Real-time Gait Classification for Persuasive Smartphone Apps: Structuring the Literature and Pushing the Limits

Oliver S. Schneider[1]
oschneid@cs.ubc.ca

Karon E. MacLean[1]
maclean@cs.ubc.ca

Kerem Altun[2]
kerem.altun@kemerburgaz.edu.tr

Idin Karuei[1]
idin@cs.ubc.ca

Michael M.A. Wu[1]
mike.wu@alumni.ubc.ca

[1]Department of Computer Science
University of British Columbia
Vancouver, Canada

[2]Department of Mechanical Engineering
Istanbul Kemerburgaz University
Istanbul, Turkey

ABSTRACT

Persuasive technology is now mobile and context-aware. Intelligent analysis of accelerometer signals in smartphones and other specialized devices has recently been used to classify activity (*e.g.*, distinguishing walking from cycling) to encourage physical activity, sustainable transport, and other social goals. Unfortunately, results vary drastically due to differences in methodology and problem domain. The present report begins by structuring a survey of current work within a new framework, which highlights comparable characteristics between studies; this provided a tool by which we and others can understand the current state-of-the art and guide research towards existing gaps. We then present a new user study, positioned in an identified gap, that pushes limits of current success with a challenging problem: the real-time classification of 15 similar and novel gaits suitable for several persuasive application areas, focused on the growing phenomenon of exercise games. We achieve a mean correct classification rate of 78.1% of all 15 gaits with a minimal amount of personalized training of the classifier for each participant when carried in any of 6 different carrying locations (not known *a priori*). When narrowed to a subset of four gaits and one location that is known, this improves to means of 92.2% with and 87.2% without personalization. Finally, we group our findings into design guidelines and quantify variation in accuracy when an algorithm is trained for a known location and participant.

Author Keywords

Gait classification; activity detection; survey; mobile; persuasive computing; exercise games

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

General Terms

Human Factors; Design; Measurement; Performance

IUI'13, March 19–22, 2013, Santa Monica, California, USA.

INTRODUCTION

As smartphones become more powerful and pervasive, developers and designers are using them to effect social or personal change — encouraging sustainable transport [9], facilitating free living for the elderly [35], and supporting individuals in being physically active [7]. The last has led to an abundance of smartphone-based exercise applications and games targeting the global concern of obesity: 10% of adults worldwide are classified as overweight or obese, a risk factor for many diseases [32]. In 2008, 31% of adults worldwide aged 15 and over were insufficiently active [33].

Modern smartphones have a wealth of sensing capabilities, such as accelerometers, gyroscopes, and Global Positioning System (GPS) units. These sensors allow for novel mobile applications to assess the user's context automatically, and permit implicit control of a system rather than requiring explicit control of an application (such as an intelligent music player that automatically pauses when a user stops running). This can support physical activity in a natural way, such as matching a song's beat to a user's running cadence (*i.e.*, step rate) to provide motivation [26], or the ambient display of physical activity, which has helped people maintain their exercise or weight management goals [6]. In particular, automatically sensing different gaits to help infer higher-level context lends itself well to applications involving physical activity. Applications can encourage activity in everyday choices, such as taking the stairs to the office rather than the elevator. Mobile exercise games can also benefit from sensing novel gaits as an input modality: accurate recognition of exercise is important for motivation [6], and many only use location-based sensing or approximations of overall physical activity [2, 11, 12, 17, 20, 30]. Different gaits are also valuable for other persuasive applications, fitting naturally with sustainable transport and unintrusive monitoring of the elderly to encourage positive life choices.

However, results for gait-related activity classification vary drastically in the related literature, obscuring the current state-of-the-art. Meanwhile, variation in methodology and problem characterization complicate finding the best match of algorithm to novel applications. To remedy this, we have identified key classification study dimensions, then organized them into a framework that makes explicit the algorithm's and/or its evaluation's intended target application, imple-

mentation details, and validation methodology. We use this framework to map closely-related work for mobile gait classification that is robust to different locations. Our map is intended as a tool for readers and researchers — a guide for different works and their key differences, a quick reference for study details, and a way to view the current state-of-the-art while placing results in the context of their methodology.

Related work posed within this framework revealed that our target of mobile exercise applications required sensing that is more robust to carrying location, and differentiates more gaits, than past work; this led to the creation of our own new algorithm. We classified 15 walking-category gaits suitable for a range of persuasive smartphone applications, especially exercise and games, and in describing this algorithm's analysis, we articulate our own study parameters within the proposed framework. Finally, we frame our results in design guidelines for using gait classification for mobile exercise applications.

RELATED WORK

Our related work is divided into two parts. In this section, we cover both the potential of gait classification for persuasive mobile applications and the current state of gait or gait-like activity classification using a triaxial accelerometer in a context analogous to a smartphone's.

Gait as Input for Mobile Exercise Applications

Using on-person sensors to support exercise is not new. Examples range from self-reported logging of exercise and nutrition to smartphone-based pedometers and GPS-based location sensing for running or cycling [10, 24, 26, 31]. More elaborate sensing solutions (in particular entailing a feedback and/or social networking element) have also been developed, such as Nike+, whose sensors can communicate with third party exercise equipment [25].

Recent studies have shown that physical activity monitoring can influence physical activity levels when coupled with an ambient display. Pedometers have been used to encourage participants to increase their daily step count via graphical fish avatars [23]. Findings from the "Houston" system, a mobile fitness journal application, suggest that though pedometers are motivating, non-walking activities must be supported. When Houston failed to report non-walking activities (such as climbing) correctly both to users themselves and via social network sharing, users were frustrated [6]. Better reporting was included with the followup system, UbiFit, which provides an ambient display of a garden growing when the user is active or achieves fitness goals [7]. UbiFit allowed automatic recording of some activities and manual entry of any activity, and this was better received by participants: the frustration encountered with Houston was not observed with UbiFit. Over 60 activities were reported, suggesting that automated systems will have to support a large number of activities.

Houston and UbiFit required external sensors, which users felt were large and unsightly — unsuitable for a pervasive application used in day-to-day life. Today's powerful smartphones have potential to reduce reliance on external sensors, but their on-board sensors require more elaborate processing. Exercise games have used accelerometers to infer overall activity level. NEAT-O-Games (Non-Exercise Activity Thermogenesis) are simple games that used hip-mounted external accelerometers to infer an overall physical activity level. Their goal is to increase "energy expenditure of all physical activities other than volitional sporting-like exercise," such as moving around to do chores [11,12,20]. Other common types of input for mobile exercise games include heart-rate monitors [5, 8] and location through GPS or wireless connection points [2, 17, 30].

However, at the time of this writing there are no exercise games that use gait as an input modality; that is, no exercise games directly take into account *how* the user is moving. Previous work on types of exercise game input has focused on situated rather than mobile contexts (*e.g.*, a stationary bike or vision-based system): a set of *abstract inputs* for existing exergames (or "active games"), later developed into a toolkit, has included gesture, stance, point, power, continuous control, and tap [3, 29]. Although some aspects of gait can be fit into these abstract specifications, emphasis was on non-mobile input techniques. There is a clear need to investigate how this new interaction modality could be used to support persuasive applications.

Accelerometer-based Gait Classification

Activity detection has been accomplished with great accuracy using an array of inertial measurement units (IMUs) attached at multiple points to the body (*e.g.*, [1]), but this approach is far too intrusive for pervasive persuasive applications. Recent results using smartphone sensors (or loose accelerometers approaching the context of a smartphone) cite a variety of success rates, up to 100% [19]. A closer look reveals that results vary depending on methodology, such as the number and type of classes (attributes to be classified; in this case, gaits or activities), the constraints placed on a phone's carrying method, or the number of participants. It is thus difficult to get an overall sense of mobile classification work, and how much progress we have made at solving various aspects of this problem. As such, we do not simply report on each study, but provide a structured survey of closely-related work.

In the next section, we present a framework for organizing mobile context-aware classification work, drawing upon specific studies for examples. An explanatory diagram of this framework can be found in Figure 1.

We then present a table of the most closely related studies organized by our framework in Table 1. Please note that we only report upon gait or gait-like classification with loosely carried smartphones. This excludes work in detecting the device's carrying location [22] or the use of gait as a biometric feature for identifying the user's identity [13, 28]. We expect that our framework will be applicable to mobile classification work in general, but charting that terrain is left to future work.

STRUCTURING MOBILE CLASSIFICATION

Our framework consists of three spaces: an *application space* that describes the overall goal and major problem description, an *implementation space* of classification approaches and

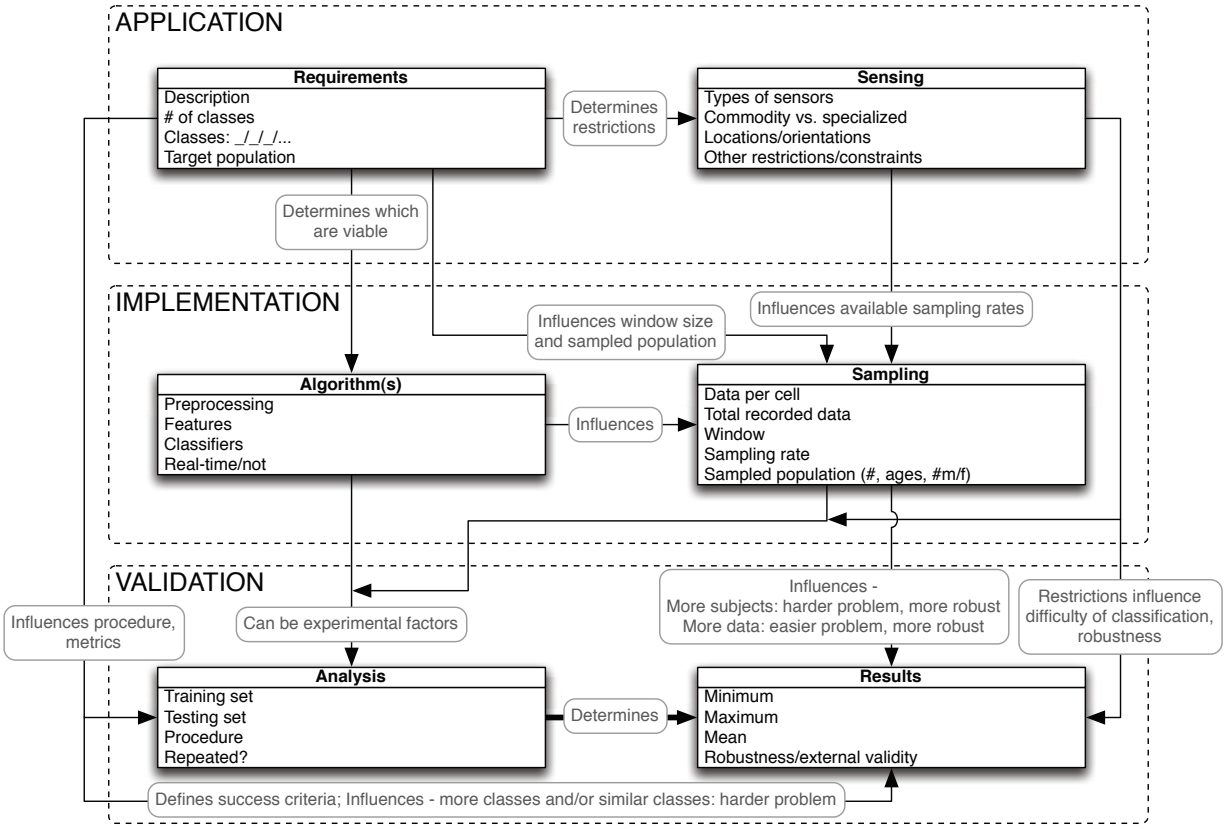


Figure 1: Framework for mobile context-aware classification work.

data collected, and a *validation space* of evaluation methodologies and results. We now describe the three spaces and six dimensions in more detail.

Figure 1 shows an explanatory diagram of this framework and the relationships between its entities. Each space is represented as a dashed box. Dimensions within each space are solid boxes (the term "dimension" is loosely defined; each is actually a collection of closely-related dimensions, but it is more useful to work with a lower-dimensional space).

Relationships between dimensions are represented with directional arrows, describing the influence of one dimension upon another. For example, when developing an interactive exercise game (described below under the Requirements dimension), this limits possible algorithms to those that can perform in real-time (influencing the Algorithm(s) dimension), which in turn restricts the available sample rate and window size, as real-time applications might require that the system reacts within a specified time period (influencing the Sampling dimension). As such, the framework can also be used as a flow chart to develop a study design – start by describing the requirements and follow the arrows.

Application Space

Classification research generally operates within a certain context to achieve a specific goal, *e.g.*, discriminating different gaits to track activity. This context can impose constraints that make classification more or less feasible. Two dimensions compose this space: Requirements and Sensing.

Requirements – The application space describes the task at hand. Is it to classify gaits, recognize a user, or detect a mode of travel? This dimension includes the number of classes and what they are. For example, Zhang et al.'s study classifies user gait and posture, recognizing 6 gaits: *walking*, *posture-transition*, *gentlemotion*, *standing*, *sitting*, and *lying* [35]. As well, the target user population are an important designation of the requirements. An exercise game designed for children will have different needs than an ambulatory monitoring system for physiotherapy in the elderly.

Sensing – What sensors are available to detect gaits, and what restrictions can we impose upon our users to facilitate sensing? Many studies seeking to use smartphones to infer context would like their system to work however the device is carried, but this is not always practical. This dimension makes chosen locations and other constraints explicit; this includes where the device is being carried or held, whether it is loose or fixed, and whether specialized equipment or commercially-available commodity hardware is required. For example, Huynh et al. use a sensor fixed to a backpack strap (a clear constraint) [15], but Kawahara, Kurasawa, & Morikawa allowed subjects carry commodity smartphones loosely in their pants pockets, chest pockets, and a personal bag [19].

	Schneider 2013	Khan 2010	Zhang 2010	Yang 2009	Kawahara 2007	Iso 2006	Huynh 2005
Requirements							
Task	Gait classification	Activity monitoring	Activity detection	Activity detection	Gait, device location classification	Activity detection	Activity detection
Target Population	Children, adults	Elderly	General	General	General	General	General
# of Classes	15	7	6	6	3-4	5	6
Classes	walk, fastwalk, jog, run, up/downstairs, walksideways/ backwards, narrow, penguin, twofoothop, toes, heels, knees, kickbum	walk, run, up/downstairs, cycle, vacuum, rest	walk, posture transition, gentle motion, stand, sit, lie	walk, run, drive, cycle, sit, stand	walk, run, sit, stand, *(sit & stand sometimes combined)*	walk, fastwalk, run, up/downstairs	walk, jog, skip, hop, ride bus, stand
Sensing							
Types of Sensors	Smartphone (Android)	Witilt 2.5 Sensor	Smartphone	Smartphone	Smartphone	Cellphone	Integrated sensor board
Specialized?	Commodity	Commodity	Commodity	Commodity	Commodity	Commodity	Specialized
Locations & Orientations	Front/back pockets, belt, arm, backpack, hand *(freely placed in each)*	Front left/front right/ rear trouser pockets, inner chest/jacket pockets	Belt *(horizontal orientation)*	Freely carried	Pants/chest pocket, bag	Breast/hip pocket	Affixed to backpack strap, consistent location
Algorithm(s)							
Preprocessing	*None*	Moving average	Kalman filter	Estimate gravity, split (x,y,z) into vertical and horizontal components	*None*	Pseudo-data created by rotating original signals	*None*
Features	Statistical features from time and frequency domains of X, Y, and Z axes	Spectral entropy, linear discriminant analysis of autoregressive coefficients, signal magnitude area	Accelerometer values	Several feature sets (factor in experiment); common statistical observations, cross correlation function	Average, variance, strongest frequency component, change in angle	Wavelet packet decomposition of periodogram, momentum from best basis	Statistical features on both time-domain and frequency-domain, light sensor, digital compass
Classifiers	Random forest with 500 trees	Hierarchical neural networks (resting, lower activity, upper activity; then activity)	Threshold rules to detect motion, SVMs for activity given motion/non-motion	Decision tree, naïve Bayes, kNN, SVM	Rule-based: detect orientation/location, then detect posture	Kohonen self-organizing map (KSOM)	K-means clustering for training, nearest cluster centroid for classification
Real-Time?	Yes	Yes	No	Yes	Yes	Yes	Yes
Sampling							
Total Recorded Data	9 hrs	24 hrs	2 hrs	192 min	~40 min	7.5 hrs	~200 min
Data Per Cell	28-42 seconds	2 to 8 minutes	*(not reported)*	10-30 minutes	~2.5 minutes	15 minutes	*(not reported)*
Window & Sampling Rate	2 seconds, ~26 Hz (variable sampling rate)	90 samples, 90 Hz	Segments of over 10s; sampling rate of 1 Hz	10 seconds, ~36 Hz	20 Hz, window varies with feature	3 seconds	0.25, 0.5, 1, 2, 4 seconds at 512 Hz
Sampled Population	12 subjects (6m/6f) Ages: 21-31 yrs Weights: 46-86 kg Heights: 155-183 cm	8 subjects (6m/2f) Age: mean 65, sd 3 yrs	10 subjects (7m/3f) Ages: 23-50	4 subjects	4 subjects	2 subjects	2 subjects
Analysis							
Training Set	For each participant, a randomly selected partition of 9 folds	60% of each subject (randomly selected)	One subject	3 subjects	3 subjects	*(not reported)*	800 cluster-means (4 of 5 folds of 1000 cluster-means)
Testing Set	For each participant, the remaining fold	40% of each subject (remaining)	All ten subjects (including training subject)	1 subject	4 subjects	*(not reported)*	remaining fold (200 cluster-means)
Procedure	10-fold cross validation	Train on several subject training sets, test on remaining subject testing sets	Train on one, test on all	10-fold cross-validation	Accuracy of gaits given device position	*(not reported)*	Cross validation
Repeated?	For each participant, 10 times (once per fold)	Until all subjects were a part of the testing set	No	For each feature set and classifier	No	*(not reported)*	5 times (one per fold)
Metric	Mean accuracy	Accuracy	Mean accuracy between motion and non-motion	Accuracy	Accuracy	Accuracy	Recall/(1-Precision) curves
Results							
Mean	78.1% - 92.2% (depending on gaitset and location)	94.4%	94.4%	Depends on classifier and feature set	N/A	~80%	N/A
Range/variation	68.1% - 100.0%	87% - 98%	87% - 98%	66.3% - 90.6%	96.7% - 100.0%	~73% - ~95%	N/A

Table 1: Map of related work, structured according to the framework presented in Figure 1. The new study reported in this paper is listed as 'Schneider 2013'.

Implementation Space

The Implementation Space describes the details of the solution, answering the question "How?" It has two dimensions: Algorithm(s) and Sampling.

Algorithm(s) – In this dimension, one describes both classification schemes as well as any preprocessing steps or unique notes about feature extraction. For example, Iso and Yamazaki generate pseudo-data by transforming their collected data rotationally [16]. This means that they may have some bias in their classifier (from similar patterns being artificially repeated), but they also making their classifier more robust to orientation, an important fact to consider when comparing different work. Because algorithms can become quite complex, we reserve this dimension to discuss high-level features, and direct readers to the papers themselves for details.

Sampling – The sampling dimension is the specific manifestation of a dataset for validation of the algorithm. As the amount of data collected has a large bearing on the results, it is important to describe both the overall amount of data collected and the amount per data cell – that is, how much data exists for each experimental factor (*e.g.*, gait/subject/location). If there is only a small amount of data in a data cell, then evaluation of the classifier may give lower results than if more data was collected. We also represent the relationship between time and number of observations in this dimension, by giving the sample rate and the window size (the amount of time or number of samples taken to constitute a sample used in classification. For example, Yang collects 10 to 30 minutes per data cell for a total of 192 minutes, and takes one observation from 10 seconds of data (with an accelerometer sampling rate of approximately 36 Hz) [34], while Khan et al. collect 2 to 8 minutes per data cell for a total of 24 hours of data, sampling at 90Hz with a window of 90 samples (1 second) [21]. Khan et al. clearly have more data, but are trying to classify observations with less data individually (1 second or 90 samples vs. 10 seconds or 36x10=360 samples), and less time spent with each gait or location. We include both sampling rate and window size to relate the number of observations to the amount of time.

Validation Space

Every technique must be evaluated. In this space, the dimensions of Analysis and Results have the most influence on the conclusions of a given study. As analysis necessarily varies with application area, generalization and comparison are difficult. However, this can be mitigated by clearly articulating both dimensions to provide context.

Analysis – The metrics that researchers use and the way that they compare results is extremely important. Most tend to use accuracy (the percent of correctly classified observations) as their metric, but classification techniques have several other ways to measure success, such as recall, precision, and F-measure. Some analyses might train on a single participant and test on several others (a difficult problem), while others might train on a subject and test on that same subject (which is expected to be easier).

Results – The actual metrics found during analysis. To facilitate comparison and access by algorithm consumers, we urge researchers to not just list an overall score (mean, etc.) but to also report the minimum, maximum, and mean. There is often a large variance of classification rates that is not captured in the mean alone.

Positioning Our Study

The goal of our user study is to explore previously unexamined areas with smartphone-based gait classification. Here, we demonstrate how the framework might be used by positioning our study within the established literature.

Looking at the Requirements dimension of Table 1, we see that previous studies have typically explored 5-7 gaits or activities, looking at common postures or modes of transport. We expand this to 15 gaits, looking at several atypical gaits suitable for exercise games. Along the Sensing dimension, we similarly examine more locations (6) than previously explored (5, [21]). With both of these dimensions, we explore a more challenging problem than does previous work.

In Algorithm(s) dimension, each study has its own algorithm, with some overlap (*e.g.*, SVM); ours introduces Random Forest. In this we are not necessarily introducing a challenge, but it illustrates how the framework can be used for a quick survey of past algorithm use and overlap between studies. In the Sampling dimension, a quick comparison establishes context for our study's results: the largest number of subjects and smallest reported length of data per data cell, a typical-to-large (but not largest) total amount of recorded data, and the use of a lower (but not lowest) sampling rate.

The Analysis dimension concisely describes how the final results were achieved, and highlights any difference from other studies. For example, in our evaluation, we both train and test our algorithm on each participant (effectively personalizing the algorithm, although we report non-personalized data) while other analyses might have other schemes, *e.g.*, training on one participant and testing on others. Finally, after examining this context, we use the Results dimension to compare how the different algorithms might perform given the analysis and dataset.

CLASSIFICATION STUDY

In order to inform gait as input for persuasive smartphone applications, in particular encouraging physical activity, we conducted a user study of novel gaits. In particular, we aim to explore rich and robust sensing by investigating more gaits and more carrying locations than in previous work. In essence, we pursued the most difficult manifestation of this classification problem to date, in order to identify baseline capabilities for mobile classification (which has been successful so far). We drew from 4 primary goals in this area: (a) to explore subtle differences for context-aware applications, investigating gaits that were similar (such as distinguishing jogging from rapid walking); (b) to explore novel gaits suitable for exercise applications, by involving different muscle groups; (c) to explore novel gaits that could be easily linked by metaphor to a child-friendly exercise game, such as marching like a soldier or walking on a tight rope; and (d) to connect

Category	Label	Verbal description
Pedestrian	walk	Walk normally
	fastwalk	Walk quickly
	jog	Jog, *if asked for clarification:* slow run
	run	Run, not necessarily a flat out sprint, but a run
	ascendstairs	Walk up the stairs as you normally would until you reach the top
	descendstairs	Walk down the stairs as you normally would until you reach the bottom
Exercise	toes	Walk on your toes
	heels	Walk on your heels
	liftknees	Walk while lifting your knees high in the air
	kickingbum	Walk while kicking your bum. You don't actually have to hit it, but do that motion
Game	twofoothop	Hop with two feet
	penguin	Walk like a penguin
	narrow	Walk as if on something narrow, like a tightrope
	walkbackwards	Walk backwards
	walksideways	Walk sideways

Table 2: Gaits and descriptions used in our data collection study.

modes of travel, such as cycling or driving. These four goals informed the generation of our gaits. As such, our target gaits were divided into *gaitsets*, sets of gaits grouped by their relationship to our four goals.

Gaits

With this in mind, we developed a set of gaits suitable for an exercise or context-aware game or application. After piloting, we retained 15 gaits organized into three gaitsets (Pedestrian, Exercise, and Game) corresponding to our first three goals; for scope and logistical reasons, our fourth goal of travel methods (the would-be Locomotion gaitset) is left to future work. We discovered during initial brainstorming that some gaits were consistently interpreted by participants (*e.g.*, walking on heels, walking like a penguin) but others were not (*e.g.*, skipping, walking like a zombie). To facilitate our user study and allow for a robust application, we decided to only use gaits that were consistently interpreted by a simple verbal or written description. To limit scope, all gaits had to involve directional movement and be bipedal - that is, we did not allow in-place activities like jumping jacks or walking on hands and knees. Our gaits are described in Table 2.

Gaits not involving stairs were performed outside in a random order. For logistical reasons, Ascend Stairs and Descend Stairs were performed either at the beginning or end of the trial; this was counterbalanced by participants, as was the order of Ascend Stairs and Descend Stairs.

Apparatus and Participants

Seven Samsung Galaxy Nexus smartphones running Android OS 4.0.1 (Ice Cream Sandwich) were used in the experiment, each loaded with a custom accelerometer logging application. Six were placed on the participant in the following locations: front pocket, back pocket, hip, hand, arm, and backpack. These six phones sampled the accelerometer at the highest rate allowed by the operating system (mean sampling period was 16.95 ms, standard deviation 28.75 ms, median 8.3 ms),

	Age (years)	Height (cm)	Weight (kg)
Minimum	21	155	46
Median	26	171.5	69.5
Mean	25.3	171.5	68
St. Dev.	2.96	9.71	14.14
Maximum	31	183	86

Table 3: Self-reported statistics from 12 participants.

and saved the sampled signals to a comma-separated value (CSV) file retrieved after the experiment. The final phone was held by the experimenter, and was used to record the start time and end time of each gait. The logging phones were synchronized with the experimenter's phone at the beginning of the experiment by having the experimenter simultaneously press buttons on the experimenter's phone and the logging phone. Analysis was conducted on a MacBook Pro laptop with a 2.7 GHz Intel i7 processor and 8 GB of RAM.

12 participants (6 female, 1 left-handed) are hereafter referred to as P1 to P12. Table 3 reports summary statistics for self-reported age, weight, and height.

Algorithm

For our classification scheme, we use Random Forest, an algorithm that uses bagging with randomly selected subsets of features, constructing a decision tree for each bootstrap sample. Decisions are made by majority response of these trees. As the number of trees approaches infinity, the accuracy of a Random Forest classifier converges [4]; through piloting we found accuracy converged between 100 and 500 trees, and so use 500 trees. We used the implementation provided by Weka [14].

We used a 2-second window with 50% overlap to compensate for our small amount of data; this yielded a total of 35806 instances for our entire dataset. Our classification scheme and window were chosen as the best performing in a pilot

study examining window sizes (2- and 4-second), classification schemes (SVM, Naïve Bayes, Multilayered Perceptron, and Random Forest), and axis-grouping (x/y/z, vertical/horizontal [34], and vector magnitude). We hereafter refer to our algorithm as "RF500".

Features are extracted using the Python programming language and Scipy software package [18]. Features were chosen to represent a large number of basic statistical observations of both time and frequency domain data: our gaits can have subtle differences. To address the variable sampling rate found in smartphones, we calculate the frequency spectrum of each window with the Fast Lombe-Scargle Periodogram (FASPER) algorithm [27].

For each axis (x, y, z), we take the following features: minimum, maximum, and mean values, variance, skewness, kurtosis, 25th percentile, median (50th percentile), 75th percentile, a ten-bin histogram (normalized to have the proportion of each value in each bin), the most powerful and least powerful spectral frequencies, a weighted average of spectral frequencies by spectral power, spectral variance, spectral entropy, and a ten-bin histogram of spectral powers (normalized to have the proportion of each value in each bin). In addition, we look at the Pearson correlation and corresponding p-value for each pairwise combination of the x, y, and z axes. Finally, we take the signal magnitude area (sum of the Euclidean or ℓ-2 norm of every x,y,z tuple).

A full analysis of feature quality is beyond our present scope. However, we observed no trends in feature contribution to results. We suspect that the Random Forest algorithm's performance might be influenced by an ability to examine subtle trends across all features.

We could not train the All Locations/All Gaits/All Participants data cell with RF500, as the 8GB of laptop memory was in sufficient for this dataset. This dataset is thus missing from our all-participants analysis. We use the equivalent algorithm with 100 trees ("RF100") for our confusion matrix involving all gaits (Figure 3); it is expected to perform similar to but slightly worse than RF500. These computational costs are only present when training the algorithm; once trained, RF500 can perform efficiently on mobile devices.

Results

We conducted our main algorithm analysis with location, gaitset, and participant as factors. That is, we examine the benefit of knowing them by training classifiers that assume specific levels (such as an algorithm that is trained only on Exercise gaits with Front Pocket as the location).

We conduct this analysis twice, once training and testing on all 12 participants ("All-Participant"), and once training and testing on each participant ("By-Participant"). For each location, gaitset, and (for by-participant analysis) participant, we conducted 10 iterations of a 10-fold cross-validation. Our results analyze the mean of each cross validation, giving 10 data points for each location/gaitset(/participant) combination. For both all-participant and by-participant analysis we planned an Analysis of Variance (ANOVA) to compare different factors. However, the Shapiro-Wilk test of normality failed with a 5% level of significance on data cell residuals in both analyses (Exercise/Arm data cell in all-participant, W=0.83, p=0.037; 55 of 336 data sets in by-participant). We thus do not detect statistical effects for the three factors (location, gaitset, and participant), and could not conduct analysis through planned contrasts.

Gaitset	Location	Mean % correct
Exercise	Back Pocket	87.2
Exercise	Front Pocket	86.6
Exercise	Arm	85.3
Game	Front Pocket	85.0
Exercise	Belt	84.8
Exercise	Backpack	84.6
Game	Back Pocket	84.5
Pedestrian	Front Pocket	83.7
Exercise	Hand	83.4
Pedestrian	Back Pocket	82.6
Game	Arm	81.7
All Gaits	Front Pocket	80.1
Game	Backpack	79.9
All Gaits	Back Pocket	79.8
Game	Belt	79.6
Pedestrian	Belt	79.5
Game	Hand	78.9
Pedestrian	Backpack	78.9
Pedestrian	Arm	78.4
Pedestrian	Hand	77.2
All Gaits	Arm	74.5
All Gaits	Belt	74.4
All Gaits	Backpack	74.3
All Gaits	Hand	73.5

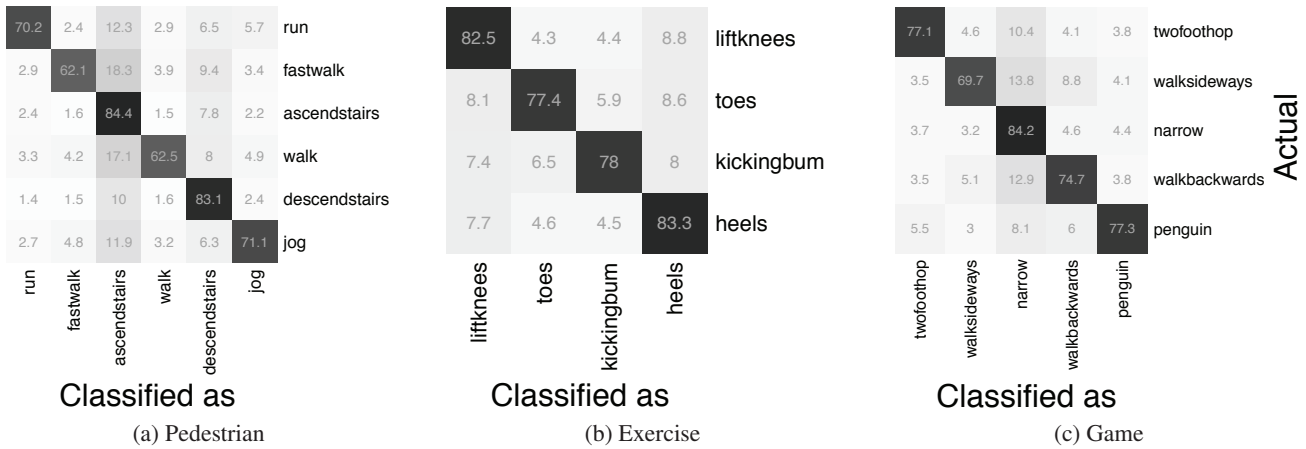
Table 4: All-participant means of gaitset/location accuracy. Chance ranges from $6.6\bar{6}\%$ ("All Gaits") to 25% ("Exercise").

All-Participants

Confusion matrices have been produced with all locations across all-participants (*e.g.*, without personalization) for each gaitset in Figure 2, and for all gaits in Figure 3. Table 4 shows mean classification rates by gaitset and location.

By-Participant

Training the RF500 algorithm on each participant improves results with the short data collection time of 30s per gait and location, and reveals variability in individual differences; see Figure 4 for graphs of the effect of participants by gaitset and location. The top performing mean data classification rate was 100.0%, present in three data cells (involving two participants): Game/Front Pocket/P3, Game/Backpack/P9, and Exercise/Backpack/P3. The worst performing mean data classification rate was All Gaits/All Locations/P10 with a mean of 63.2% classification rate. Table 5 shows mean classification rates by gaitset and location. Due to space considerations, we only report the mean of each personalized Gaitset/Location result over all participants; as such, the extreme values of 63.2% and 100.0% do not appear in any tables.



(a) Pedestrian (b) Exercise (c) Game

Figure 2: All-participant confusion matrices over all carrying locations for each gaitset (distinguishing only between gaits in the gaitset). Darker squares represent a higher classification rate. Numbers presented are percents of classification rates for each actual gait (rows sum to 100%).

Gaitset	Location	Mean % correct
Exercise	Belt	92.2
Exercise	Hand	91.0
Exercise	Arm	90.1
Exercise	Back Pocket	89.8
Exercise	Front Pocket	89.7
Exercise	Backpack	88.7
Game	Back Pocket	88.6
Game	Front Pocket	88.2
Pedestrian	Hand	87.9
Game	Hand	87.5
Game	Arm	87.4
Exercise	All Locations	87.3
Pedestrian	Front Pocket	87.0
Game	Belt	87.0
Pedestrian	Belt	86.7
Game	Backpack	86.3
Pedestrian	Back Pocket	86.0
Pedestrian	Arm	85.6
Pedestrian	Backpack	84.7
Game	All Locations	84.1
All	Front Pocket	83.0
Pedestrian	All Locations	82.7
All	Hand	82.3
All	Back Pocket	82.1
All	Belt	81.6
All	Arm	80.7
All	Backpack	80.1
All	All Locations	78.1

Table 5: By-participant means of gaitset/location accuracy rate. "All Gaits" results are highlighted in grey. Chance ranges from $6.6\bar{6}\%$ ("All Locations") to 25% ("Exercise").

DISCUSSION

Overall, the classification scheme performed with a varying range of success. In this section, we list our major conclusions to inform designers. We make an effort to compare our results with previous work whenever possible; please refer to our map (Table 1) for additional context.

Overall success

When trained and tested over all participants, accuracy ranged from 73.5% (All Gaits/Hand) to 87.2% (Exercise/Back Pocket), although we notably must exclude All Gaits/All Locations (due to limitations of computational resources), expected to be the worst performer. All are far larger than chance.

Overall, accuracy improved by about 5% when the algorithm was personalized (by-participant analysis). Although we cannot confirm statistical significance of this effect, it suggests that a short personalization session might improve classification rates. Even with personalization, though, there was a great deal of variation in accuracy: mean classification rates of each Gaitset/Location/Participant had extremes of 63.2% (minimum) to 100.0% (maximum). All classification rates are well above chance (the lowest classification rate, 63.2% compares to $6.6\bar{6}\%$ for all 15 gaits). For all locations and all gaits, after training on an individual for 30 seconds for each gait, the mean classification rate across all 12 participants was 78.1%.

For post-hoc analysis (analyzing a day's data for overall activity after-the-fact) we expect that both personalized and non-personalized algorithms (that is, trained on the user's data or not) could be sufficient for extremely high classification rates: Kunze et al. used a majority vote of several 1 second windows with 82% classification rate in 1 or more minute sequences to achieve 100% classification rate of walking vs not walking [22]. For real-time applications, a classification rate of 78.1% could be used with novel application design

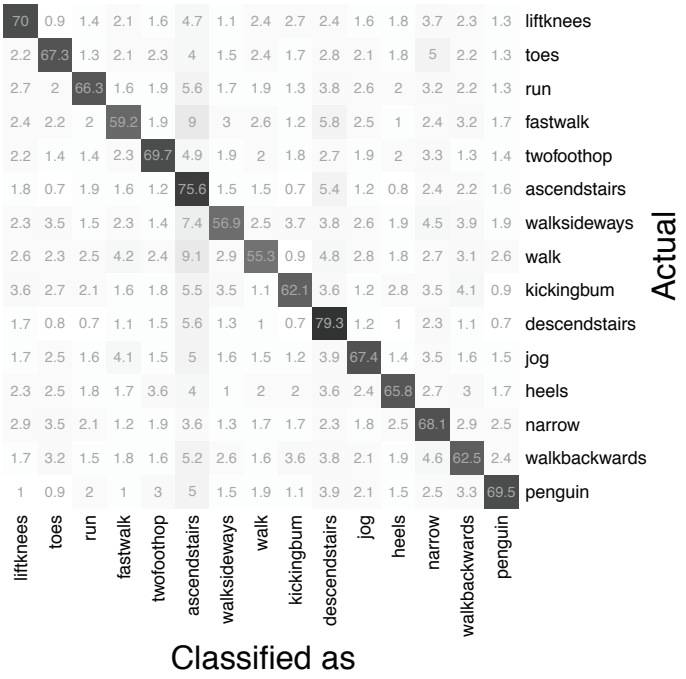


Figure 3: All-participants confusion matrix over all gaits and all carrying locations. Darker squares represent a higher classification rate. Numbers presented are percents of classification rates for each actual gait (rows sum to 100%).

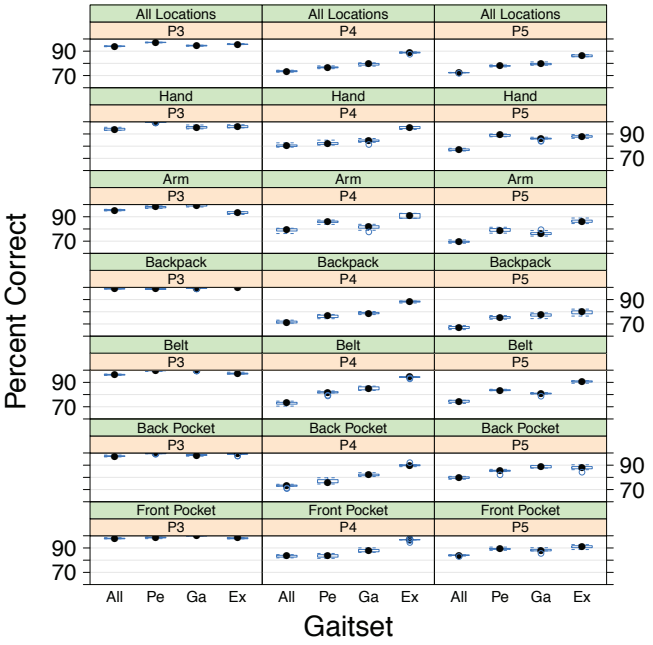


Figure 4: Bar-and-whiskers plot of gaitset performance for each of participants 3, 4, and 5 (chosen for illustrative purposes; P3's results are indicative of the high-performing participants P2, P3, and P9, while P4 and P5 are representative of the interactions present in the other 9 participants). Note the 3-way interactions between gaitset, location, and participant.

or game mechanics to improve perceived recognition rates. We also note that many of our gaits are very similar, and that careful pruning of the selection of gait could improve performance: our Exercise gaitset had a mean classification rate of 87.3% without knowledge of location when personalized, and achieved a mean classification rate of 92.2% when it was known to be mounted on the user's belt. Thus, our results are promising for robust classification of a wide variety of gaits.

Gait is a complex ecosystem of gaitset, location, and user

In Figure 4, we note interactions between location, gaitset, and participant. Different locations perform well with different gaitsets, and these vary depending on participant. With P3, we see a flat pattern across gaitsets and location — basically, the algorithm performed well for P3 regardless of location and gaitset. This pattern was observed for three participants: P2, P3, and P9; in fact, the 18 best performing data cells were from P3 and P9. No distinguishing demographic features stand out for these participants: both sexes are represented, heights and weights vary (155-183cm and 51-64kg respectively), and no notable behaviours were observed. This suggests that these gaits may work very effectively for certain individuals with minimal personalization (30 seconds per gait and location). If this is from gait interpretation, then it is possible that, if instruction is given to users, a wider variety of gaits could be effectively classified.

However, with other participants we see more variation (P4 and P5 are representative of the other 9 participants). Accuracy varies by participants and gaitsets when location is fixed. When the smartphone is carried in the hand, all gaitsets except for Exercise perform similarly for P4, but all gaitsets except for All Gaits perform similarly for P5. With different locations, more interactions appear; performance of all four gaitsets is similar for both P4 and P5 (but quite different from P3). We conclude that the performance of RF500 depends upon location, gaitset, and user, with intricate interactions.

In aggregate, Exercise does best and All Gaits worst

In general, the best performing gaitset tended to be Exercise, and the worst was the expected laggard, All Gaits. In the all-participant analysis, the top three performing data cells are Exercise data sets, and the worst four are from All Gaits (Table 4). In the by-participant analysis, the top 6 performing algorithms are Exercise sets, and the worst 6 performing algorithms are from All Gaits (Table 5). In both all-participant and by-participant analyses, the best performer for every location was Exercise, and the worst for every location was All Gaits. A close look at the by-participant results (Figure 4) shows that these results might change dramatically depending on the user even when the algorithm is personalized.

Gait similarity affects classification performance

Though there could be several reasons for the difference in performance of different gaitsets, such as the number of gaits

within each gaitset, our findings suggest that the *types* of gaits chosen within each gaitset are an important factor. Major patterns in the all gaits confusion matrix follow, using "→" to indicate mis-classification. Mis-classification is not necessarily bidirectional.

As we can see in Figure 2, Exercise's gaits have fewer standout confusions with each other than the gaits in Pedestrian and Game, which is not surprising considering that Exercise is the best performing gaitset. Pedestrian has a number of confusions that suggest differences between these gaits are quite subtle, such as *walk* → *fastwalk* and *jog* → *fastwalk*. This is an expected result, and is consistent with the design of the Pedestrian gaitset to include very similar gaits. For the Pedestrian gaitset, then, we note that distinguishing between different categories of interpreted speeds of walking is challenging, and it may be best to only classify walking and jogging or running, or refer to cadence directly in applications.

In the Game gaitset we notice a dichotomy – two groups of gaits. Most gaits were mis-classified as *narrow*, especially *walksideways* and *walkbackwards*. The least confused gaits were *narrow* → *walksideways* and *penguin* → *walksideways*. We thus suggest that *narrow*, *walksideways*, and *walkbackwards* are similar, but different from *twofoothop* and *penguin* (which, unexpectedly, had a strong confusion with *penguin* → *twofoothop*). Ultimately, many of these gaits are similar to normal walking, and those that differ were more easily recognizable.

Pedestrian shows a number of confusions. As in the All Gaits confusion matrix (Figure 3), many were mis-classified as *ascendstairs*, and several were mis-classified as *jog*. Both *stairs* were often correctly distinguished from *jog* and *run*. Few gaits were confused with *run* and *walk*. Exercise demonstrates few standout confusions between its gaits. The most confused gait was *heels*, in that it has the darkest column excluding the diagonal element, followed by *liftknees*. The strongest individual confusion is *liftknees* → *heels*. *Toes* and *kickingbum* were confused with each other more than with the other two gaits.

Individual differences strongly influence gait recognition

Individual differences add an important dimension to the relationship between location and gaitset. This wide variation in individual differences for activity recognition is consistent with literature (Zhang et al. report 69% to 95.3% with their best classifier [35], Kunze et al. report 72% to 93% with a binary classifier [22]), but we are not aware of any previous work that has commented on or explored individual differences and their impact on gait. Furthermore, previous work has used a non-personalized algorithm for all participants; we demonstrate this effect *even when it is personalized via training on each participant, location, and gaitset*, suggesting that it is not merely error from the classifier, but could be an intrinsic component of each participant's gaits.

We thus suggest that future work attempting gait classification with gaits that might be susceptible to individual differences, such as the interpretation of unusual gaits or gaits with subtle differences, must accommodate these differences. We aim to pursue this in our future work.

CONCLUSION

In this work, we have given an argument for the role of gait classification in persuasive applications. We have identified a need for better comparison between different studies in this area, and provided a framework for comparison based differences in application area, implementation details, and validation methodology. We have used this framework to provide a map of relevant classification studies to help guide readers through the literature. Finally, we have provided our own contribution to real-time gait classification, looking at the difficult problem of classifying 15 novel, similar gaits considering 6 different carrying locations with only 30 seconds of data collected for each data cell.

There are limitations to our study that must be considered for interpretation and future work. As revealed by our framework, we have only a single algorithm, and a limited amount of data per data cell. More data could improve results, or make them more trustworthy. As well, there are two confounded variables in our study. First, our data was only collected once from each participant - individual differences could be a result of clothing or the phone orientation in addition to personal gait characteristics. Second, gaitsets do not have a consistent number of gaits: Pedestrian has 6 gaits, Game has 5, and Exercise has 4. We leave the investigation of these variables to future work. Finally, although we approach the problem of gait classification by exploring the limits of sensing, we suggest another angle for future work: examining how robust and accurate classification algorithms must be for use in persuasive mobile applications.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments and our participants for their time and effort. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and received support from the Institute for Computing, Information and Cognitive Systems (ICICS) at UBC.

REFERENCES

1. Altun, K., Barshan, B., and Tunçel, O. Comparative study on classifying human activities with miniature inertial and magnetic sensors. *Pattern Recognition 43*, 10 (Oct. 2010), 3605–3620.
2. Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D., Benford, S., Capra, M., and Hampshire, A. Interweaving mobile games with everyday life. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, no. April in CHI '06, ACM Press (2006), 417–426.
3. Brehmer, M., Graham, T. C. N., and Stach, T. Activate Your GAIM : A Toolkit for Input in Active Games. In *ACM FuturePlay 2010* (2010), 151–158.
4. Breiman, L. Random Forests. *Machine Learning 45* (2001), 5–32.

5. Buttussi, F., and Chittaro, L. Smarter Phones for Healthier Lifestyles: An Adaptive Fitness Game. *Pervasive Computing 9*, 4 (2010), 51–57.

6. Consolvo, S., Everitt, K., Smith, I., and Landay, J. A. Design requirements for technologies that encourage physical activity. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*, ACM Press (New York, New York, USA, Apr. 2006), 457–466.

7. Consolvo, S., Klasnja, P., McDonald, D. W., Avrahami, D., Froehlich, J., LeGrand, L., Libby, R., Mosher, K., and Landay, J. A. Flowers or a Robot Army? Encouraging Awareness & Activity with Personal, Mobile Displays. In *Proceedings of the 10th international conference on Ubiquitous computing - UbiComp '08*, ACM Press (New York, New York, USA, Sept. 2008), 54–63.

8. Davis, S. B., Moar, M., Jacobs, R., Watkins, M., Riddoch, C., and Cooke, K. Ere Be Dragons : heartfelt gaming. *Digital Creativity 17*, 3 (Jan. 2006), 157–162.

9. Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., and Landay, J. A. UbiGreen: investigating a mobile tool for tracking and supporting green transportation habits. In *Proceedings of the 27th international conference on Human factors in computing systems (CHI 2009)*, D. R. Olsen, Ed., vol. 09 of *CHI '09*, ACM SIGCHI, ACM (2009), 1043–1052.

10. Fujiki, Y. iPhone as a physical activity measurement platform. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, ACM Press (New York, New York, USA, 2010), 4315–4320.

11. Fujiki, Y., Kazakos, K., Puri, C., Buddharaju, P., Pavlidis, I., and Levine, J. NEAT-o-Games: Blending Physical Activity and Fun in the Daily Routine. *Computers in Entertainment 6*, 2 (July 2008), Article 21.

12. Fujiki, Y., Kazakos, K., Puri, C., Pavlidis, I., Starren, J., and Levine, J. NEAT-o-Games: Ubiquitous Activity-based Gaming. In *CHI '07 extended abstracts on Human factors in computing systems - CHI '07*, ACM Press (New York, New York, USA, Apr. 2007), 2369–2374.

13. Gafurov, D., and Bours, P. Improved Hip-Based Individual Recognition Using Wearable Motion Recording Sensor. In *Security Technology, Disaster Recovery and Business Continuity*, T.-h. Kim, W.-c. Fang, M. K. Khan, K. P. Arnett, H.-j. Kang, and D. Ślzak, Eds., vol. 122 of *Communications in Computer and Information Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, 179–186.

14. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter 11*, 1 (2009), 10–18.

15. Huynh, T., and Schiele, B. Analyzing features for activity recognition. In *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence innovative context-aware services: usages and technologies - sOc-EUSAI '05*, ACM Press (New York, New York, USA, Oct. 2005), 159–164.

16. Iso, T., and Yamazaki, K. Gait analyzer based on a cell phone with a single three-axis accelerometer. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services - MobileHCI '06*, ACM Press (New York, New York, USA, Sept. 2006), 141–144.

17. Jensen, K. L. v., Krishnasamy, R., and Selvadurai, V. Studying PH. A. N. T. O. M. in the wild. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction - OZCHI '10*, ACM Press (New York, New York, USA, Nov. 2010), 17–20.

18. Jones, E., Oliphant, T., Peterson, P., and Others. Scipy: Open source scientific tools for Python, 2012.

19. Kawahara, Y., Kurasawa, H., and Morikawa, H. Recognizing User Context Using Mobile Handsets with Acceleration Sensors. In *2007 IEEE International Conference on Portable Information Devices*, IEEE (May 2007), 1–5.

20. Kazakos, K., Fujiki, Y., Pavlidis, I., Bourlai, T., and Levine, J. NEAT-o-Games: Novel Mobile Gaming Versus Modern Sedentary Lifestyle. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services - MobileHCI '08*, ACM Press (New York, New York, USA, Sept. 2008), 515–518.

21. Khan, A. M., Lee, Y.-K., Lee, S., and Kim, T.-S. Accelerometer's position independent physical activity recognition system for long-term activity monitoring in the elderly. *Medical & Biological Engineering & Computing 48*, 12 (Dec. 2010), 1271–1279.

22. Kunze, K., Lukowicz, P., Junker, H., and Tr, G. Where am I : Recognizing On-body Positions of Wearable Sensors. In *LOCA05: International Workshop on Location and Context-Awareness*, vol. 3479 (2005), 264–275.

23. Lin, J., Mamykina, L., Lindtner, S., Delajoux, G., Strub, H., Dourish, P., and Friday, A. FishnSteps: Encouraging Physical Activity with an Interactive Computer Game. In *8th International Conference on Ubiquitous Computing (Ubicomp 2006)*, P. Dourish and A. Friday, Eds., vol. 4206 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg (Berlin, Heidelberg, 2006), 261–278.

24. MapMyFitness Inc. MapMyRun, 2012.

25. Nike Inc. Nike+ Project, 2012.

26. Oliver, N., and Flores-Mangas, F. MPTrain: a mobile, music and physiology-based personal trainer. In *8th*

International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06), M. Nieminen and M. Röykkee, Eds., ACM (2006), 21–28.

27. Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical recipes in C: The art of scientific computing: Second edition*. Cambridge University Press, 1992.
28. Sprager, S., and Zazula, D. A cumulant-based method for gait identification using accelerometer data with principal component analysis and support vector machine. *WSEAS Transactions on Signal Processing 5*, 11 (Nov. 2009), 369–378.
29. Stach, T., Graham, T. C. N., Brehmer, M., and Hollatz, A. Classifying Input for Active Games. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology - ACE '09*, ACM Press (New York, New York, USA, Oct. 2009), 379.
30. Stanley, K. G., Livingston, I., Bandurka, A., Kapiszka, R., and Mandryk, R. L. PiNiZoRo : A GPS-based Exercise Game for Families. In *Futureplay '10 Proceedings of the International Academic Conference on the Future of Game Design and Technology* (2010), 243–246.
31. Tudor-Locke, C. Taking Steps toward Increased Physical Activity: Using Pedometers To Measure and Motivate. *President's Council on Physical Fitness and Sports Research Digest 17*, 3 (2002), 3–10.
32. World Health Organization. WHO — Obesity and overweight, 2012.
33. World Health Organization. WHO — Physical Inactivity: A Global Public Health Problem, 2012.
34. Yang, J. Toward Physical Activity Diary: Motion Recognition Using Simple Acceleration Features with Mobile Phones. In *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics - IMCE '09*, ACM Press (New York, New York, USA, Oct. 2009), 1–10.
35. Zhang, S., Mccullagh, P., Nugent, C., and Zheng, H. Activity Monitoring Using a Smart Phones Accelerometer with Hierarchical Classification. In *6th International Conference on Intelligent Environments* (2010), 158–163.