graph is plotted against the number of frames in the video. Here we can clearly observe the number and duration of each of the interactions. This suggests that the change from no interaction to interaction corresponded to transitions between activities as seen in the example frames. This example revealed that an interaction detector could be used to capture meaningful metrics for the measurement of hand usage at home. Note that in Fig. 4 at frame 1 to 139, the hand segmentation fails to detect the hand and instead detects cardboard and wooden objects as hands. This resulted in the descending pattern in the predicted interaction. The results highlight the importance of accurate hand detection and segmentation.

In the ADL tasks and EDHSK, the Visible Hand method outperformed our algorithm in ADLs activities. In contrast, our algorithm had considerably better performance than Visible Hand for the non-interactive tasks, and higher overall classification accuracy on our dataset with all frames included (Table III). The success of the Visible Hand approach in the videos with activities is likely an artefact of the datasets used, which contain ADLs that are realistic in execution but not in frequency of occurrence. A study on hand use based on accelerometer by Lang et al. found that healthy control participants used their upper extremities 8–9 hours per day. However, hemiparetic participants used their affected and unaffected upper extremities substantially less than control participants, 3.3 and 6.0 hours per day, respectively [39].

While that study does not explore hand use in SCI, it is expected that hand use would be similar or lesser than in hemiparetic individuals. This difference was shown by another accelerometer study with sensors on both the upper and lower extremities, which revealed that individuals with SCI had the lowest levels of activity among chronic physical conditions, i.e. 34% compared to able-bodied and 50% compared to stroke [7]. In other words, real recordings of daily life will contain much larger stretches with no interactions than our synthetic dataset, particularly in individuals with SCI. Thus, it is important to correctly classify hands at rest on a table or a lap, as this is expected to constitute the majority of video frames, particularly considering the field of view of many existing commercial cameras. Our algorithm has been shown to be able to distinguish idle hands from hands engaged in functional activities, and is therefore likely to translate better to real-life applications.

Another conceivable method for interaction detection would be to use an activity recognition approach, with a "no activity" class used to denote the absence of interaction, and any other activity class considered an interaction. However, this approach has several shortcomings. Activity recognition is typically a multiclass classification problem that is inherently limited to a predefined set of activities, which may be too limiting for unconstrained data collection in the home or community. Furthermore, many previous studies in egocentric activity recognition did not include a class for no activities [28]-[33]. Activity recognition also typically required multiple preprocessing steps, which may increase computational complexity. For example, a study by Pirsiavash et al. designed a hand manipulation detector after an object detection step, by determining if the object is active or non-active, based on scale and location of the object with respect to the hand [32]. Similarly, Matsuo et al. used an attention classifier as a preprocessing step for hand manipulation, by considering the salience of the detected objects [33].

We explored the contribution of different features to the classification performance in the interaction detection problem. Our main finding on this point, as shown in Table II, is that both the optical flow and HOG features individually were able to provide information about hand-object interactions. The average F-score when using only optical flow was lower than when using only HOG by 0.12 in the leave-one-activity-out and lower by 0.13 for in EDHSK. The comparison of overall leave-one-activity-out and EDHSK results in Tables I and II suggests that HOG alone may be sufficient for interaction detection in able-bodied individuals. However, future examination is needed to determine how well these findings can be generalized, since the results are only different by 0.01 and 0.02 respectively. In particular, in rehabilitation applications, hand postures will generally be impaired and may therefore deviate significantly from those observed in the able-bodied population [40]. It is possible that relying on hand shape alone in this context may lead to poor performance or require classifiers tailored specifically to different types and severities of injury. Further work in clinical populations will be needed to elucidate these issues.

Several challenges remain in the segmentation of the hand from egocentric recordings. Factors that can reduce the performance of an egocentric wearable sensor include glare, faulty segmentation due to objects in the background with colour similar to skin (such as from a wooden floor, door and table), as well as high computational time. However, research on the development of a robust segmentation approach is progressing, as described in the related work section. Since the work presented in this paper is designed to be a proof of concept for the interaction detection problem, we have used a coloured glove for easier segmentation and to minimize errors that may arise from poor segmentation. The use of the glove will be eliminated in future work to make the system practical for clinical use. Nevertheless, we have shown a reasonable performance of our interaction detection in a public dataset, EDSHK, while using a previously reported segmentation algorithm [20], which demonstrates our system is generalizable. Future work is also needed to further test the system in a greater variety of environments and segmentation situations.

It is also important to explore the interactions on a hand-by-hand basis when multiple hands are present in a frame. This consideration is important to ensure that all interactions by either hand are appropriately captured, as well as to better measure the user's independence and reliance on attendant care. This includes implementing an algorithm for differentiating object manipulations performed by the user's right and left hands as well as manipulations performed by a caregiver [41].

Another limitation of this study is the use of artificial no-interaction situations. The ANS Able-Bodied dataset contained sections with no interactions (an overall balance of 51% interaction and 49% no interaction in the dataset), including resting hands and moving hands without interaction. However, within the individual ADL interaction tasks of this dataset (Fig. 1a-n), proportionally there is a larger number of frames with interactions (81%) than frames without interaction (19%), which creates a risk for overfitting, in particular when the classifier is used in real world applications. Rather than intentionally resting or moving the hand, future work will need to focus on providing more natural no-interaction frames through recordings in the home or community.

Future work will also need to explore possible reasons why some activities performed better than others. Based on visual observation, activities that involved complex bimanual tasks (e.g. washing dishes) or handling small objects, where the hand covers the object (e.g. grabbing a tennis ball), more often had low accuracy in detection. A controlled experiment of different object sizes and hand movements is to be explored.

Moreover, future work will need to address the computational limitations. The current algorithm suggested in this study remains computationally expensive for a mobile device. The segmentation of the hand using Li and Kitani's method took 18.84 seconds per frame on average, while the feature extraction for hand-object interaction took 0.18 seconds for optical flow and 0.13 seconds for HOG with PCA feature extraction. The hand-object interaction classification took 0.0029 second per frames (Intel® Xeon® E3-1241 v3: 3.50GHz, DDR3-1600MHz ECC: 16GB). Ideally, the system will need to process videos in real time, such that no video needs to be stored, only the extracted metrics. This will address privacy and confidentiality concerns for the user. Nevertheless, the current approach of off-line processing is still useful in many research applications, for example, using this tool to assess the impact on daily life of a new UE intervention after SCI. Even in the absence of real-time processing capabilities, privacy issues can be effectively managed by giving users the ability to turn off the camera at any time and to review videos before they are shared with investigators.

VI. Conclusion

This study has shown that it is possible to use an egocentric wearable camera to monitor interactions between the hand and objects in the environment. This work may serve as the basis for a wearable system to monitor functional hand use at home in neurorehabilitation applications. The system consists first of a pre-processing step, where the hand is segmented out from the background. We then extract features that are relevant to hand interactions, which include motion cues in the form of optical flow and hand shape information based on HOG. Our study shows that these features provide a strong basis for a hand-object interaction detection classifier that is reliable in a variety of activities and environments. Furthermore, after smoothing, changes between the interaction and no interaction states may serve as good indicators of the occurrence of discrete activities performed by the hand.

Acknowledgement

The authors would like to thank Meng (Melinda) Lu, Emily Qiu, Ryan G. L. Koh, Elizabeth R. Sumitro, and Seung Yun (Susan) Choi for their valuable assistance in the data labeling process. We would also like to thank all the participants of the study.

References

[1] K. D. Anderson, "Targeting recovery: priorities of the spinal cord-injured population," *J. Neurotrauma*, vol. 21, no. 10, pp. 1371-1383, Oct. 2004.

[2] M. R. Popovic et al., "Functional electrical stimulation therapy of voluntary grasping versus only conventional rehabilitation for patients with subacute incomplete tetraplegia: a randomized clinical trial," *Neurorehabil. Neural Repair*, vol. 25, no. 5, pp. 433-442, Jun. 2011.

[3] S. Kalsi-Ryan et al., "The graded redefined assessment of strength sensibility and prehension: reliability and validity," *J. Neurotrauma*, vol. 29, no. 5, pp. 905-914, Mar. 2012.

[4] N. Kapadia et al., "Toronto Rehabilitation Institute–hand function test: assessment of gross motor function in individuals with spinal cord injury," *Top. Spinal Cord Inj. Rehabil.*, vol. 18, no. 2, pp. 167-186, 2012.

[5] R. J. Marino et al., "Reliability and validity of the capabilities of upper extremity test (CUE-T) in subjects with chronic spinal cord injury," *J. Spinal Cord Med.*, vol. 38, no. 4, pp.498-504, Oct. 2014.

[6] A. Catz et al., "A multicenter int. study on the Spinal Cord Independence Measure, version III: Rasch psychometric validation," *Spinal Cord*, vol. 45, no. 4, pp. 275-291, Apr. 2007.

[7] R. J. Van Den Berg-Emons et al., "Accelerometry-based activity spectrum in persons with chronic physical conditions," *Arch. Phys. Med. Rehabil.*, vol. 91, no. 12, pp. 1856-1861, Dec. 2010.

[8] S. Patel et al., "A review of wearable sensors and systems with application in rehabilitation," *J. Neuroeng. Rehabil.*, vol. 9, pp. 21, Apr. 2012.

[9] S. Patel et al., "A novel approach to monitor rehabilitation outcomes in stroke survivors using wearable technology," *Proc. of the IEEE*, vol. 98, no. 3, pp. 450-461, Feb. 2010.

[10] V. T. Cruz et al., "A novel system for automatic classification of upper limb motor function after stroke: an exploratory study," *Med Eng Phys.*, vol. 36, no. 12, pp. 1704-1710, Dec. 2014.

[11] M. Noorkõiv, H. Rodgers, and C. I. Price, "Accelerometer measurement of upper extremity movement after stroke: a systematic review of clinical studies," *J. Neuroeng. Rehabil.*, vol. 11, pp. 144, Apr. 2014.

[12] R. J. Lemmens et al., "Accelerometry measuring the outcome of robot-supported upper limb training in chronic stroke: a randomized controlled trial," *PLoS One*, vol. 9, no. 5, e96414, May 2014.

[13] A. Erol et al., "Vision-based hand pose estimation: A review," *Comput. Vision Image Understanding*, vol. 108, no. 1-2, pp. 52-73, Oct.-Nov. 2007.

[14] N. Friedman et al., "The manumeter: a wearable device for monitoring daily use of the wrist and fingers," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 6, pp. 1804-1812, Jun. 2014.

[15] J. B. Rowe et al., "The variable relationship between arm and hand use: A rationale for using finger magnetometry to complement wrist accelerometry when measuring daily use of the upper extremity," in *Conf. Proc. IEEE. Eng. Med. Biol. Soc.*, Chicago, IL, 2014, pp. 4087-90.

[16] G. R. S. Murthy and R. S. Jadon, "A review of vision based hand gestures recognition," *Int. Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, pp. 405-410, Jul.-Dec. 2009.

[17] A. Betancourt et al., "The evolution of first person vision methods: a survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 774-760, May 2015.

[18] A. R. Doherty et al., "Wearable cameras in health: the state of the art and future possibilities," *Am. J. Prev. Med.*, vol. 44, no. 3, pp. 320-323, Mar. 2013.

[19] A. Betancourt et al., "Towards a unified framework for hand-based methods in first person vision," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Turin, Italy, 2015, pp. 1-6.

[20] C. Li and K. M. Kitani, "Pixel-level hand detection in ego-centric videos," in *IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR)*, Portland, OR, 2013, pp. 3570-3577.

[21] C. Li and K. M. Kitani, "Model Recommendation with Virtual Probes for Ego-Centric Hand Detection," in *IEEE Int. Conf. on Comput. Vision (ICCV)*, Sydney, NSW, 2013, pp. 3570-3577.

[22] G. Serra et al., "Hand segmentation for gesture recognition in ego-vision," in *Proc. of the 3rd ACM int. workshop on Interactive multimedia on mobile & portable devices*, Barcelona, 2013, pp. 31-36.

[23] X. Zhu et al., "Pixel-Level Hand Detection with Shape-Aware Structured Forests," in *Comput. Vision—ACCV*, vol. 9006, Singapore, 2014, pp. 64-78.

[24] J. Zariffa and M. R. Popovic, "Hand contour detection in wearable camera video using an adaptive histogram region of interest," *J. Neuroeng. Rehabil.*, vol. 10, pp. 114, Dec. 2013.
[25] A. Betancourt et al., "A sequential classifier for hand detection in the framework of egocentric vision," in *IEEE Conf. on Comput. Vision and Pattern Recognition Workshops (CVPRW)*, Columbus, OH, 2014, pp. 600-605.
[26] S. Bambach et al., "Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions," in IEEE I Conf. comp. vis., Santiago, Chile, 2015, pp. 1949-1957.
[27] T. Ishihara et al., "Recognizing Hand-Object Interactions in Wearable Camera Videos," in *Int. Conf. on Image Process. (ICIP)*, Québec City, QC, 2015.
[28] A. Fathi et al., "Learning to recognize objects in egocentric activities," in *IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR)*, Providence, RI, 2011, pp. 3281-3288.
[29] A. Fathi et al., "Understanding egocentric activities," in *International Conference on Computer Vision (ICCV)*, Las Vegas, NV, 2011, pp. 407-414.
[30] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, 2013, pp. 2579-2586.
[31] X. Ren and M. Philipose, "Egocentric recognition of handled objects: Benchmark and analysis," in *IEEE Computer Society Conf. on Comput. Vision and Pattern Recognition workshops (CVPR)*, Miami, FL, 2009, pp. 1-8.
[32] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, 2012, pp. 2847-2854.
[33] K. Matsuo et al., "An attention-based activity recognition for egocentric video," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, 2014, pp. 565-570.
[34] S. S. Roley et al., "Occupational therapy practice framework: domain & practice, 2nd edition," *Am. J. Occup. Ther.*, vol. 62, no. 6, pp. 625–683, Nov.-Dec. 2008.
[35] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," in *IEEE Comput. Society Conf. on Comput. Vision and Pattern Recognition*, vol. 1, Fort Collins, CO, 1999, pp. 81-96.
[36] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *SCIA'03 Proc. of the 13th Scandinavian conf. on Image analysis*, Gotenborg, 2003, pp. 363-370.
[37] M. Cai et al., "A Scalable Approach for Understanding the Visual Structures of Hand Grasps" in *Int. Conf. on Robotics and Automation (ICRA)*, Seattle, WA, 2015, pp. 1360-1366.
[38] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*. Berlin, Germany: Springer, 2009.
[39] C. E. Lang et al., "Upper extremity use in people with hemiparesis in the first few weeks after stroke," *Journal of Neurologic Physical Therapy*, vol. 31, no. 2, pp. 56-63, Jun. 2007.
[40] E. C. Fiel-Fote, "Upper extremity training for individuals with cervical spinal cord injury: functional recovery and neuroplasticity" in *Spinal Cord Injury Rehabilitation, 1st ed.* F.A. Davis Company, 2009, ch. 11, pp. 272-3.
[41] J. Likitlersuang and J. Zariffa, "Arm Angle Detection in Egocentric Video of Upper Extremity Tasks," in *IFMBE Proc. World Congress on Medical Physics and Biomedical Engineering*, Toronto, ON, 2015, pp. 1124-1127.

Jirapat Likitlersuang is a Ph.D. candidate at the Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto Ontario, Canada. He previously received a B.A.Sc. in engineering science, with a biomedical engineering major at the University of Toronto.

He is interested in the field of rehabilitation engineering. Specific areas of interest are the development of novel assistive devices and the evaluation of treatments for children and adults living with a disability. He has worked on various biomedical projects in collaboration with scientists, engineers, and clinicians in the Greater Toronto Area. As an undergraduate, he worked to develop a tracking microscope to examine neuronal activity in *C. elegans*. He also worked on an automatic rehabilitation assessment system for upper limbs, as well as educational tools that teach high-functioning children with autism spectrum disorder how to safely cross the road. In collaboration with scientists at the Bloorview Kids Rehabilitation Hospital, he helped to develop a portable pressure sensor device that can be used to examine the biomechanics of mobility assistive devices. Currently, he is working to develop a wearable system capable of monitoring hand function for adults with upper limb dysfunction at the Toronto Rehabilitation Institute – University Health Network.

José Zariffa (M'01) received the Ph.D. degree in 2009 from the University of Toronto's Department of Electrical and Computer Engineering and the Institute of Biomaterials and Biomedical Engineering. He later completed post-doctoral fellowships at the International Collaboration On Repair Discoveries (ICORD) at the University of British Columbia in Vancouver, Canada, and at the Toronto Rehabilitation Institute – University Health Network in Toronto, Canada.

He is a currently a Scientist at the Toronto Rehabilitation Institute – University Health Network and an Assistant Professor at the Institute of Biomaterials and Biomedical Engineering at the University of Toronto in Toronto, Canada. He is also affiliated with the Rehabilitation Sciences Institute, the Edward S. Rogers Sr. Department of Electrical and Computer Engineering and the Collaborative Program in Neuroscience at the University of Toronto. His research interests include technology for upper limb rehabilitation after spinal cord injury, neural prostheses, and interfaces with the peripheral nervous system.

Dr. Zariffa was awarded the 1st place award (Research Category) at the 2012 National Spinal Cord Injury Conference for his work on the motor control of the upper limb after spinal cord injury.