Learning Characteristics of Smartphone Users from Accelerometer and Gyroscope Data

David Molay
dmolay@stanford.edu

Fan-Hal Koung
fanhal@stanford.edu

Kingston Tam
ktam@stanford.edu

Abstract

The presence of increasingly advanced sensors on modern smartphones provides the opportunity to leverage sensor data to enhance the user experience with biometric security features and user-targeted advertising. As a first step towards achieving this goal, we attempt to determine several key user traits based on smartphone accelerometer and gyroscrope data collected while the user walks with the device in his or her pants pocket. We chose to analyze data collected from the pants pocket because it is one of the most likely locations that a user would place the device. The traits we initially attempted to predict were weight, height and gender. However, we fortuitously discovered that we could identify individual users from their gait data with a high degree of accuracy. We approached predicting weight and height as classification problems (with the classes being small, medium and large) rather than regression problems because precision and recall are frequently much easier to interpret than mean squared error or mean absolute error.

To accomplish this task, we extracted and analyzed a variety of features from the sensor data. The features ranged from basic metrics, such as the mean and standard deviation of accelerometer readings, to the coefficients of a Fast Fourier Transform. By applying several classification methods, we were able to predict the weight, height, gender and identity of a smartphone user with leave-one-out cross validation accuracies of 57%, 56%, 83% and 96%, respectively.

1 Introduction

Smartphones are powerful computing devices owned by over 50% of American adults [1]. As devices that are almost always within reach, smartphones provide instant access to a wealth of content. Indeed, smartphones connect us to others, put the internet at our fingertips and even stream live video. The way we interact with smartphones is as interesting as the content they deliver. Smartphones are one of the few interactive devices that we frequently carry in a non-interactive manner. This made us wonder, while smartphones are clearly consummate content deliverers during periods of use, can they be capable content gatherers during periods of inactivity? In this paper, we investigate the application of machine learning techniques to determine the weight, height, gender and identity of a smartphone user while the device remains in its user's pocket.

There are a number of applications to which the weight, height, gender or identity of a smartphone user would be particularly useful. A straight-forward application would be to fitness trackers (e.g. estimating calories burned, metabolic rate, etc.), but there are others as well. For example, with this additional information, website and app advertisements could target users more effectively, potentially increasing the likelihood of a click-through. Further, when navigating to a clothing retailer's website, the smartphone user could automatically be directed to the appropriate section (men's or women's) where items have been pre-selected for the user based on his/her weight and height. Additionally, the ability to recognize individual users could enhance both the security and convenience of the device by sending an alert if an unauthorized user is carrying the phone or remaining unlocked if an authorized user is carrying the device. The rise of the smartphone as a powerful computing device generally carried on one's person provides the potential of broadening the use of accelerometer and gyroscope data beyond traditional gaming applications to improve the user experience in general.

2 Discussion of prior work

In recent years, there has been significant research on applying machine learning techniques to smartphone sensor data. Much of this research has focused on activity recognition [2, 3, 8] (e.g. determining if a smartphone user is standing, walking, running or biking) and, more recently, user identification [4, 5] (e.g. using the sensor data as a biometric). However, there has been considerably less research on using smartphone sensor data to identify general user characteristics such as weight, height and gender.

Our work on identifying individual users is most comparable to the wearable sensor-based recognition research discussed by Gafurov [12], while our work on identifying weight, height and gender is most similar to the research conducted by Weiss and Lockhart [6]. Weiss and Lockhart, however, focused only on classifying the extremes of the population (e.g. "short" v. "tall" and "light" v. "heavy"). They did not attempt to classify people of average weight and height and, thus, did not include examples of "average" people in their training or test sets. Additionally, Weiss and Lockhart used solely accelerometer data in their analysis whereas we incorporated gyroscope data as well.

3 The data

3.1 Overview

Our data set consists of a combination of smartphone gait data from two unique sources. Initially, we began prototyping our learning algorithms using data obtained from the Human Activity Sensing Consortium (HASC) [7]. This data set was

a unique fit for our purposes because it not only contained sensor data for a variety of activities (including walking), but also the weight, height and gender of each test subject. However, the HASC data was largely homogeneous. Indeed, after extensive preprocessing of the 2012 HASC corpus, 36 samples remained of which only five were female. The distributions of weight and height were similarly skewed. This lack of diversity prompted us to develop our own web application (`http://gyro.ktam.org/`) for additional data collection. We subsequently made recording one's gait data a task on Amazon Mechanical Turk for expedient data collection (we denote the resulting data set TURK). When collecting data we specified a specific orientation of the device (pointed downward in the user's front pant pockets with the screen facing the user's leg). For consistency, we also filtered the HASC data to only consist of gait data with the same device orientation. From our integrated data source we eliminated examples with fewer than fifteen seconds of gait data. We then considered only the middle ten seconds of the time series to remove the data that was collected as the device was being transferred to and from the user's pocket.

3.1.1 Predicting weight, height and gender

We wanted to include features from both the gyroscope and accelerometer data when predicting weight, height and gender. While the TURK data set and 2012 HASC corpus contain data from both the accelerometer and gyroscope sensors, HASC data from previous years contains only accelerometer data. Consequently, our integrated data set for predicting weight, height and gender consisted of 144 examples in total (108 examples from the TURK data and 36 from the 2012 HASC corpus).

3.1.2 Identifying individual users

The TURK data does not contain multiple instances of gait data from the same individual. Thus, we could not use this data to identify individual users. Accordingly, we decided to supplement the 2012 HASC corpus with the 2011 HASC corpus. The 2011 corpus, however, does not contain gyroscope data, so we only selected features from the accelerometer data when performing our analysis. When identifying users, we considered two orientations of the device: 1) the phone was mounted on the waist and 2) the phone was mounted on the waist or placed in the pant pocket. The former is a classification task with 75 unique individuals while the latter has 169 unique individuals.

3.2 Distribution of weight, height and gender

Even after supplementing the 2012 HASC corpus with the TURK data, the distribution of gender was still heavily skewed towards males. Of the 144 individuals, 107 were male (74%) and 37 were female (26%). We discretized weight and height into three classes each (small, medium and large) according to the following table:

	Small	Medium	Large
Weight (kg)	< 65	[65, 80)	≥ 80
Height (cm)	< 165	[165, 180)	≥ 180

We choose these cutoffs to be similar to those employed by Weiss and Lockhart [6], so our results would be able to be compared when we eliminate the medium class from our data and predict on only small and large individuals. This discretization resulted in a fairly even distribution of weight among the three classes (33.3% were considered small, 39.6% medium and 27.1% large). The discretiztion of height, on the other hand, was uneven with 27.7% considered small, 54.9% medium and 17.4% large. The discretization of height proved to be more difficult since its distribution in our data set was much tighter than that of weight. So there was a trade off between giving different labels to individuals that only differ in height by one or two centimeters, or having distinct classes that do not have an evenly distributed number of individuals in each class. Ultimately, we elected to have distinct classes as this provided the benefit of being able to compare our results to those of Weiss and Lockhart [6].

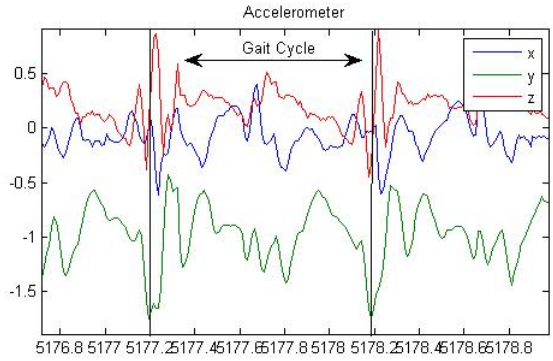

Figure 1: Plot of accelerometer data

4 Learning

4.1 Feature extraction

As a preprocessing step, we applied a median filter and a 3rd order low-pass Butterworth filter with a corner frequency of 20 Hz to remove noise from the accelerometer and gyroscope signal [8]. 20 Hz is an appropriate corner frequency since 99% of the energy associated with human gait is contained below 15 Hz [8]. We subsequently applied another low-pass Butterworth filter with a corner frequency of 0.3 Hz to separate the acceleration signal into body and gravity components as this has shown to be successful in prior work [8].

We then drew 49 features from the filtered accelerometer and gyroscope signals. These features were derived from the time and frequency domains and were selected largely on the basis of prior work in activity recognition and gait identification [8, 9]. A tabular summary of our time features can be found in Figure 2. Our frequency features were drawn from the total acceleration series and consist of DC component, signal energy, dominant frequency and coefficients sum.

Feature	Acceleration	Gyroscope
Mean	$X_{b,g,j}, Y_{b,g,j}, Z_{b,g,j}$,T	β_b
Std Dev	$X_{b,g,j}, Y_{b,g,j}, Z_{b,g,j}$	β_b
Dist btwn peaks	Y_b	β_b
Max point	X_b, Y_b, Z_b	β_b
Min point	Y_b	β_b
Mean square	X_b, Y_b, Z_b	β_b

Figure 2: Selected Time Domain Features. X, Y, and Z refer to the components of acceleration, T refers to the magnitude vector of acceleration, and β refers to the β rotational component. Subscripts b and g refer to the Body and Gravity components while j refers to the Jerk signal.

4.1.1 Time Domain

We first considered features in the time domain. Most of our features were drawn directly from filtered time series, but some were drawn from the derivative time series. These include the mean, standard deviation, maximum, minimum and root mean square of acceleration, jerk and gyroscope signals as well as the sequence of acceleration magnitudes. An additional feature we included was the length of the gait cycle (the elapsed time between consecutive contacts of a single leg with the ground). A typical gait cycle is shown in Figure 1. From the vertical component of acceleration we extracted a_{max} and a_{min}, the maximum and minimum values of vertical acceleration in a single gait cycle, to compute $\sqrt[4]{a_{max} - a_{min}}$ which has been shown to be proportional to step-size (which is proportional to height) [10]. We hoped that this feature would be predictive of height.

4.1.2 Frequency Domain

We also considered several features in the frequency domain. We first sampled the total acceleration time signal in fixed-width sliding windows of 2.5 seconds each (250 timesteps at 100 Hz), with 50% overlap between successive windows. A window size of 2.5 seconds is suitable because the cadence of an average person is at least 1.5 steps per second [8]. Thus, we are able to capture close to a full walking cycle (2 steps) in each window. We then mapped the signal samples to the frequency domain using a Fast Fourier Transform (FFT) and extracted the desired features using the generated coefficients.

4.2 Learning algorithms

Our primary learning algorithms were AdaBoost with decision stumps and Linear Discriminant Analysis (LDA), both of which were selected because they are easy to generalize to the multi-class setting (for AdaBoost, we used the SAMME.R algorithm for multi-class classification proposed by Zhu et al. [11] and implemented in SKLearn). AdaBoost with decision stumps was particularly well suited to our analysis because it inherently performs feature selection. On the other hand, we need to perform feature selection before fitting our model with LDA to avoid overfitting.

4.3 Selection and Importance

Gender	Weight	Height
Std Dev X_g	Max Y_b	Dist btwn peaks Y_b
Mean Z_j	Std Dev X_b	Std Dev β_g
DC Component	Std Dev Y_b	Dist btwn peaks β_b

Figure 3: The three most important features for each trait

We employed feature selection algorithms to both improve the accuracy of our classifiers and gain insights into which aspects of gait are associated with a user's physical characteristics. The three most informative features when predicting weight, height and gender are shown in Figure 3. As previously noted, we elected to use AdaBoost for predicting weight and gender as it inherently performs feature selection. We considered the most informative features to be the ones that appeared most frequently in the decision stumps. On the other hand, we used LDA for predicting height. Since LDA does not perform feature selection on its own, we selected important features using stability selection as described by Meinshausen et al. and implemented in SKlearn [13].

By examining Figure 3, we notice that there does not appear to be a set of uniformly most informative features, but rather different features were important for each physical characteristic. This may indicate that different aspects of gait are more linked to different characteristics. Accordingly, we performed our classification tasks by using separate models for predicting each trait. We found it interesting and intuitive that our model for height benefited from the inclusion of the length of the gait cycle, as taller individuals tend to take longer strides.

5 Individual User Identification: A Fortuitous Discovery

The HASC data set contains multiple distinct examples of gait data per individual. Since we initially desired to classify individuals on the basis of weight, height and gender (and not on the basis of other confounding factors such as identity), we filtered the data to ensure that there is only one example of gait data from each individual in either the training or test set (but never both). Initially, however, we did not have these filters in place and our accuracies when predicting weight, height and gender were significantly higher. This gave us reason to believe that our previous accuracies were due to the ability of our classifier to identify individual users.

To explore this possibility, we performed Principal Component Analysis (PCA) to visualize the data. From the projection of the data onto the first three principal components, it is evident that individuals form fairly distinct clusters (Figure 4). Furthermore, since the first five principal components explain over 95% of the variance in the data, we decided to use PCA as a dimensionality reduction technique and k-nearest neighbors with $k = 3$ as our classifier. From PCA we found that the first principal component is weighted towards y-acceleration features, while the second is weighted towards x-acceleration features. These correspond to vertical

movement and abduction/adduction of the hips [4], respectively, so it was interesting to see the principal components weighted in this fashion.

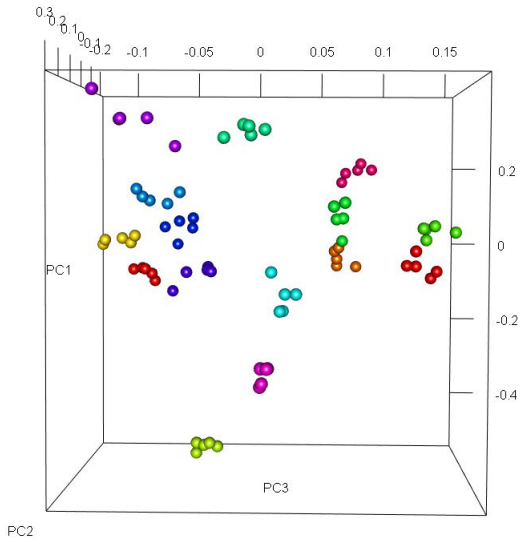

Figure 4: Projection of gait data from 15 individuals onto the first three principal components. Each color represents a different individual.

6 Results

6.1 Soft Biometrics

Our results are summarized by the following confusion matrices, which display our leave-one-out cross validation results on the combined data set (HASC 2012 + TURK). Each confusion matrix also shows precision and recall for each class, which are denoted P and R, respectively. Let S, M, L denote small, medium and large and S_a and S_p denote actual small and predicted small. Similarly define M_a, M_p, L_a, L_p.

6.1.1 Gender

	$Male_p$	$Female_p$	R
$Male_a$	**99**	8	*93%*
$Female_a$	17	**20**	*54%*
P	*85%*	*71%*	

Figure 5: Gender

We were able to identify users' gender with a high degree of accuracy. Indeed, we achieved both higher precision and recall for males than did Weiss and Lockhart [6] (precision: 85% v. 72%, recall: 93% v. 82%). Our performance on females was nearly identical. We obtained slightly higher precision (71% v. 70%), but slightly lower recall (54% v. 57%). Despite the fact that our data set has a disproportionate number of males, our results are comparable to those obtained by Weiss and Lockhart (who used a far more balanced data set). Though our overall accuracy is not significantly higher than baseline (83% v. 74%), our classifier achieved both high precision and recall and, thus, is far superior to the baseline predictor.

6.1.2 Weight

	S_p	L_p	R
S_a	**37**	9	*80%*
L_a	11	**28**	*72%*
P	*77%*	*76%*	

Figure 6: Weight (excluding middle class)

	S_p	M_p	L_p	R
S_a	**28**	16	4	*58%*
M_a	6	**38**	13	*67%*
L_a	8	15	**16**	*41%*
P	*67%*	*55%*	*48%*	

Figure 7: Weight

We were also fairly accurate when predicting weight. First, we will compare our results to those of Weiss and Lockhart when we exclude the middle class and only attempt to classify users of extreme weight. Our overall accuracy was 76%, which was comparable to their performance. We achieved higher recall on the small class, but lower precision, while the opposite was true on the large class. Thus, our results highly resemble those of Weiss and Lockhart. Our data set, however, is significantly larger and our data collection was not as controlled, so it is encouraging to be able to reproduce their results.

For the three class problem, we achieved an overall accuracy of 57% with satisfactory marginal performance for each class. We again outperformed the baseline predictor (which achieves 40% accuracy on this data set). From the confusion matrix, we see that there is little confusion among the small and large classes, which further suggests that weight can be accurately predicted from smartphone gait data.

6.1.3 Height

	S_p	L_p	R
S_a	**22**	6	*79%*
L_a	9	**16**	*64%*
P	*71%*	*73%*	

Figure 8: Height (excluding middle class)

	S_p	M_p	L_p	R
S_a	**17**	16	7	*43%*
M_a	22	**51**	6	*65%*
L_a	8	5	**12**	*48%*
P	*36%*	*71%*	*48%*	

Figure 9: Height

Our results for height are less positive than those for weight and gender. First, we will compare our results to those of

Weiss and Lockhart when we exclude the middle class and only attempt to classify users of extreme height. We achieved an overall accuracy of 72%, while they achieved accuracy of approximately 80%. They also achieved higher precision and recall for both the small and large classes.

For the three class problem, we achieved an overall accuracy of 56% which is just marginally better than the baseline predictor in this instance (55%). This is likely due to the skewed nature of our height data, which is described in Section 3.2. Improving our results on height prediction would likely require a more varied data set and more predictive features.

6.2 Identity

To access our accuracy when identifying individual users, we again performed leave-one-out cross validation. Specifically, we held out a single walking example from one individual (while including all other examples from that individual as well as the examples from everyone else in the training set) and then tried to identify which individual is associated with the held out example. We achieved an overall accuracy of 96% when identifying user gait data collected from the mounted position. Furthermore, 63 of the 75 individuals were correctly identified 100% of the time, 10 of the 75 individuals were identified 80% of the time and the remaining 2 were correctly identified 60% of the time. When identifying user gait data from either the fixed or mounted position, we correctly identified the user associated with the gait data 76% of the time. Our accuracy of 96% is on waist-mounted gait data is on par with the work achieved by prior researchers [4,12], which is impressive considering our data set contained significantly more individuals. However, we also note there was a recent Kaggle competition [5] that showed accelerometer data could be used effectively as a biometric.

7 Summary

As smartphone accelorometer and gyroscope sensors grow more and more ubiquitous, the potential uses for applications that can predict a user's identity and physical characteristics from a smartphone lying passively in one's pocket are seemingly endless. One potential application is user-targeted advertising, where knowledge of an individual's gender or specific identity would greatly increase the precision of the targeting. Considering the success we had in identifying individuals, we can also envision a smartphone application that can detect and report potential theft of the device by recognizing when the phone is being carried by an unauthorized user. Further, we believe our research demonstrates that there are many untapped applications for this data that can lead to a significantly enhanced user experience.

8 References

[1] Smith, Aaron. "Smartphone Ownership 2013." Pew Internet & American Life Project. http://pewinternet.org/Reports/2013/Smartphone-Ownership-2013/Findings.aspx (accessed December 11, 2013).

[2] Kwapisz, Jennifer, Gary Weiss, and Samuel Moore. "Activity recognition using cell phone accelerometers." ACM SigKDD Explorations Newsletter 12, no. 2 (2010): 74-82.

[3] Nham, Ben, Kanya Siangliulue, and Serena Yeung. "Predicting Mode of Transport from iPhone Accelerometer Data." CS229 Machine Learning.

[4] Nowlan, Michael. "Human Identification via Gait Recognition Using Accelerometer Gyro Forces." Yale Computer Science. http://www.cs.yale.edu/homes/mfn3/pub/mfn_gait_id.pdf (accessed November 12, 2013).

[5] "Accelerometer Biometric Competition." Kaggle. http://www.kaggle.com/c/accelerometer-biometric-competition (accessed December 14, 2013).

[6] Smith, Aaron. "Smartphone Ownership 2013." Pew Internet & American Life Project. http://pewinternet.org/Reports/2013/Smartphone-Ownership-2013/Findings.aspx (accessed December 11, 2013).

[7] "Data Hub for HASC." Human Activity Sensing Consortium. http://hub.hasc.jp/menu (accessed December 14, 2013).

[8] Anguita, Davide, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Reyes-Ortiz. "A Public Domain Dataset for Human Activity Recognition Using Smartphones." In 21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges, Belgium: ESANN 2013, 2013.

[9] Silva, Joana. "Smartphone Based Human Activity Prediction." (2012).

[10] Pratama, A.R.; Widyawan; Hidayat, R., "Smartphone-based Pedestrian Dead Reckoning as an indoor positioning system," System Engineering and Technology (ICSET), 2012 International Conference on , vol., no., pp .1,6, 11-12 Sept. 2012

[11] Zhu, Ji, Hui Zou, Saharon Rosset, Trevor Hastie. "Multi-class adaboost." Statistics and Its Interface Volume (2009).

[12] Gafurov, Davrondzhon. "A survey of biometric gait recognition: Approaches, security and challenges." Annual Norwegian Computer Science Conference. 2007.

[13] Meinshausen, Nicolai, and Peter B hlmann. "Stability selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72.4 (2010): 417-473.