WISDM Smartphone and Smartwatch Activity and Biometrics Dataset

Gary M. Weiss
Department of Computer and Information Science
Fordham University
441 East Fordham Road, Bronx NY 10458
gaweiss@fordham.edu

Abstract—Members of the WISDM (Wireless Sensor Data Mining) Lab in the Department of Computer and Information Science of Fordham Unversity collected data from the accelerometer and gyroscope sensors of a smartphone and smartwatch as 51 subjects performed 18 diverse activities of daily living. Each activity was performed for 3 minutes, so that each subject contributed 54 minutes of data. These activities include basic ambulation-related activities (e.g., walking, jogging, climbing stairs), hand-based activities of daily living (e.g., brushing teeth, folding clothes), and various eating activities (eating pasta, easting chips). The data set contains the low level time-series sensor data from the phone's accelerometer, phone's gyroscope, watches' accelerometer, and watches' gyroscope. All of the time-series data is tagged not only with the activity that was being performed, but with a subject identifier, which means that the data be used for building and evaluating biometrics models, as well activity recognition models. Researchers in the WISDM Lab subsequently used a sliding window approach to transform the time-series data into labeled examples, and the scripts for performing the transformation, as well as the transformed data, are also provided with the data set. The data set is available from the UCI Machine Learning Repository as the "WISDM Smartphone and Smartwatch Activty and Biometrics Dataset."

1 Overview

The "WISDM Smartphone and Smartwatch Activity and Biometrics Dataset" includes data collected from 51 subjects, each of whom were asked to perform 18 tasks for 3 minutes each. Each subject had a smartwatch placed on his/her dominant hand and a smartphone in their pocket. The data collection was controlled by a custom-made app that ran on the smartphone and smartwatch. The sensor data that was collected was from the accelerometer *and* gyrocope on both the smartphone *and* smartwatch, yielding four total sensors. The sensor data was collected at a rate of 20 Hz (i.e., every 50ms). The smartphone was either the Google Nexus 5/5X or Samsung Galaxy S5 running Android 6.0 (Marshmellow). The smartwatch was the LG G Watch running Android Wear 1.5. The general characteristics of the data and data collection process are summarized in Table 1. More detailed information is presented later in this document.

TABLE 1
SUMMARY INFORMATION FOR THE DATASETS

Number of subjects	51
Number of activities	18
Minutes collected per activity	3
Sensor polling rate	20Hz
Smartphone used	Google Nexus 5/5x or Samsung Galaxy S5
Smartwatch used	LG G Watch
Number raw measurements	15,630,426

2 The Activities

Table 2 lists the 18 activities that were performed. The actual data files specify the activities using the code from Table 2. Similar activities are not necessarily grouped together (e.g., eating activities are not all together). This mapping can also be found at the top level of the data directory in *activity_key.txt*.

TABLE 2
THE 18 ACTIVITIES REPRESENTED IN DATA SET

Activity	Code
Walking	A
Jogging	B
Stairs	C
Sitting	D
Standing	E
Typing	F
Brushing Teeth	G
Eating Soup	H
Eating Chips	I
Eating Pasta	J
Drinking from Cup	K
Eating Sandwich	L
Kicking (Soccer Ball)	M
Playing Catch w/Tennis Ball	O
Dribblinlg (Basketball)	P
Writing	Q
Clapping	R
Folding Clothes	S

In some of our research articles [3, 4] we partition these activities into 3 groupings to facillitate analysis, as follows:

Non-hand-oriented activities:
{walking, jogging, stairs, standing, kicking}

Hand-oriented activities (General):
{dribbling, playing catch, typing, writing, clapping, brushing teeth, folding clothes}

Hand-oriented activities (eating):
{eating pasta, eating soup, eating sandwich, eating chips, drinking}

3 The Raw Time-Series Sensor Data

The raw time-series sensor data is recorded by the accelerometrer and gyrsocope on both the phone and watch at a rate of 20Hz. These sampling rates are just suggestions to the operating systems and polling of the sensors can be delayed if the processor is busy. Each sensor measurement that we record is for one sensor on one device. Thus there are effectively four sensors, which we abbreviate as: phone-accel, phone-gyro, watch-accel, and watch-gyro. The data for each of the four sensors are recorded in different subdirectories, as shown visually in Figure 1.

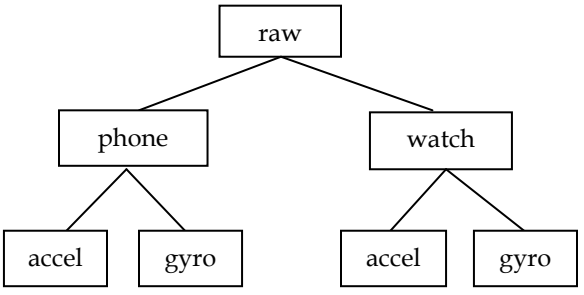

Figure 1. Under the main data directory is a subdirectory called raw that contains all of the raw sensor data. The phone and watch sensor data are stored in separate subdirectories, and the accelerometer and gyroscope data for each device are stored in the corresponding two subdirectories.

Within each subdirectory there is a file per subject, so there will be 51 files in each of the four subdirectories. The files are all named using a standard naming convention with four components. Sample filenames are:

- data_1600_accel_phone.txt
- data_1634_gyro_watch.txt

The first component is fixed and is always "data_". The second component identifies the subject and varies from 1600 to 1650 (i.e., there are 51 subjects and labeling starts at 1600). The third component is either "accel" or "gyro" to identify the sensor, and the fourth component is "phone" or "watch" to identify the device on which the sensor resides. All of these files end with ".txt" since they are text files. Note that technically the third and fourth components of the filenames are not needed since the device and sensor type is implied by the directory structure. However it is convenient to have all of the necessary information encoded in the filename.

Within each time-series sensor file is one sensor reading per line. The format of this line is identical across both devices and both types of sensors. The data on each line is comma separated and each line ends with a semicolon. The format of each line is as follows: *Subject-id, Activity Code, Timestamp, x, y, z;*

Table 3 defines each of the six features. The actual sensor values have different units for the different sensors. For the accelerometer sensor, the units are m/s^2, while for the gyroscope sensor, the units are radians/s. Note that the force of gravity on Earth, which affects the accelerometer readings, is $9.8m/s^2$.

TABLE 3
Definition of Elements in Raw Data Measurements

Field name	Description
Subject-id	*Type: Symbolic numeric identifier.* Uniquely identifyies the subject. Range: 1600-1650.
Activity code	*Type: Symbolic single letter.* Identifies a specific activity as listed in Table 2. Range: A-S (no "N" value)
Timestamp	*Type: Integer.* Linux time
x	*Type Numeric: real.* Sensor value for x axis. May be positive or negative.
y	Same as x but for y axis
z	Same as x but for z axis

The filename "data_1600_accel_phone" corresponds to the time-series sensor data for subject 1600 from his/her phone accelerometer. The file contains 64,311 lines/sensor readings. Given a sampling rate of 20Hz this corresponds to 53.59 minutes of data, which is very close to the expected 54 (18 × 3) minutes of data. If precisely 54 minutes of readings were recorded that would correspond to 64,800 lines). If we look at the first 3 lines in the "data_1600_accel_phone" file we see the following, which corresponds to walking (code A in field 2) activity data:

1600,A,252207666810782,-0.36476135,8.793503,1.0550842;

1600,A,252207717164786,-0.8797302,9.768784,1.0169983;

1600,A,252207767518790,2.0014954,11.10907,2.619156;

The raw data directory contains a total of 15,630,426 lines of measurements across the four subdirectories. Ideally, since there should be 64,800 lines for each of the 51 subjects, there should be 3,304,800 lines of data per subdirectory. However, the data collection pro-

cess is not perfect, so we expect these numbers will vary. The actual number of lines of data per subdirectory is as follows:

- raw/phone/accel: 4,804,403
- raw/phone/gyro: 3,608,635
- raw/watch/accel: 3,777,046
- raw/watch/gyro: 3,440,342

We are unsure exactly why there are so many more sensor readings for the phone accelerometer. Next, Table 4 shows the distribution of sensor readings by activity. This was computed by going into the subdirectory for each of the four sensors and executing the following shell command:

cat data_* | cut -d"," -f2 | sort | uniq -c

This command extracts the second field in the comma delimited files (the activity code), then sorts it, and then counts the number of entries for each unique value/code. The "Total" field in Table 4 is the sum of the values preceding it in each row, and the "Class %" is computed as this total value divided by 15,630,426, the total number of sensor readings. Thus "Class %" represents the class distribution for the activity class. Table 4 shows that the class distribution over these values is close to the ideal amount, which would be 5.55% (100/18).

TABLE 4
DISTRIBUTION OF RAW SENSOR DATA

Activity	Phone		Watch		Total	Class %
	Accel	Gyro	Accel	Gyro		
Walking	279,817	203,919	210,495	192,531	886,762	5.7%
Jogging	268,409	200,252	205,787	187,833	862,281	5.5%
Stairs	255,645	197,857	207,312	180,416	841,230	5.4%
Sitting	264,592	202,370	213,018	195,050	875,030	5.6%
Standing	269,604	202,351	216,529	194,103	882,587	5.6%
Typing	246,356	194,540	205,137	187,175	833,208	5.3%
Brush Teeth	269,609	202,622	208,720	190,759	871,710	5.6%
Eat Soup	270,756	202,408	209,483	187,057	869,704	5.6%
Eat Chips	261,360	197,905	210,048	192,085	861,398	5.5%
Eat Pasta	249,793	197,844	203,112	189,609	840,358	5.4%
Drinking	285,190	202,395	215,879	197,917	901,381	5.8%
Eat Sandwich	265,781	197,915	203,684	190,191	857,571	5.5%
Kicking	278,766	202,625	209,491	191,535	882,417	5.6%
Catch	272,219	198,756	210,107	187,684	868,766	5.6%
Dribblinlg	272,730	202,331	212,810	194,845	882,716	5.6%
Writing	260,497	197,894	215,365	197,403	871,159	5.6%
Clapping	268,065	202,330	208,734	190,776	869,905	5.6%
Fold Clothes	265,214	202,321	211,335	193,373	872,243	5.6%
Total	4,804,403	3,608,635	3,777,046	3,440,342	15,630,426	100%

4 THE TRANSFORMED ACTIVITY EXAMPLES

Activity recognition and biometric models can be built directly from the time-series data, but most machine learning and data mining methods require the data be in the form of labeled examples and not labeled time series values. This data set includees labeled examples in addition to the raw time-series data. There are many possible transformation processes that can be applied, and the process reflected in the examples associated with the data set reflects just one process. Therefore we view the raw time series data as the main contribution of this data set, since researchers can use it to generate any higher level representation that they want.

The transformation process that was used to generate the labeled examples is described in Section 4.1, whichh also describes the features that are generated and the layout/format of each generated example. Section 4.2 then supplies some meta-data about the data, such as the number of examples generated and the distribution over the various activity values.

4.1 The Data Transformation Process

The basic data transformation process utilized to form the labeled examples provided with this data set has been used by our WISDM Lab since 2010 and has been used in many research papers-- although normally on smaller data sets. The process was described in our first article on activity recognition [2], although the transformation process applied to generate the examples in this data set include some additional features.

The raw time series data for each sensor (per subject and per activity) is divided into 10-second non-overlapping segments and then high-level features are generated based on the 200 (10s × 20 readings/s) readings contained within each segment. A 10-second window was chosen because we felt that it provided sufficient time to capture several repetitions of those actitions that involve repititious movements, and still is small enough to provide quick response (i.e., a prediction every 10 seconds).

The transformation process described in our prior research yields 43 features excluding the activity label, or 44 features with the activity label. However, the examples included in this data set have 93 features (with the label), because some experimentation was done and additional features were added, but never used in published research. For completeness we described the 93 features, but identify the 49 features that were not used in any published papers.

Our transformed examples are placed into ARFF (Attribute-Relation File Format) files, which is a file format specified by the WEKA suite of data mining tools [1]. ARFF files contain both formatting information for the data and then the data itself. The formatting information specifies information about all of the attributes, so we will use the ARFF header to introduce and describe the features generated by the transformation process. The start of every ARFF file will have the same header information, with one exception to be described shortly (line 95 in Table 5).

The very first line will always be the following line which defines the relation, and this line will always be followed by a blank line:

@relation person_activities_labeled

The remainder of the header, starting with line 3, is described by Table 5. The header goes from line 3 to line 95, with one attribute per line, and thus we get 93 attributes/features. Every line in the header from line 3 to line 95 will begin with "@attribute". The first column in Table 5 specifies the corresponding line number in the ARFF file. The second column specifies the name of the attribute, which follows the "@attribute" text that starts each line. The attribute name is always included in double quotes. The third column specifies the attribute type, or, in the case of a categorical feature, a list of all possible feature values. This information follows the attribute name on the line. Some of the attributes/features have many variants (e.g., several have a value per axis so come in multiples of three), so in some cases the table entries are compressed into a single row (but with multiple line numbers in the first column). For example, feature X0 appears on line 4, X1 on line 4, …, and X9 on line 13. Similarly, XAVG is on line 34, YAVG on line 35, and ZAVG on line 36.

The 49 features in Table 5 that are denoted with an asterisk (*) and highlighted in bold are not used in any of our published research. Descriptions of all of the features are provided following Table 5. Recall that each of the featuresare based on the sensor readings of one sensor over a 10 second window.

TABLE 5
LAYOUT OF ARFF HEADER FILE

Line #	Attribute Name	Attribute Type or Values
3	ACTIVITY	{A,B,C,D,E,F,G,H,I,J,K,L,M,O,P,Q,R,S}
4-13	X{0-9}	numeric
14-23	Y{0-9}	numeric
24-33	Z{0-9}	numeric
34-36	{X,Y,Z}AVG	numeric
37-39	{X,Y,Z}PEAK	numeric
40-42	{X,Y,Z}ABSOLDEV	numeric
43-45	{X,Y,Z}STANDDEV	numeric
46-48	**{X,Y,Z}VAR***	numeric
49-61	**XMFCC{0-12}***	numeric
62-77	**YMFCC{0-12}***	numeric
75-87	**ZMFCC{0-12}***	numeric
88-90	**{XY, XZ, YZ}COS***	numeric
91-93	**{XY, XZ, YZ}COR***	numeric
94	RESULTANT	numeric
95	**class***	{16XX}

* The value for class is the subject identifier for the file and is a single value between 1600 and 1650.

The descriptions of the features are below:

- ACTIVITY: specifies the activity performed using one of the activity codes from Table 2.
- X{0-9}; Y{0-9}, Z{0-9}: These 30 features collectively show the distribution of values over the x, y, and z axes. We call this a binned distribution. For each axis we determine the range of values in the 10s window (max – min value), divide this range into 10 equal-sized bins, and then record the fraction of values in each bin.
- {X,Y,Z}AVG: Average sensor value over the window (per axis).
- {X,Y,Z}PEAK: Time in milliseconds between the peaks in the wave associated with most activities. Heuristically determined (per axis).
- {X,Y,Z}ABSOLDEV: Average absolute difference between the each of the 200 readings and the mean of those values (per axis)
- {X,Y,Z}STANDDEV: Standard deviation of the 200 values (per axis)
- {X,Y,Z}VAR: Variance of the values (per axis)
- XMFCC{0-12}, YMFCC{0-12}, ZMFCC{0-12}: MFCCs represent short-term power spectrum of a wave, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. There are 13 values per axis.
- {XY, XZ, YZ}COS: The cosine distrances between sensor values for pairs of axes (three pairs of axes).
- {XY, XZ, YZ}COR: The correlation between sensor values for pairs of axes (three pairs of axes).
- RESULTANT: Average resultant value, computed by squaring each matching x, y, and z value, summing them, taking the square root, and then averaging these values over the 200 readings.
- Class: This is a very different feature than the rest. It simply is set to the subject-id.

After the header is complete, there will be a blank line and then on line 97 will be "@data", which signals the start of the actual data (which begins on line 98). Then there wil be one example per line, comma-separated, with no delimiter at the end of the line.

The transformation process was accomplished using scripts that are available under the arffmagic-master directory at the top level of the data directory. The code can be used to provide additional details about how the examples were generated. However, the scripts have not continued to be maintained and may require some effort before they can be executed.

4.2 Detailed Information about the Examples

The transformed examples are stored in a manner analogous to the raw time-series data files, with the only difference being that the file names end with ".arff" rather than ".txt". That is, the data files are stored in a directory structure identical to that shown earlier in Figure 1, except that the data is stored under a directory called "arff_files" rather than "raw". Thus the transformed examples for the phone accelerometer sensor are found under "arff_files/phone/accel" and that subdirectory will have 51 files corresponding to the 51 subjects. The phone accelerometer examples for subject 1600 will be found in the following file in this subdirectory: *data_1600_accel_phone.arff*. The arff extension indicates that the file is a valid "arff" file, as specified by the Weka Data Mining Toolkit [1].

The actual number of examples in each subdirectory is provided below:

- arff_files/phone/accel: 23,173
- arff_files /phone/gyro: 17,380
- arff_files /watch/accel: 18,310
- arff_files /watch/gyro: 16,632

We expect to have 45.9 hours of examples per sensor (since there are 51 subjects performing 18 activities for 3 minutes per activity). To determine the number of 10-second intervals in 45.9, we multiply 45.9 by 60 (minutes/hour) and then by 6, since there are 6 10-second intervals in a minute. This yields 16,524 examples. We see that is about what we have for three of the four sensors, although we have more for the phone accelerometer, which is consistent with what we say with the raw time series data.

TABLE 6
DISTRIBUTION OF EXAMPLES

Activity	Phone		Watch		Total	Class %
	Accel	Gyro	Accel	Gyro		
Walking	1,271	936	1,011	915	4,133	5.5%
Jogging	1,314	966	993	902	4,175	5.6%
Stairs	1,180	946	997	865	3,988	5.3%
Sitting	1,263	984	1,028	939	4,214	5.6%
Standing	1,283	969	1,046	934	4,232	5.6%
Typing	1,180	938	988	900	4,006	5.3%
Brush Teeth	1,282	954	1,006	918	4,160	5.5%
Eat Soup	1,252	974	1,012	899	4,137	5.5%
Eat Chips	1,236	947	1,011	922	4,116	5.5%
Eat Pasta	1,179	959	978	911	4,027	5.4%
Drinking	1,310	976	1,044	954	4,284	5.7%
Eat Sandwich	1,242	949	980	915	4,086	5.4%
Kicking	1,466	971	1,009	919	4,365	5.8%
Catch	1,431	944	1,015	903	4,293	5.7%
Dribblinlg	1,413	972	1,027	939	4,351	5.8%
Writing	1,241	948	1,038	948	4,175	5.6%
Clapping	1,270	978	1,009	917	4,174	5.6%
Fold Clothes	1,261	970	1,019	933	4,183	5.6%
Total	23,074	17,281	18,211	16,533	75,099	100.0%

The distribution of examples across the different activities is provided in Table 6, just as it was done in Table 4 for the raw sensor readings. This was computed by going into the subdirectory for each of the four sensors and executing the following shell command (it is similar to the command used earlier for the time-series data but here we start processing the file on line 98 which is the first line of actual data after the ARFF header):

```
tail -n +98 data_* | cut -d"," -f1 | sort | uniq -c
```

The results in Table 6 demonstrate that the class distribution for each activity is close to the expected value of 5.55% (100/18). The variation in the number of readings per sensor is expected given the distribution of raw time-series data entries in Table 4 (which showed more readings for the phone acclerometer).

5 PAST USAGE OF DATA

The data set described in this document has currently been used in two published studies, both conducted by the author's WISDM Lab. One of these studies is a biometrics study [3] and the other is an activity recognition study [4]. Extended versions of those two conference papers have been submitted to journals and are currently under review. We hope to update this section once those articles are published, but in the meantime interested parties may check with the author.

ACKNOWLEDGMENT

The author wishes to thank all members of the WISDM Lab who assisted with data collection throughout the years, but especially Abby O'Neill and Kenichi Yoneda, who were instrumental in collecting this data set. While this data set was not collected using funding from the US National Science Foundation, our earlier research and associated data collection efforts were funded by NSF grant 1116124, which did help us develop the tools and experience necessary to efficiently collect this data.

REFERENCES

[1] E. Frank, M.A. Hall, and I.H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition.

[2] J.R. Kwapisz, G.M. Weiss and S.A. Moore. Activity Recognition using Cell Phone Accelerometers, *ACM SIGKDD Explorations*, 12(2):74-82.

[3] K. Yoneda and G.M. Weiss (2017). Mobile Sensor-based Biometrics using Common Daily Activities, *Proceedings of the 8th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, New York, NY, 584-590.

[4] G.M. Weiss, J.L. Timko, C.M. Gallagher, K. Yoneda, and A.J. Schreiber (2016). Smartwatch-based Activity Recognition: A Machine Learning Approach, *Proceedings of the 2016 IEEE International Conference on Biomedical and Health Informatics (BHI 2016)*, Las Vegas, NV, 426-429.