Automated Detection of Puffing and Smoking with Wrist Accelerometers

Qu Tang
College of Engineering
Northeastern University
360 Huntington Avenue
Boston, MA 02115
1-617-320-1212
tang.q@husky.neu.edu

Damon J. Vidrine, Eric Crowder
Department of Behavioral Science,
The University of Texas MD Anderson
Cancer Center, Unit 1330
1155 Herman Pressler Street
Houston, TX 77030-3721
1-713-792-8270, 1-713-745-3648
dvidrine@mdanderson.org,
elcrowder@mdanderson.org

Stephen S. Intille
Colleges of Computer & Information
Science & Health Sciences
Northeastern University
360 Huntington Ave
Boston, MA 02115
1-617-373-3711
s.intille@neu.edu

ABSTRACT
Real-time, automatic detection of smoking behavior could lead to novel measurement tools for smoking research and "just-in-time" interventions that may help people quit, reducing preventable deaths. This paper discusses the use of machine learning with wrist accelerometer data for automatic puffing and smoking detection. A two-layer smoking detection model is proposed that incorporates both low-level time domain features and high-level smoking topography such as inter-puff intervals and puff frequency to detect puffing then smoking. On a pilot dataset of 6 individuals observed for 11.8 total hours in real-life settings performing complex tasks while smoking, the model obtains a cross validation F1-score of 0.70 for puffing detection and 0.79 for smoking detection over all participants, and a mean F1-score of 0.75 for puffing detection with user-specific training data. Unresolved challenges that must still be addressed in this activity detection domain are discussed.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems –*human factors*

General Terms
Algorithms, Measurement, Experimentation, Human Factors

Keywords
Behavior recognition, health, real-time, smoking, cigarette, random forest, supervised learning, ubiquitous computing

1. INTRODUCTION
Cigarette smoking is the principal cause of premature death and disability in the United States [5]. Tobacco smoking is causally linked to 13 different types of neoplastic disease [26]. Of the 230,000 Americans who are killed by cancer every year, over two-thirds of the deaths are attributable to cigarette smoking [1]. Smoking is also reported to be the cause of 6 million preventable deaths yearly [5]. These statistics underscore the important role that smoking cessation research plays in reducing life-threatening diseases and death. Pervasive technologies create new health behavior measurement and intervention opportunities [23], which may also be applicable in tobacco addiction research and treatment.

Smoking monitoring has been studied with technologies ranging from computer vision to text-messaging [16, 17, 25], but the widespread adoption of smartphones and the maturity of wearable sensing technology may provide an alternative, powerful, and practical way to monitor smoking and environmental factors that may influence smoking in daily life. For instance, computer vision systems are restricted to locations where cameras can be affordably installed and may raise privacy concerns; Text messaging systems require user input and cannot proactively detect or intervene on behavior. Wearable technology may provide an affordable way to gather rich data about behavior, anywhere, and respond proactively to passively detected behavior, without creating privacy concerns.

This paper describes an algorithm for automatic detection of puffs and *smoking* activity. *Puffing* is the *hand-to-mouth gesture* (HMG) while *smoking*. And *smoking* is a heterogeneous activity that may include multiple *puffing*, hand *resting* and *tapping* gestures. The algorithm uses a two-layer model, combining a supervised puff-detection classifier and a model based on common smoking topography, such as inter-puff interval and puff frequency. We explore detection of *puffing* and *smoking* from body-mounted motion sensors, where data are gathered from real smokers engaged in *smoking* and also other complex, similar activities, such as *drinking, eating* and *using phone*. To investigate the utility of various sensor positions, the datasets were collected with four accelerometers on the left and right wrists, dominant upper arm and dominant ankle. The *smoking* data includes examples of superposition, interleaving and ambiguity [10]. Despite the complexity of the datasets relative to those used in prior work, the results outperform previously reported systems. Nevertheless, challenges were identified that researchers must ultimately resolve to support useful deployment of the algorithms in future applications. This work demonstrates why the *smoking* detection problem is deceptively complex.

2. BACKGROUND
Wearable sensor devices that measure or recognize health-related behaviors may advance health measurement research and, ultimately, lead to novel, real-time interventions delivered on mobile devices. The behavior recognition problem can typically be divided into three phases: gesture (movement) recognition [8, 11, 24], activity recognition [3, 9, 14, 24] and routine recognition [7]. Gestures (or movements) are described as short term (milliseconds to several seconds) signal sequences [24]. Some gestures are stationary like *resting*; others are non-stationary like *hand-to-mouth gestures*

PervasiveHealth 2014, May 20-23, Oldenburg, Germany

DOI 10.4108/icst.pervasivehealth.2014.254978

Table 1. Prior work in puffing, smoking or eating and drinking detection with accelerometers

Ref	Purpose	Sensor	Dataset				Model			Performance	
			P	Dur	Env	Challenges	Feat	Cls	Alg	Acc	Others
Junker et al., 2008	Drinking/ eating	2 wrist, 2 arm, 1 upper torso	4	4h	Lab	*Drinking*, *eating*	6T	6	Similarity spotting and HMM	N/P	Recall: 80%-90% Precision: 70%
Varkey et al., 2011	Puffing	1 wrist, 1 ankle	3	N/P	Lab	*Eating*, *drinking*	3T	3	SVM	80%	CMR > 20%
	Smoking							6		91%	CMR: 5-9%
Scholl & Laerhoven, 2012	Puffing & smoking detection	1 wrist	4	N/P	Real world	*Smoking* while *sitting*, *standing* or *walking*	2T	2	GMM	N/P	Precision: 51.2% User specific recall: 70%
This study	Puffing	2 wrist, 1 arm, 1 ankle (only wrist used)	6	11.8h	Real world	*Smoking* while *eating*, *walking*, *talking*, *drinking* or *standing*	17T	10	RF and thresholding	83%	F1-score: 0.70
	Smoking							2		N/P	F1-score: 0.79

P: number of participants; Dur: duration; Env: environment used; Feat: number of features; Cls: classes; Alg: algorithm; Acc: accuracy
N/P: not provided; N/A: not applicable; CMR: cumulative misclassification rate; T: time domain features; RF: Random Forest; GMM: Gaussian Mixture Model; SVM: Support Vector Machine; HMM: Hidden Markov Model.

(HMG). Examples of HMGs are raising the hand to mouth for *drinking*, *eating* or *puffing*. Activities can be described as a sequence of gestures that may last for many minutes [24]. An example of an activity is *walking*, which can be formed with many *stepping* gestures. *Eating a meal* is a more complex activity that consists of many HMGs and, probably, *resting* gestures. Lastly, routines are sequences of related activities and contexts [10]. For example, *taking the commuter train to work* can be decomposed into *walking* to the station, *sitting* on a train and *walking* to the office. In our case, *smoking* is considered as a complex activity that consists of a sequence of *puffing* gestures, *hand resting* gestures and, potentially, *hand tapping* gestures. *Smoking* is less homogenous than activities like *walking* and *sitting*.

The bulk of prior work on behavior recognition focuses on detection of homogenous or periodic activities such as *walking*, *standing* and *lying* [3, 9, 13, 18]; Some work went further to detect less homogeneous gestures like HMGs and activities such as *eating* and *drinking* [8, 24], but algorithms are evaluated on datasets collected under controlled conditions in the lab – activity superposition or interleaving is not included, and gestures are nicely segmented from other activities. This work, instead, focuses on developing algorithms to detect a specific complex activity – *smoking* – and uses carefully-annotated data collected from the naturalistic environment for evaluation. The naturalistic dataset forces us to address challenges omitted from prior work.

Kim *et al*. identified three activity recognition challenges: activity superposition, activity interleaving, and activity ambiguity [10]. Superposition refers to the concurrency of multiple activities, such as *walking* while *smoking*. Activity interleaving occurs when an activity is interleaved with other activities. For example, *puffing* may be interleaved with *drinking* HMGs during *smoking* while *drinking* coffee. No prior work related to *smoking* detection specifically addresses this challenge. Finally, activity ambiguity refers to the ambiguous activities based on available sensor data. In our case, *drinking* and *eating* activities are ambiguous as they contain HMGs that look similar to *puffing* given the information collected with our sensors. In this work, we characterize the challenges in the domain of *puffing* and *smoking* detection, testing on data collected in a natural environment by real smokers.

Wearable sensing technologies, such as proximity sensors, respiration bands, wireless cigarette lighters, personal carbon monoxide monitors and accelerometers, have been adopted for *puffing* and *smoking* monitoring [2, 4, 12, 21, 22]. Using a proximity sensor mounted on the wrist and chest for HMG recognition, a sensitivity of 90% has been obtained [12]. Respiration bands provide information only on deep breathing [2]. Instrumented cigarette lighters capture only the starting point of a *smoking* episode [21], but not *puffing*. Personal carbon monoxide monitors may respond to second-hand smoke and may not pinpoint smoking start/stop times. Wrist or arm-worn accelerometers have been used successfully for detection of homogeneous and heterogeneous activities, such as *walking, drinking and eating* [8, 19]; they may be easier to wear, and more informative, energy efficient and affordable than other options for some detection tasks.

Table 1 summarizes prior work on *puffing*/*smoking* detection or *eating* or *drinking* using accelerometers. Junker *et al*. described a two-stage spotting-classification method to detect similar HMGs in activities like *using phone*, *eating* and *drinking*. They obtained 70% precision and 80-90% recall, but the result was only validated with data collected from a semi-naturalistic environment [8]. Varkey *et al*. described a two-level, window-based support vector machine (SVM) algorithm to detect ambiguous activities (*smoking*, *eating*, and *drinking*) and obtained 5-9% cumulative misclassification rate. They also differentiated *puffing* from other gestures within *smoking* with the same algorithm and obtained 20% cumulative misclassification rate [24]. Their analysis, however, did not include any activity concurrency, and the *puffing* detection was evaluated only on samples from within *smoking* episodes. Scholl and van Laerhoven described a Gaussian mixture model (GMM) with features constructed from the "upper" and "lower" gestures in *smoking*. They obtained precision of 51% and user specific recall of 70% with user-annotated naturalistic dataset [22], yet they did not specifically characterize challenges like *smoking* while *eating*. None of the prior work in Table 1 characterized and exploited the topography of smoking.

This work advances prior work 1) by using data collected in a naturalistic way, 2) by addressing, analyzing and handling challenges of superposition, interleaving and ambiguity in both *puffing* and *smoking* detection, 3) by exploiting smoking topography in the smoking detection model, and 4) by developing a new two-level hierarchical classification model for smoking detection.

3. EXPERIMENT

3.1 Sensors and Placement

Table 2. Dataset characteristics

Session	1	3	4	5	6	7	Overall	Typical (see [15])
Number of cigarettes	5	3	5	6	4	4	34	N/P
Number of right puffs(right prototypical puffs)	44(30)	0(0)	46(38)	34(23)	23(15)	38(10)	185(116)	N/P
Number of total puffs(total prototypical puffs)	99(74)	52(35)	55(45)	57(40)	36(19)	56(16)	355(229)	N/P
Puff duration (s)	4.1	6.1	4.3	6.1	6.9	5.2	5.4 ± 1.1	5.1 ± 1.8
Inter-puff Interval (s)	38.4	35.4	30.2	42.4	34.2	34.1	33.1 ± 13.0	25.7 ± 14.2
Number of puffs per cigarette	19.8	17.3	11.0	9.5	9.0	14.0	13.7 ± 6.8	17.8 ± 6.0
Smoking duration per cigarette (min)	14.0	10.8	6.2	7.0	6.2	8.9	8.6 ± 2.8	N/P
Hand swaps rate (per min)	0.16	0.0	0.06	0.16	0.12	0.35	0.13 ± 0.11	N/P
Proportion of concurrent activity (%)	68.0	82.4	51.3	66.4	49.1	40.2	61.0 ± 27.2	N/P
Proportion of drinking/eating (%)	8.9	3.2	2.7	4.0	8.9	7.2	8.0 ± 18.8	N/P

3-axis accelerometer data were collected using an Actigraph GT3X+ monitor sampling at 40 Hz, which is sufficiently high to capture details of limb movement. It has a dynamic range of +/- 6G with resolution 2.93mG. HMGs during *smoking* are wrist and arm movements, and hand *swapping* may also occur. Thus, three monitors were placed on the participant's dominant wrist, dominant upper arm, and non-dominant wrist (see Figure 1(right)). A sensor placed on the ankle captured lower body movements. Figure 1(left) shows sensor orientation on the dominant wrist (the 3-axis orientation follows the "right hand" rule).

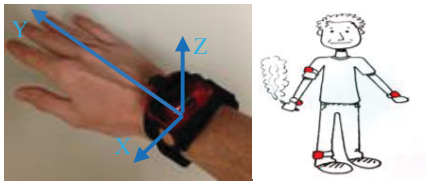

Figure 1. Actigraph GTX3+ and placement on body

3.2 Experimental Protocol

The pilot study was approved by The University of Texas MD Anderson Cancer Center Institutional Review Board. Smokers were recruited who were at least 18 years of age. Participants were observed at a casual restaurant with an outdoor patio allowing smoking at lunchtime. Experimenters met a participant at a table and placed the sensors on the participant's body, ensuring all sensor clocks were synchronized. No restrictions were given to the participant except to remain in the general vicinity of the restaurant and parking area; they were told to smoke cigarettes as they normally would during this period. Two trained research staff observed the participant during the session. Staff used customized annotation software on a tablet computer to objectively document the start and stop times of each activity.

The annotation protocol was designed to allow recording concurrent and interleaving activities while *smoking*. There are four categories: Posture, Activity, Smoking and Puffing. Posture was mutually exclusive and included *sitting*, *standing* and *walking* states. Activity was also mutually exclusive and included *eating*, *drinking*, *reading*, *talking*, *using phone* and *using computer* states. Smoking (true or false) was labeled independently of other activities, allowing for superposition and interleaving. Puffing included three mutually exclusive states: left puffing, right puffing and no puffing. *Smoking* activity was defined from the lighting of a cigarette to the time when the cigarette was thrown away. If a participant put down the cigarette instead of throwing it away, the period of time was still marked as *smoking*. *Drinking* and *eating* were labeled starting from holding cups/forks and ending with finishing food and putting them down. *Puffing* normally lasts 3-7s, which makes it difficult to annotate in real-time. The experimenters marked the start time of *puffing* when the cigarette was put on the mouth and the end time about 2s after the hand was removed from the mouth. Note that hand movement was required for a puff as defined in this experiment. For example, if a participant left the cigarette in the mouth without holding it with the hand, that period was labeled as "no-puff."

3.3 Dataset

Data of 6 two-hour sessions were collected on 6 different right-handed participants (50% male), ages 31 to 51, each with 10+ years of smoking history and an average daily consumption of 20+ cigarettes for the last year. 710 minutes (11.8 h) of data were collected, including 34 *smoking* episodes (287 min) and 481 *puffing* instances. 61% of *smoking* activity time is concurrent with other activities, and 8% of data are ambiguous activities (see Table 2).

4. CHALLENGES

4.1 Puff Inspection

4.1.1 Puff Duration

Typically *puffing* lasts from 3-7s [15, 22] and *puffing* events have a duration of 5.4s ± 1.1s in our dataset (see Table 2). Inspection of the data reveals two observations. First, puff duration varies within and among different individuals (see Table 2). Second, outlier puff durations occur throughout the dataset, but atypically during *smoking* while *using phone/computer* for some participants. For instance, during the 3rd cigarette of Participant 6, 84.8% of the *smoking* episode overlaps with *using phone/computer*, and puff durations of this cigarette have a large standard deviation of 3.7 s.

4.1.2 "Prototypical Puff"

We divide all instances of *puffing* in the dataset into prototypical puffs and non-prototypical puffs, based on visual inspection. A prototypical puff is a normal puff when no other factors (e.g., *eating*) are impacting the signal. A non-prototypical puff is any other puff (see Figure 2(a) and (b)). Prior work has indicated that a puff can be divided into two stages: "hand rising" and "hand lowering" [22, 24]. Based on our data, we decompose a puff into three stages: "hand rising," "hand holding," and "hand lowering" (see Figure 2(a)). The "hand holding" stage normally lasts for 2-4 s when all three axes are stable. Corresponding with each stage, the kinetics of hand motion would be: 1) the hand accelerates to the mouth then suddenly decelerates as it approaches mouth with some rotation (depending upon on the initial orientation of hand), and 2) after the smoker inhales, the hand accelerates down from the mouth and then again decelerates, along with some rotation. The hand transitions and rotations may help to detect *puffing*.

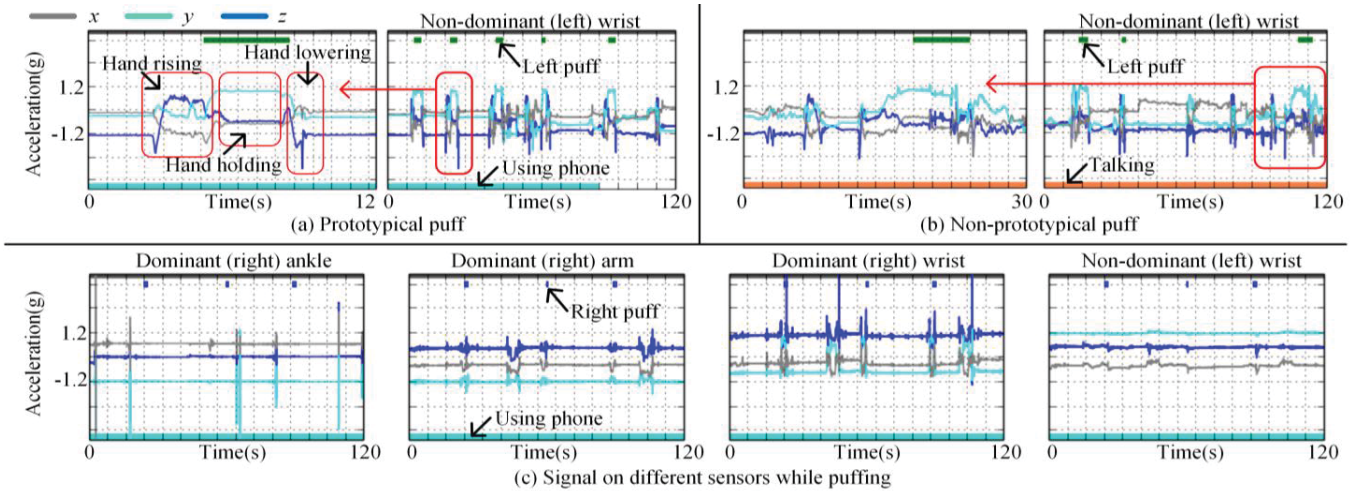

Figure 2. Examples of *puffing*

Figure 2(c) shows a prototypical puff when the participant is *sitting*, *using phone* and *puffing* using the dominant right hand (apparently the left hand is holding the phone). The dominant arm accelerometer shows smaller intensity of movement than that of the dominant wrist, but it does seem to capture some information about the *puffing*. The dominant ankle accelerometer indicates the lower body is not moving. Not surprisingly, the most informative sensor locations for *puffing* and *smoking* are the two wrists.

4.1.3 "Non-prototypical Puff"

35.5% of observed puffs in our data are not prototypical puffs (see Table 2). For Participant 7, the non-prototypical puff percentage is 71.4%. Visual inspection suggests that three factors contribute to non-prototypical puffs.

First, *puffing* is affected by lower body postures: Figure 3(a), from left to right, shows a short period of a *smoking* episode with several puffs while the participant is *sitting*, *standing* and *walking*. The amplitude of *smoking* while *standing* is larger than that of *smoking* while *sitting*, and puff wrist movement is impacted by the *walking*.

Second, *puffing* is affected by concurrent or interleaved hand activities: Figure 3(b) shows three representative cases where the participant is *smoking* while *talking*, *using phone*, and *eating and using computer*. Puffs are mixed with the hand gestures of these activities. Some concurrent activities have less impact, such as *using phone,* likely because the participant is *using phone* with one hand and *puffing* with the other.

Lastly, *puffing* is strongly affected by puffing styles. The participant could start a puff with hand *resting on the thigh* (see Figure 3(c-left)), with *hand hanging near mouth* (see Figure 3(c-middle)), or *with hand hanging aside the ear* (Figure 3(c-right)). A different participant was asked to puff 5 times for each puffing style in these figures. With *hand hanging near mouth*, the amplitude of acceleration change is small. With *hand resting on thigh*, the acceleration values are inverted from *hand holding aside the ear*.

Visual inspection reveals that (1) *puffing* varies within and across individuals, and (2) *puffing* is regularly interleaved with other activities. We suspect that a puffing classification system trained with prototypical puffs, such as might be collected in controlled lab settings or with "simulated" smoking, may not work well on field data of real *smoking* behaviors.

4.2 Smoking Characterization

Smoking behavior can be characterized by 4 highly correlated parameters: 1) puff duration, 2) inter-puff interval per cigarette, 3) number of puffs per cigarette and 4) smoking duration per cigarette. Although related, these parameters are not sufficient proxies for each other [15].

Inter-puff interval characterizes the "*not puffing*" state of *smoking*. Variation in inter-puff interval results from activity superposition and interleaving. For instance, the mean and standard deviation of the inter-puff interval of the 3rd and 4th cigarettes (*using computer*) of Participant 1 exceed those for other cigarettes. Inter-puff interval also varies individually. Participant 5 has a mean inter-puff interval of 42.4 s, which is 1.8 σ from the mean (see Table 2).

Number of puffs per cigarette and smoking duration are used to characterize the frequency of puffs and the duration of smoking (see Table 2). The variability among different individuals is substantial. Participant 1 has the largest average number of puffs (19.8±7.0) and longest average smoking duration (14min±6.3min), resulting from a large inter-puff interval (>2min). "Puff frequency" is the number of puffs divided by smoking duration, which is modeled as a Gamma distribution (shape: κ=8.7, scale: θ=0.2).

Hand swap rate refers to the count of hand changes per minute (from left puffing to right and vice versa) during a cigarette. Along with the proportion of left and right puffs, it reflects hand preference during *smoking*. A person's dominant hand is not necessarily preferred (see Participants 1&3 in Table 2). Hand swap rate ranges from 0 to 0.35 for all the cigarettes smoked in the study.

The topography of smoking behavior might be used to improve a model, but even so, the variability in *smoking* topography in non-lab conditions is likely to impact detection results.

4.3 Concurrent Activities

Besides the variety of puff signatures, and the variability of the way people puff during *smoking*, the context also contributes to ambiguity in *puffing* or *smoking* detection. Contexts that involve *hand-to-mouth gestures* or complex hand movements are particularly challenging. In our scenario, such activities include *drinking beverage*, *eating a meal*, *talking*, *using computer* and *using phone*.

Figure 4(a) shows an *eating a meal* activity with both dominant (right) and non-dominant wrist accelerometers. *Hand-to-mouth*

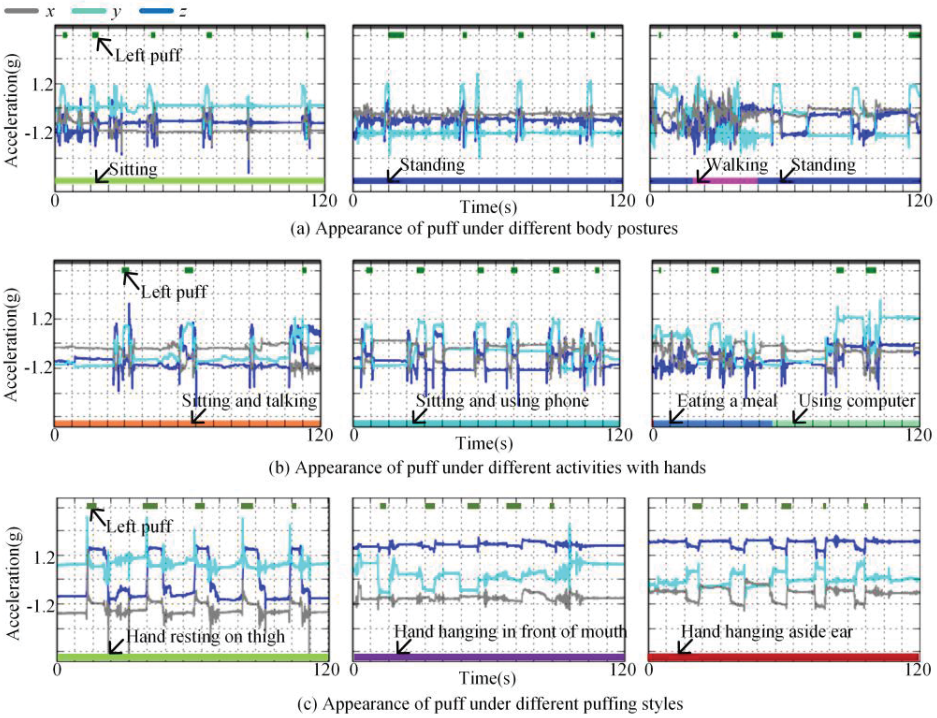

Figure 3. Collection of various non-prototypical puffs

gestures appear repeatedly. The shape of the *eating* HMG is different from that of a prototypical puff. In particular, the "hand holding" stage of *eating* HMGs appears to be more variable than that of *puffing* because chewing food may contain more hand movement than *puffing*. Further, the intervals between *eating* HMGs appear shorter than that of *smoking* in the figure. Lastly, *eating* patterns can be observed in both accelerometers at the same time, while *puffing* only happens with a single hand. Another ambiguous activity is *drinking* (see Figure 4(b)).

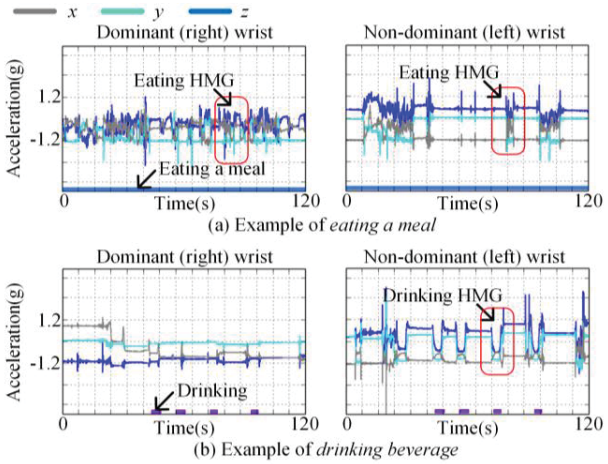

Figure 4. Examples of ambiguous activities

If the participant is *smoking* while *eating*, *drinking* or *using phone*, these concurrent activities can affect the pace of *smoking* and inter-puff intervals (see Figure 5). The accelerometer trajectory of puffs is not impacted by *eating* or *drinking* activity because in the example different hands were used. However, the frequency of *puffing* is affected and becomes slower because of *eating* and *drinking*. The inter-puff interval is 55 s in Figure 5 (about 1.7 σ from the mean, according to Table 2). A similarly long inter-puff interval is observed in Figure 3(b-right). Predictably, these ambiguous activities may trigger false positive predictions, especially if a sensor is used only on a single wrist, as in past work.

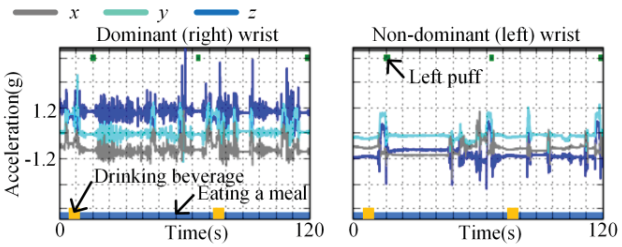

Figure 5. Example of smoking while eating and drinking

5. PUFF CLASSIFICATION

The proposed puff classification system consists four parts: 1) segmentation, 2) feature construction, 3) filtering of segments with no movement and 4) movement classification.

5.1 Model Description

Segmentation. Prior work applied both sliding window [22] and adaptive segmentation [8] to puff detection. Adaptive segmentation may provide more accurate segments for short-term movements, but it may not work well on puffs given that start and end times are ambiguous for non-prototypical puffs (see §4.1.3). The whole raw sequence of each axis of each sensor is filtered using a median filter with size 125 ms to remove spikes due to sensor noise, normalized to have similar range (with zero mean and unit standard deviation), smoothed by a Gaussian filter (σ=2.5), and lastly segmented into fixed length segments by with size *W,* and a step length of 50%*W*.

Features. In total, 51 candidate features for each wrist are computed (17 features on each of three axes). Besides mean, standard

Table 3. Puff classification experiments and results

Exp	Name	Training	Eval	Feat	Model	Acc (%)	Prec (%)	Rec (%)	F1-score	F1-score (puff)	TPR non-proto.(%)	TPR proto. (%)
M1	Individual	Individual	CV	BW	RF	87.7	83.3	83.4	0.83	0.75	72.1	78.0 ±12
M2	Baseline	All	CV	BW	RF	83.0	77.4	81.9	0.79	0.70	72.7	83.7
M3	Single side	All	CV	SW	RF	89.9	79.1	76.9	0.78	0.62	46.8	65.9
M4	Prototypical	All*	CV	BW	RF	84.5	80.9	74.7	0.77	0.64	31.9	68.4
M5	SVM	All	CV	BW	SVM	85.8	76.2	80.2	0.78	0.68	65.2	87.4
M6	LOSO	All	LOSO	BW	RF	67.4	58.7	59.8	0.59	0.40	46.6	43.6

Eval: evaluation; Feat: feature set; Acc: accuracy; Prec: average precision; Rec: average recall; F1-score: average F1-score; TPR non-proto.: true positive rate of non-prototypical puff; TPR proto.: true positive rate of prototypical puff; CV: 5-fold cross validation; LOSO: Leave-one-subject-out validation; BW: both wrists; SW: single wrist; RF: Random Forest; All*: All participant data except for all prototypical puffs.

deviation, maximum, minimum, median, kurtosis, skew, percentile, SNR and RMS of each window, the following additional features have been used. *Peak-peak amplitude* (see also [24]) reflects how strong the hand/body tends to change its speed and may be helpful to differentiate slow transitions (like *reading* and *using computer*) from fast transitions (like *drinking*, *eating* and *puffing*). *Peak rate* is the percentage of local minima or maxima points, which reflect the sudden stop or change of acceleration, corresponding to the edge points between "hand rising" and "hand holding" or between "hand holding" and "hand lowering." Local peak points are detected with a Laplacian of Gaussian (LoG) filter. *Correlation coefficients* (see also [3]) measure the correlation between two axes, which may be an important feature of a prototypical puff (see Figure 2(a)). *Crossing rate between axes* is the percentage of crossing over points between each two axes. Axis crossings have been observed consistently during puffs (see Figure 2(a) & (b)). In short, crossing rate reveals the frequency of rotation/flipping of the hand. Finally, *Slope, MSE and R-squared* are calculated on the linear regression of each segment and may capture relative trend change for a sequence.

Filtering movement. The feature set is imbalanced because the *puffing* instances are rare (in percentage of time). To reduce the imbalance of the training set, we exclude segments with little movement by thresholding the standard deviation of the magnitude of each segment; 25% of segments with little motion are thereby excluded. This step makes the analysis task more complex, because the easy-to-detect no-puff conditions are removed.

Classification. This is a two-class classification problem: *puffing* or *not-puffing*. Random Forests (RF) and SVM were candidates for classification. The RF algorithm is based on the scikit-learn package [20], and the SVM implementation is based on the Python version of LibSVM [6]. The hyper parameters of the RF and SVM are tuned through randomized search of 20 iterations for all experiments, and the criteria used for parameter search is the F1-score. The optimal window size in our case, based on the F1 score for puffing, is 25 s.

5.2 Results and Discussion

Table 3 summarizes the experimental classification results. Note that individual model (M1) refers to models trained and evaluated on each individual, so metrics are averaged. The single-side model (M3) was trained and evaluated with all participant data from the left and right wrist separately. The prototypical model (M4) was trained with only prototypical puff samples but evaluated on the whole dataset. We will use the baseline model (M2) to compare with other models. M2 was also used to evaluate performance for different window sizes (see Figure 6).

Only using accuracy is insufficient for measurement. Accuracy can be biased when a testing set is imbalanced. Comparing between M2 and M4, the higher accuracy (biased to the *non-puffing* class) of M4 may not reflect the actual performance of puffing detection, considering the poor precision and recall measurement. Thus, we choose F1-score as our primary metric.

Longer windows provide more contextual information The F1-score for puffing increases until window size reaches 25 s, because the puff recall increases faster than the decrease of puff precision (see Figure 6). At a window size of 15 s, precision and recall are balanced. With longer window size, each segment includes not only puffing (lasts 2-6 s) but also signals before and after a puff. This information could be useful, when, for example, distinguishing *puffing while just sitting* from activity like *eating*. However, it also increases false alarms (precision drops), because *puffing while eating* and *eating* have similar contexts.

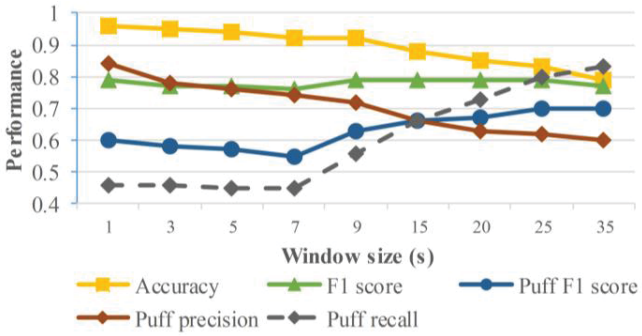

Figure 6. Performance under different window sizes

Variability across individuals affects the performance. The small drop in all metrics (except TPR for prototypical puffs due to its high standard deviation) from M1 to M2 is expected (see Table 3), due to the variability across individuals. For both, the performance was impacted by Participant 6 who has TPRs for non-prototypical and prototypical puffs of 63.9% and 57.1%, respectively with 47.2% of puffs non-prototypical (see Table 2). The performance of puff detection is affected by the non-prototypical puffs.

Using both wrists improves recognition. Using both left and right wrist sensors results in better performance than with a single wrist model (0.1 improvement in overall F1-score and puffs F1-score; an improvement about 20% in TPR for both prototypical and non-prototypical puffs because two-handed activities (*eating*) can be ambiguous with one-handed *puffing* when considering only one wrist (see Figure 4(a-right) as an example).

Be cautious of generalizing prototypical models. Training only on prototypical puffs, TPR for non-prototypical puff recognition drops severely by more than 40%. This finding further emphasizes prototypical vs. non-prototypical puff diversity, consistent with our observations in §4.1. It is a cautionary note against applying models trained with simplified datasets to real-world datasets.

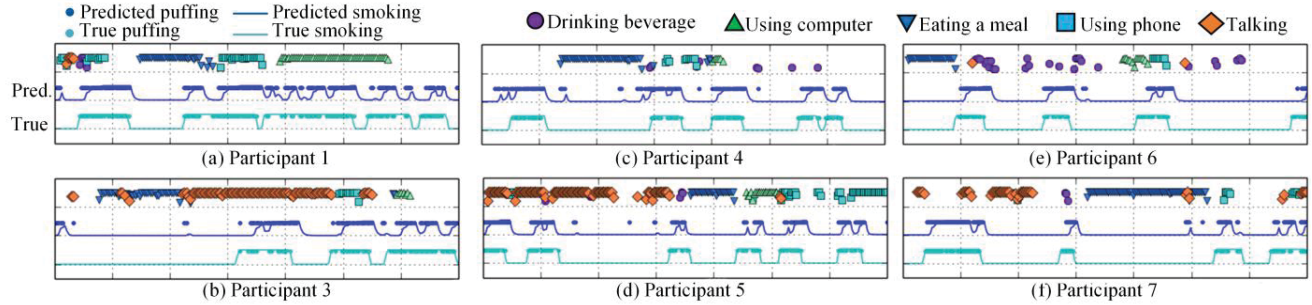

Figure 7. Smoking predictions with activity annotations

SVM is as good as RF. Random Forest and SVM performed equally well for all tests, suggesting that the limitation of puff classification is not the model selection but the complexity of the data (and perhaps features included in the original selection set).

Performance is limited under LOSO validation. Leave-one-subject-out (LOSO) validation tests how well a model built with general training set would work on a new person. The results show huge drops in all metrics. Cross validation may overfit compared to LOSO because training examples may be selected close in time to those used for testing, looking "familiar." M6 suggests that instead of building a generalized puff detection model across all individuals, systems may need to build models specific to individuals.

Concurrent activity results in misclassification. Examining the misclassification cases for M2 shows that non-prototypical *puffing* is misclassified most often, especially when *talking* and *eating-a-meal*. Among the remaining activities, *using phone* followed by *using computer* are most ambiguous, with 66% of examples misclassified as *puffing*. These two activities resemble one-handed *puffing* more than *talking* and *eating-a-meal*.

6. SMOKING DETECTION

The smoking detection model exploits higher-level *smoking* topography (see §4.2) to predict *smoking* status.

6.1 Smoking Prediction Model

The smoking detection model integrates information over longer windows of time in the past to predict the state of smoking at the current time point, using three specific history lengths: 1min, 4min and 8min; The longest history length is set to represent the average smoking duration (see Table 2).

Within each history window, the puff frequency (see §4.2) is calculated by counting the number of puffs from puff classification results. Then the score of smoking for the current window is estimated using the Gamma distribution of puff frequency (see §4.2). Lastly we compute a weighted average of the scores of smoking for the three different window lengths -- weights .4, .3, and .3 for 1min, 4min, and 8min, respectively -- as the final score of smoking at current time point. The model is biased towards the most recent 1 min period in time. The state of *smoking* is then decided by thresholding on the score of *smoking* (0.3 as optimal threshold).

6.2 Results and Discussion

6.2.1 Improvement in Accuracy of Puff Classification

Using smoking typography information for *puffing* detection improves *puffing* recognition 6 percentage points (from 71% to 77%) for non-prototypical puffs, but only 2 percentage points (from 85% to 87%) for prototypical puffs (with window size 25 s). This suggests the smoking detection model can correct some false positives, especially non-prototypical puffs, using temporal patterns of smoking encoded by puff frequency and inter-puff interval.

6.2.2 Investigation of Smoking Prediction

F1-score for *smoking* prediction is 0.79, which is 0.09 higher than *puffing* prediction (see Table 4). Concurrent *eating*/*drinking* activities result in the lowest smoking F1-score (0.65), indicating that temporal patterns of smoking behavior are affected by activities with complex hand movements (see Table 4). Figure 7 shows example predictions of smoking and puffs with activities.

Table 4. F1-score for *smoking* prediction during different concurrent activities

Concurrent activities	All	Eating/ drinking	Using computer	Using phone	Talk-ing	Others
F1-Score	0.79	0.65	0.81	0.88	0.78	0.61
# samples	546	37	82	84	80	144

Sporadic false positive (FP) puffs have limited impact on smoking detection. FPs of puffs are observed before and during the 1st cigarette of Participant 3, between the 1st and 2nd cigarettes of Participant 4, and between the 2nd and 3rd cigarettes of Participant 5 and 7 (see Figure 7). Most do not trigger FPs in smoking prediction because they occur intermittently. But when FPs occur continuously in a short period, they can trigger FP smoking prediction, such as between the 2nd and 3rd cigarette of Participant 7 (see Figure 7).

Reliable detection of smoking can only be accomplished with latency, as we cannot rely on puff detection and must detect multiple puffs. For example, the 1st cigarette of Participant 3 has a 5 min latency (average is about 1-2 min) before detection. All three time scales of the prediction model fail to respond strongly to long inter-puff interval (>2 min) at the beginning of *smoking*.

Prediction fails with unusually-long inter-puff intervals. The 2nd and 3rd cigarettes of Participant 1 have several long inter-puff intervals (>2 min). Our model regards these as gaps between two cigarettes, because the Gamma distribution used to model inter-puff interval suggests these long inter-puff intervals are outliers.

Overall, although the smoking prediction model is able to improve the performance of both puff classification and smoking prediction, it can do so only modestly because of the variability of temporal patterns of smoking behavior.

7. CONCLUSION

The proposed puffing classification model is able to maintain a competitive performance with prior work despite a complex, realistic dataset – F1-score 0.70 for *puffing*, and F1-score 0.79 for *smoking* – and this is when "easy" intervals with little motion have been removed from the dataset prior to analysis.

Analysis of our dataset leads to several findings. First, puff signatures are diverse among and within individuals in our data, due to different puffing styles and concurrent activities. By dividing them into two categories: prototypical puff and non-prototypical puff, we verified that the TPR of non-prototypical puffs is lower than that of prototypical puffs (see Table 3). **This is a caution against using simulated smoking data or data without other ongoing activities included in future smoking detection work.** It also suggests that models trained with prototypical puff instances may perform poorly on non-prototypical puff instances.

Second, the temporal pattern of smoking behavior varies across cigarettes, partially because of concurrent activities. The upper level smoking detection model is able to capture "puff frequency" in multiple history lengths as long as the inter-puff intervals are not too long (>2 min). Modeling smoking independently of puffs is helpful in correcting the false positives of puffing classification as long as false positives occur intermittently (as opposed to clustered in time).

Despite a small dataset, the work presented suggests three future research directions. First, it may be more beneficial to build individual-specific smoking models in real-time, given smoking variability among people as indicated by the results of individual model and LOSO validation. Second, the complexity in detecting smoking from wrist accelerometer data – which on the surface seems a relatively easy recognition problem given the repetitiveness and distinctiveness of puffing – suggests that any human-computer system using automatic puffing/smoking detection from the wrist will need to accommodate a significant amount of error during activities such as *eating* and *talking*. Lastly, work could explore use of other data to help improve the performance such as daily smoking patterns, psychological status (gathered from a just-in-time survey), and walking/sedentary recognition.

8. ACKNOWLEDGEMENTS AND DATA

This work was supported in part by start-up funds awarded to D. Vidrine from the University of Texas MD Anderson Cancer Center; the tools used in this work were supported by National Heart, Lung and Blood Institute, National Institutes of Health award #5UO1HL091737. All MATLAB code and data used to generate these results can be found at http://mhealth.ccs.neu.edu.

9. REFERENCES

[1] World Health Organization. 2012. Last visited 3/15/14, http://www.who.int/features/factfiles/tobacco_epidemic/en/

[2] Ali, A.A., Hossain, S.M., Hovsepian, K., Rahman, M.M., Plarre, K., and Kumar, S., 2012. mPuff: Automated detection of cigarette smoking puffs from respiration measurements. In *Proc. of Conf. on Inf. Proc. in Sensor Networks*, 269–280.

[3] Bao, L. and Intille, S., 2004. Activity recognition from user-annotated acceleration data. *Pervasive Computing*, 1–17.

[4] Beard, E. and West, R., 2012. Pilot study of the use of personal carbon monoxide monitoring to achieve radical smoking reduction. *J. of Smoking Cessation*. 7(1), 12–17.

[5] Centers for Disease Control and Prevention, 2008. Smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 2000-2004. *MMWR*. 57(45), 1226–1228.

[6] Chang, C.C. and Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM TIST*. 2(3), 27.

[7] Huynh, T., Fritz, M. and Schiele, B., 2008. Discovery of activity patterns using topic models. *UbiComp*, 10–19.

[8] Junker, H., Amft, O., Lukowicz, P. and Troster, G., 2008. Gesture spotting with body-worn inertial sensors to detect user activities. *Pattern Recognition*. 41(6), 2010–2024.

[9] Karantonis, D.M., Narayanan, M.R., Mathie, M., Lovell, N.H. and Celler, B.G., 2006. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Tran. on Inf. Technol. in Biomed.* 10(1), 156–167.

[10] Kim, E., Helal, S., and Cook. D., 2010. Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, 48–53.

[11] Lementec, J.C. and Bajcsy, P., 2004. Recognition of arm gestures using multiple orientation sensors: Gesture classification. *Intl. Conf. on Intell. Transport. Sys*., 965–970.

[12] Lopez-Meyer, P., Patil, Y., Tiffany, T. and Sazonov, E., 2013. Detection of hand-to-mouth gestures using a RF operated proximity sensor for monitoring cigarette smoking. *The Open Biomed. Eng. J.* 9, 41.

[13] Mannini, A., Intille, S.S., Rosenberger, M., Sabatini, A.M. and Haskell, W., 2013. Activity recognition using a single accelerometer placed at the wrist or ankle. *Med. Sci. Sports Exerc.* 45(11), 2193–2203.

[14] Mannini, A. and Sabatini, A.M., 2010. Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*. 10(2), 1154–1175.

[15] Marian, C., O'Connor, R.J., Djordjevic, M.V., Rees, V.W., Hatsukami, D.K. and Shields, P.G., 2009. Reconciling human smoking behavior and machine smoking patterns: Implications for understanding smoking behavior and the impact on laboratory studies. *CEBP*. 18(12), 3305–3320.

[16] Obermayer, J.L., Riley, W.T., Asif, O. and Jean-Mary, J., 2004. College smoking-cessation using cell phone text messaging. *J. of Am. College Health*. 53(2), 71–78.

[17] Odetallah, A. and Agaian, S., 2012. Human visual system-based smoking event detection. *Proc. of SPIE*, 840607:1-12.

[18] Oner, M., Pulcifer-Stump, J.A., Seeling, P. and Kaya, T., 2012. Towards the run and walk activity classification through step detection-An android application. *IEEE Intl Conf. of EMBC*, 1980–1983.

[19] Park, J., Patel, A., Curtis, D., Teller, S. and Ledlie, J., 2012. Online pose classification and walking speed estimation using handheld devices. *UbiComp*, 113–122.

[20] Pedregosa, F. and Varoquaux, G., 2012. *Scikit-learn: Machine learning in Python*. Report #1201.0490. arXiv.

[21] Scholl, P.M., Kucukyildiz, N. and Laerhoven, K., 2013. When Do You Light a Fire?: Capturing Tobacco Use with Situated, Wearable Sensors. *Proc. of ACM Conf. on UbiComp Adjunct*, 1295–1304.

[22] Scholl, P.M. and van Laerhoven, K., 2012. A feasibility study of wrist-worn accelerometer based detection of smoking habits. *Intl. Conf. on Innov. Mobile and Internet Services in UbiComp*, 886–891.

[23] Teicher, H., 1995. Actigraphy and motion analysis: New tools for psychiatry. *Harvard Review of Psychiatry*. 3(1), 18–35.

[24] Varkey, J.P., Pompili, D. and Walls, T., 2011. Human motion recognition using a wireless sensor-based wearable system. *Personal and Ubiquitous Computing*. 16, 897–910.

[25] Wu, P., Hsieh, J.W., Cheng, J.C., Cheng, S.C. and Tseng, S.Y., 2010. Human smoking event detection using visual interaction clues. *ICPR*, 4344–4347.

[26] World Health Organization IARC, 2004. *Tobacco smoke and involuntary smoking*.