Human Activity Prediction Across Multiple Datasets

Colin M. Adams[*]
Detection & Estimation
Areté Associates
Northridge, CA 91324
cadams@arete.com

Kai S. Kaneshina
Detection & Estimation
Areté Associates
Northridge, CA 91324
kkaneshina@arete.com

Eli J. Weissler
Detection & Estimation
Areté Associates
Northridge, CA 91324
eweissler@arete.com

Abstract

Human Activity Recognition has recently become a field of focus for machine learning researchers due to the wide-scale proliferation of smartphones and the sensors that they possess. While many researchers have shown success predicting activity based on accelerometer data, their primary focus has on achieving the highest possible accuracy on a single dataset, gathered under controlled conditions. Our goal is to expand on previous work by creating a classifier capable of identifying activities across multiple data sets, with an eye towards an algorithm robust enough to handle real-world data. This approach requires us to take into account the phone's orientation, which has a large impact on accelerometer readings. The phone's orientation is kept consistent within a single academic dataset, so it is usually not taken into account. We attempted to create an orientation invariant representation of accelerometer data by rotating all acceleration axes to a common frame and then aligning them with the principal components of the acceleration over a single feature vector. While we believe this approach is promising, our initial attempt performed worse than simply using the unprocessed data, primarily due to misidentifying sitting and standing. Future work will look to utilize time invariant and coordinate invariant features, in addition to separate classifiers for active and sedentary activities in order to improve our cross dataset activity recognition accuracy.

1 Introduction

As of 2019, it was estimated that more than five billion people have cell phones and over half of these devices are smartphones. The amount of data generated from these devices is enormous; most of them are equipped with a myriad of sensors, e.g. accelerometer, gyroscopes, magnetometers, and so forth [1]. The information from these sensors has enabled the study of Human Activity Recognition (HAR), i.e. the determination of what physical activity a person is performing (for example, HAR determines if a person is walking, running, sitting, standing, falling, etc.) based upon the outputs of these sensors. HAR has a variety of applications. For instance, in healthcare, determining when the elderly have had a potentially-devastating fall; another, in sports, where techniques can be refined to perfection; in cybersecurity, for continuous-user authentication by keeping track of who is physically manipulating the smartphone; and, finally, in generating a more secure biometric key which are harder to spoof than the traditional fingerprint and iris scanners [2, 3].

The sheer number of smart phones opens the door to big data approaches. Often, HAR data is non-linear so typical machine-learning techniques—such as support vector machines (SVM), principal component analysis (PCA), etc.—which rely on linearity for analysis, do not perform as well as deep neural networks which are not constrained linearly [4, 3, 5, 6]. Neural network models, with

[*] The names are ordered alphabetically.

33rd Conference on Neural Information Processing Systems (NIPS 2020), Vancouver, Canada.

complicated, unintuitive architectures, merely provides an output from a set of inputs and allow for very little in the way of understanding the fundamental aspects of the problem. Deep neural networks also rely on vast amounts of high-fidelity, labeled training data, which often is not feasible to gather for most problems. As a result, neural networks are often over-fit, and their accuracy can be highly dependent on the specific conditions in which the data was taken (e.g. data taken in controlled conditions).

Our hope is to improve the performance of traditional machine-learning methods in the context of HAR by accurately accounting for the inherent non-linearity of the data through extensive preprocessing. This would both allow for higher performance and elucidate what our algorithm is actually doing to classify activities. The hope is that we will perform better than the both the supervised and unsupervised machine-learning methods used in the other papers with a stretch goal of approaching neural network performance.

With this in mind, we have a simple—albeit rather difficult—goal. By combining multiple datasets, we aim to develop a robust algorithm to classify six activities—walking up stairs, walking down stairs, walking, jogging, sitting, and standing—using only the accelerometer (and maybe gyroscopic) data recorded by various smartphones. To measure the accuracy of our classifiers, we will cross-validate across users, so that we *never* predict a user's activity using a classifier already trained on that user. By combining multiple datasets for training and testing our model, with different sensor types, sampling rates etc., we are hoping to improve the reliability of our classifier in novel situations. If we are able to perform well, despite the fundamental differences in the datasets, then our model may be robust enough for real-life use.

2 Our Data

2.1 The Raw Data

We have chosen to combine multiple datasets into a single large dataset in the hope that it will lead to more reliable and robust results for accurately classifying activities of daily life. Originally, we decided to use three datasets that are relatively well known within the HAR community. These datasets have been used previously throughout HAR technical literature and are described in detail below.

1. MOTION SENSE: Twenty-four volunteers—of various ages, genders, heights, and weights—recorded six activities of daily life (walking, walking up- and downstairs, jogging, sitting, standing) using an iPhone 6S. The iPhone 6S recorded tri-axial acceleration, orientation, and gyroscopic data. It was kept in the user's front pocket and collected data at a rate of 50 Hz [2]. Data was taken on Queen Mary University of London's campus (51.524 °N, 0.040 °W).
2. MOBIACT: Sixty-six volunteers—of various ages, genders, height, and weight—performed four different types of falls and twelve different types of activities of daily life (six of which were walking, walking up- and downstairs, jogging, sitting, standing) resulting in more than 3200 trials. The data was captured using a Samsung Galaxy S3 and includes tri-axial accelerometer, orientation, and gyroscopic data (among others). Data was collected at a rate of 87 Hz. The phone was also in the user's front pocket during data collection [7]. Data was taken at Technological Educational Institute of Crete's campus (35.319 °N, 25.102 °W). The acceleration due to gravity was removed from the acceleration of the phone, but it is stored within the data separately.
3. UNIVERSITY OF CALIFORNIA AT IRVINE'S HUMAN ACTIVITY RECOGNITION (UCI–HAR): Data was collected by thirty volunteers—of various ages, genders, heights, and weights—with a Samsung Galaxy S2 performing six activities of daily life (walking, walking up- and downstairs, laying down, sitting, standing). They captured the tri-axial acceleration and tri-axial angular velocity from the phone's accelerometer and gyroscope. The data was taken at a sample rate of 50 Hz. A notable difference, however, is that the phone was attached on the user's waist with a strap [8]. Data was taken by SmartLab at Università degli Studi di Genova (44.415 °N, 8.927 °W).

After investigating the circumstances in which the UCI–HAR data was taken, we decided to not include it in our data combination. Specifically, the UCI–HAR data used a high-pass filter to remove the gravity components of the accelerometer data, which is impossible to reverse. In order to make our HAR algorithm as effective and robust as possible, we want to do all of our data processing on the raw accelerometer data—gravity included. This effectively allows us to take any accelerometer (and orientation) data from a smartphone and classify the user's activity.

2.2 Initial Data Processing

The raw data is initially broken up into segments with a 50% overlap with adjacent segments. While it would be ideal to have completely uncorrelated examples, we believed a 50% overlay was a good trade off between more samples and higher quality. We chose to test 128 and 256 length segments—corresponding to 2.56 s and 5.12 s each. Each one of these segments becomes a single feature vector. For a given trial measuring activity of daily life, we take the first 2.56 s as our first feature vector; our next feature vector begins in the middle of the first feature vector (e.g. our second feature vector begins at 1.28 s) and ends 2.56 s later (e.g. our second feature vector ends at 3.84 s). In other words we have a 50% overlap for each segment and the next. If there is not enough data for the last feature vector, then we do not make one. The datasets were originally kept separate for initial testing.

While the UCI-HAR and MOTION SENSE datasets were both sampled at 50 Hz, the MOBIACT data needs to be downsampled from 87 Hz to 50 Hz. Why would anyone sample at 87 Hz? It is unclear—perhaps that was the maximum sample-rate for the gyroscope in Samsung's Galaxy S3. Regardless, we need a consistent sample rate between datasets. Furthermore, all three of the datasets report different measurements for acceleration.

Lastly, we need to make sure that the tri-axial accelerometer data is measuring the same thing across datasets. It appears that some of the data has taken gravity into account whereas others have not. MOTION SENSE reports the acceleration from gravity and from the user in a single vector, whereas MOBIACT has them separately. Unfortunately, UCI–HAR removed gravity by using a high-pass filter with a 0.3 Hz threshold, removing any components below that—as such, it was not possible to use this dataset in our cross dataset comparison. However, we were able to use it to initially validate the results of our classifiers. We also adjusted the acceleration to have the same units across all datasets. Lastly, it was important to evaluate our cross-data comparison based on activities that were in both dataset. As a result, we only evaluated ourselves on the activities that were in both the MOBIACT and MOTION SENSE datasets, specifically walking, jogging, sitting, standing, walking upstairs, and walking downstairs.

After briefly looking into the magnitude the Coriololis effect would have our our acceleration data—dependent on the latitude where the data was taken by the phone—we concluded it was negligible for the 2.56 s to 5.02 s timescales we were interested in [9].

We also encountered one other issue with the MOBIACT dataset. When examining the data, it was clear that for about the first 5 sof each experiment, the data was anomalous, maybe because they were setting the phone into place. In order to prevent these outliers from affecting our training, we ignored the first 10 sof each experiment.

3 Literature

Due to the large amount of data that has been generated by smart phones, there have been many papers published on human activity recognition. One group of researchers compared the performance of three different classifiers (Random Forest, Support Vector Machine, Naïve Bayes) and three different deep neural network architectures, namely Multilayer Perceptron, Deep Convolutional Neural Network, and Long-Short Term Memory on labeled accelerometer, gyroscope, magnetometer and electrocardiogram data [10]. The researchers found that the deep neural networks performed much better than other machine-learning methods. Another group compared the performance of traditional machine learning methods (k-NN, SVM) to that of a residual neural network (ResNet) [1]. These researchers used a variety of datasets, including Motion Sense, UCI–HAR, and MobiAct. They used only the accelerometer data, and compared the classifier performance on the raw data vs hand-crafted features. For the Motion Sense dataset, the hand-crafted features improved the performance of the machine learned classifiers, but the ResNet applied to the raw data still achieved the highest accuracy. ResNets allow for skipping between network layers; these skips contain non-linearities, which the researchers claim allow them to better represent the non-linear relationships present in accelerometer data [5].

One theme we noticed in the literature was that researchers appeared to be in an accuracy arms race. Rather than focus on creating an algorithm that can function in a variety of real-world settings, they, by and large, aimed for the most accurate classifier within a single academic dataset. Perhaps no paper embodied this more than [7], which claimed an accuracy of 99.88% on the MobiAct dataset using an intense period of trial and error with different window sizes, overlaps, and combinations of hand-crafted features for their neural network.

This brings us back to our main goal: we do not wish to join the accuracy arms race. Instead we are aiming to develop a robust classifier that is not overfit to one specific academic dataset. Rather, we wish to successfully predict human activity across multiple academic datasets in order to create a method capable of functioning in real-world settings.

4 Initial Results

Our preliminary testing focused on the raw, unprocessed accelerometer data from all three datasets. Our first goal was to replicate some of the standard machine-learning classifier results we found in previous literature [1].

We initially wanted to confirm the results of training and testing on a raw data set [1]. We processed the data to be in the same format at the paper. The data was broken into 128-dimensional segments; there are 128 samples of each sensor (for each axes) within a single feature vector. In addition, each segment has a 50% overlap with the previous feature vector. We ensured that the feature vectors were only made from data that was recorded consecutively, thus the overlap corresponded to their recording time. These vectors were created for each activity for each user in each of the datasets that we used, namely the Motion Sense and MobiAct datasets.

Our initial results were within one standard deviation of the results for the k-Nearest Neighbors (k-NN) and Support Vector Machine (SVC) classifiers [1]. Our results are summarized in Table 1 on the Motion Sense dataset and Table 2 on the MobiAct dataset. Our initial results were consistent with the results of previous studies in the instances of SVC and k-NN [1, 5, 6]. There was no overlap between the feature vectors in the train and test data sets. Notably, there has not been much interest in using the extra trees methods for classifying this type of data, despite our initial findings that it produces the best results and runs the fastest.

4.1 User Cross Validation

For each dataset, we performed a cross validation across individual users in the dataset. In other words, for each user, we predicted their activity using a classifier trained on the data derived from all other uses. We believe that this is a better estimate of the robustness of the classification than a mixed or random training/test set because generally we cannot count on having prior examples of the behavior of persons whose behavior we wish to classify. The feature vectors consisted of only raw accelerometer data, as done in certain parts of [1]. We compared the results for the following classifiers: k-NN (3 nearest neighbors), SVC (using a one vs. rest scheme for more than two classes), and extra trees (100 estimators, minimum of 2 samples per split, and a minimum of 1 sample per leaf node).

4.1.1 Motion Sense

For the Motion Sense dataset, the extra trees classifier performed the best and had the lowest standard deviation of the three classifiers we tested. The SVC and k-NN performed about the same. The extra trees classifier was the best classifier for 70% of the users. There was no readily apparent correlation between user features (height, weight, age, gender) and classifier performance. We looked at all six activities of daily life.

4.1.2 MobiAct

For the MobiAct dataset, the extra trees and k-NN performed similarity. We only looked at the following activities: jogging, sitting, standing, walking upstairs, and walking downstairs. For brevity,

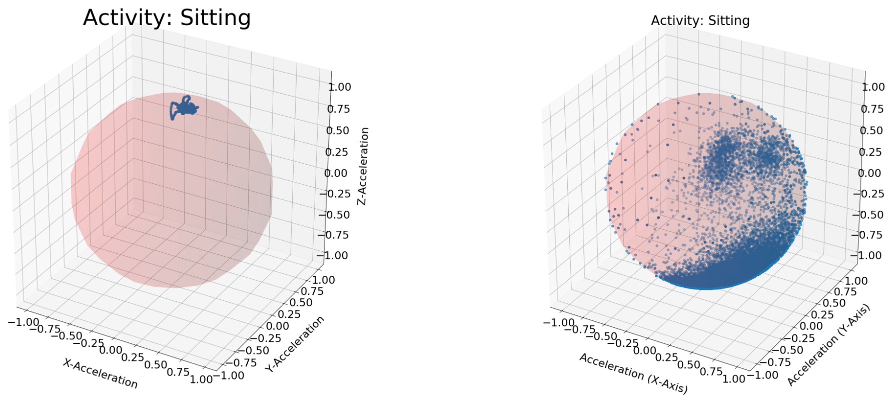
Table 1: Motion Sense

User	SVC	k-NN	Trees	User	SVC	k-NN	Trees
1	64.8%	58.6%	81.5%	13	77.8%	75.6%	82.9%
2	70.3%	69.8%	85.6%	14	77.0%	82.7%	82.6%
3	88.5%	77.4%	91.0%	15	82.0%	72.3%	80.6%
4	71.2%	39.8%	77.3%	16	69.7%	79.7%	76.6%
5	66.5%	75.3%	85.3%	17	69.2%	89.0%	80.0%
6	82.3%	82.0%	85.9%	18	82.8%	84.7%	82.2%
7	69.6%	69.0%	79.1%	19	76.4%	72.5%	81.9%
8	68.4%	67.7%	81.8%	20	73.1%	62.4%	80.2%
9	86.8%	69.8%	93.1%	21	82.8%	85.6%	85.9%
10	73.6%	62.2%	78.5%	22	62.6%	60.2%	87.7%
11	83.2%	88.4%	85.1%	23	64.5%	87.4%	78.5%
12	79.%	74.7%	89.1%	24	39.8%	84.3%	85.9%
Mean	73.4%	73.8%	83.3%	**SD**	10.3%	11.7%	4.3%

the only the mean and standard deviation results are shown in Table 2. While only doing the bare minimum data processing as described in Sec. 2.2, we ran simple classifiers on the raw accelerometer data. Our k-nearest neighbors and extra trees classifiers performed comparably to the Motion Sense data and our results were within a standard deviation of the results found in other literature [1].

Table 2: MobiAct

	SVC	k-NN	Trees		SVC	k-NN	Trees
Mean	N/A	92.19%	94.09%	**SD**	N/A	12.20%	13.89%



(a) Mobi Act Dataset

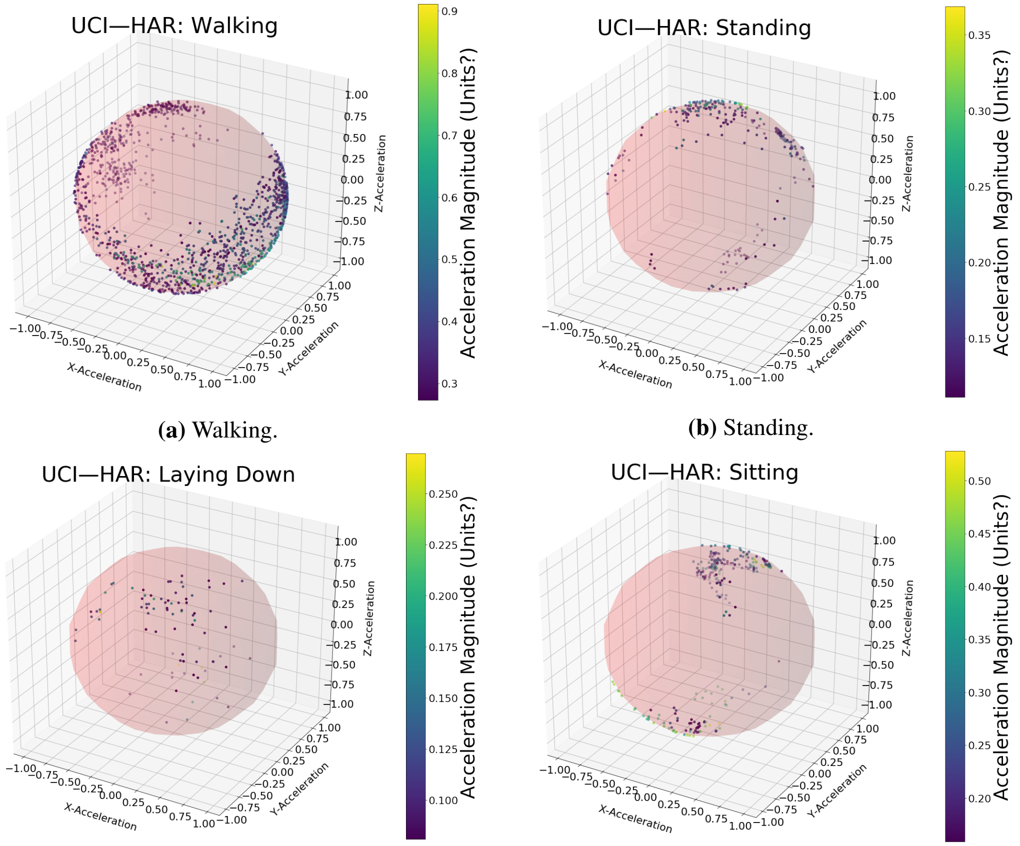
(b) Motion Sense Dataset

Figure 1: The images above were made for a random user in each dataset. As can be seen, there is significant difference in the clustering of the acceleration data between the two datasets used. The direction of the given acceleration vector $\hat{\mathbf{a}}$ is mapped to (x, y, z) and is plotted on the spheres.

4.2 Visualization

To move towards new representations of our data, we investigated different visualizations.

For the first visualization in Figure 1a, we plotted the direction of accelerometer (x, y, z) observations onto the surface of a sphere. A comparison was made for the sitting patterns of a random user from both the MobiAct and Motion Sense datasets. Comparing the two (Figs. 1a and 1b), the Motion Sense data is consistent with what we might expect from the waist/torso of someone who is sitting in a chair—the accelerometer data is predominately located at the poles (z-acceleration) as someone might shift her weight back and forth. In contrast, the MobiAct data features a much tighter distribution that is suggestive of the very small side to side back and forth that a phone may experience in the front pocket of a pair of pants. In this case, a rotationally invariant representation of the data would not resolve the inconsistency. These would be two different modes of sitting that would require distinct training examples to predict. Surprisingly, both datasets recorded data from the users' front pocket. Despite the same location on the user, the images look nothing alike.



(a) Walking.

(b) Standing.

(c) Laying.

(d) Sitting.

Figure 2: The direction and magnitude acceleration measured by the phone's accelerometer plotted on the surface of a unit sphere with the much of the low magnitude noise filtered out. The point on the sphere describes the x-,y-,z-direction of the vector itself. By filtering, we can remove much of the noise. For example, laying down has points scattered all around the surface of the sphere, essentially random whereas the sitting has accelerations on two opposite poles, suggesting a small adjustment (e.g. the shifting of a leg). More active movements like walking trace out arcing paths along the surface of the sphere.

Second, we performed a similar visualization, separating based on the labeled activity within the UCI dataset for the same user as in Figure 1. Again, the direction of the given acceleration vector (x, y, z) was plotted on a unit sphere, plus each point was colored to represent the magnitude of the acceleration. Only points with a magnitude of at least 30% of the maximum magnitude value was shown. This was a completely arbitrary value, however it allows us to get rid of much of the noise associated with the data and to more clearly see the trends of each activity. The sitting data is again consistent with the swaying back and forth of a torso while sitting in a chair. The standing data appears to show that the strongest acceleration occurs swaying side to side. The walking data appears clustered around a ring in the center of the sphere, possibly reflective of this individual's stride. Lastly, the laying down is difficult to interpret, although appears to show few data points with large accelerations. However, if you note the scale bar, it is significantly lower than any of the other activities of daily life.

The next visualization that was performed involves the time-integration of the accelerometer data which gives a proxy for the phone's velocity as a function of time. We can integrate again to get a proxy for the phone's position as a function of time. Figure 3 provides a nice example of this below. Compare the parameterized velocity curves for jogging and walking. The walking velocity quite clearly shows a downward drift in the velocity. We suspect this is an artifact of either the sensor or our current data analysis, as the velocity of a phone in a front pocket should not have systematic decrease over the course of a few seconds. Another important feature of the velocity parameterization of jogging(and less obviously in the walking data) is the sharp, periodic, looping structure.

An interesting observation to note in the jogging data is that over the course of the 2.56 s, the velocity curve has nearly three full repetitions. This is a well-known characteristic of running cadence where a runner is expected to take 150–180 steps per minute [11]. Our intuition tells us that we should expect a somewhat coiled velocity that is quasi-periodic, which can be achieved if we can eliminate or mitigate the drift. Regardless, these nearly periodic coils provide a strong signal for a phone's movement over short time frames.

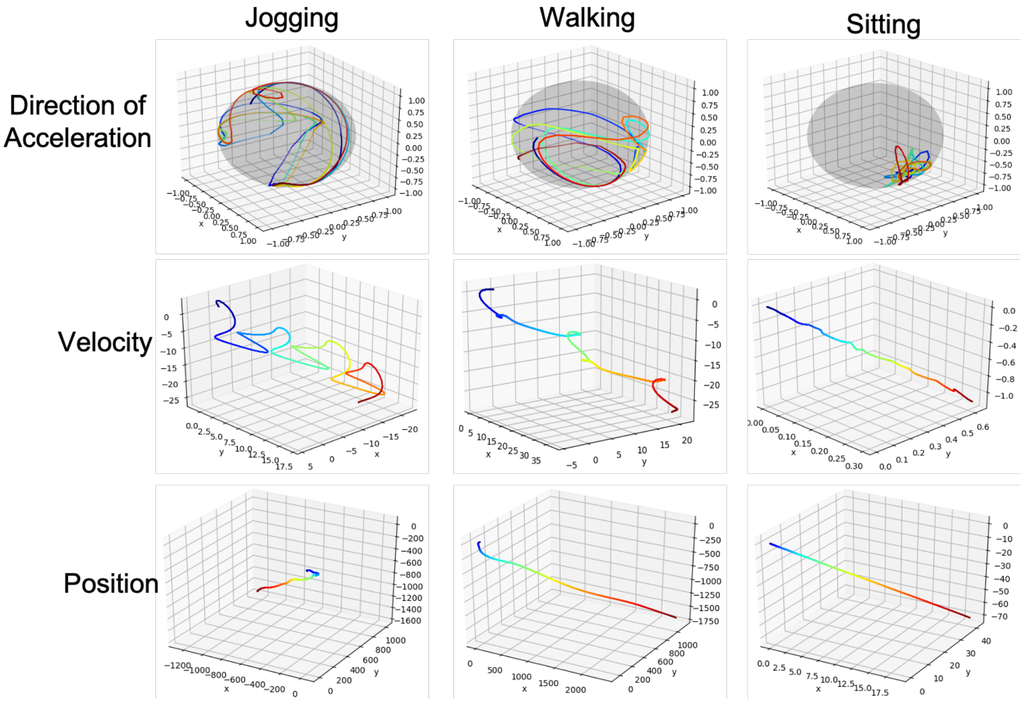


Figure 3: A 2.56 s window of accelerometer data (obtained at 50 Hz) from User 1 in the Motion Sense dataset is examined for jogging, walking, and sitting. The direction of acceleration is shown in the top row, a velocity obtained from integrating the acceleration (with $\mathbf{v}(0) = (0, 0, 0)$) is shown in the middle row, and a position obtained from integrating the velocity is shown in the bottom row (with $\mathbf{x}(0) = (0, 0, 0)$). Each component of acceleration is smoothed using a Gaussian filter with $\sigma = 0.04\,\text{s}$ to reduce high-frequency noise. For each entry, time is implicitly represented using a jet color map going from blue (beginning) to green (middle) to red (end).

Finally, in Figure 4, we examined the x, y, and z components of the accelerometer data separately for the same 2.56 s window as Figure 3. Again, the individual accelerations are smoothed using a Gaussian filter with $\sigma = 0.04\,\text{s}$ to eliminate high frequency noise. The acceleration for the jogging and walking segments much more clearly reflects the inherently periodic nature of these movements than the direction of acceleration plotted in three dimensions in Figure 3. The user appears to undergo

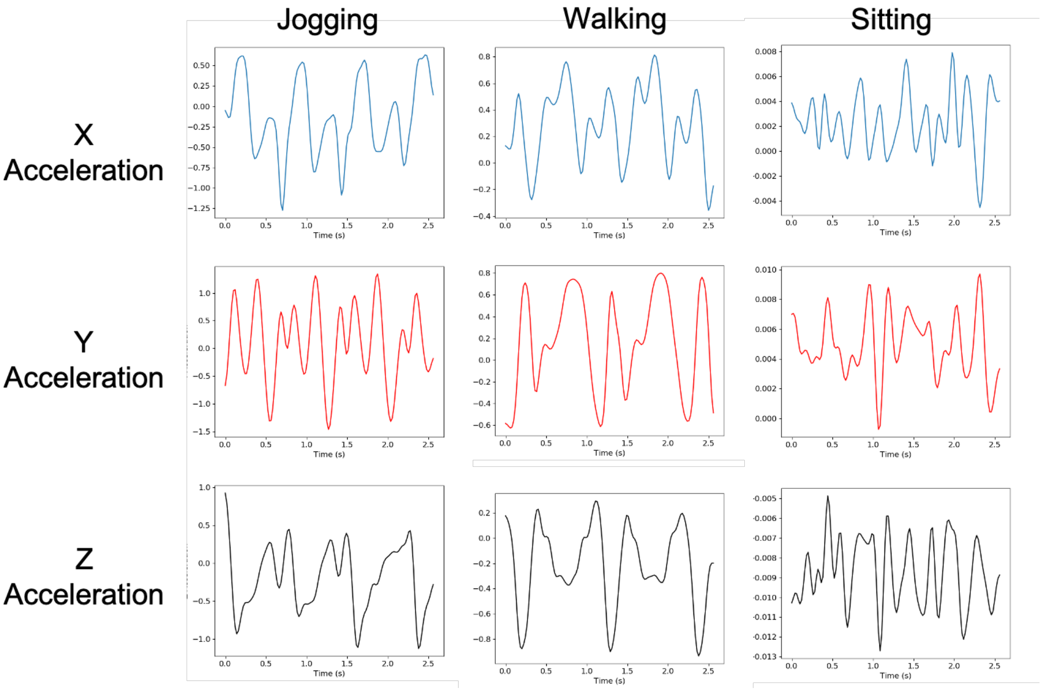


Figure 4: A 2.56 s window of accelerometer data (obtained at 50 Hz) from is examined for jogging, walking, and sitting. The x-, y-, and z-components of acceleration are shown separately and smoothed using a Gaussian filter with $\sigma = 0.04$ s.

about 3.5 cycles of movement in the jogging data and 2.5 cycles in the walking. By contrast, the sitting data does not appear to have a clear periodic structure. Furthermore the amplitude of the accelerations during jogging and walking are orders of magnitude greater than that experienced during the sitting (~ 1 vs. ~ 0.01).

5 Results

5.1 Combined Datasets

After finding consistent results between classifiers for each dataset (Motion Sense and MobiAct), we performed a quick test of how similar the datasets are—we naïvely attempted to use a classifier trained on Motion Sense to predict activity in MobiAct and vice versa. We used only used the accelerometer data, in line with our previous test classifiers. The Motion Sense dataset had removed the acceleration due to gravity from the accelerometer data while the MobiAct did not. In order to have consistent data, we added back in the acceleration due to gravity into the Motion Sense data. After this, we decided to combine the datasets and run our classifiers.

To make as fair a comparison as possible, we only attempted to predict entries that were labeled as an activity that appeared in both datasets. We only evaluated ourselves on the following activities: walking, jogging, sitting, standing, walking upstairs, and walking downstairs. After some deliberation, we decided to increase the feature vector length to 5.12 seconds (i.e. 256 samples), in order to better capture the periodicity of the signals. A summary of our results for the 128- and 256-dimensional feature vectors are given in Tables 3 and 4 below.

Table 3: Cross Dataset Accuracy, 128 Feature Length

Test Dataset	k-NN	Trees
MOTIONSENSE	39.4%	66.5%
MOBIACT	42.9%	66.7%

Table 4: Cross Dataset Accuracy, 256 Feature Length

Test Dataset	**k-NN**	**Trees**
MOTIONSENSE	36.8%	68.6%
MOBIACT	44.3%	77.0%

The naïve approach was surprisingly accurate, after carefully taking into account the difference between them in preprocessing. Unfortunately, our approach for utilizing orientation data and principle component analysis to improve data alignment between datasets worsened the results.

5.2 Data Orientation Processing

As shown in Figure 1b, we tried to devise a method to account for variation in the phone's orientation across datasets. The reported acceleration data is taken in the body frame of the phone. As a result, the x,y, and z acceleration axes shift when one moves, and are different depending on how one holds the phone. The MobiAct and Motion Sense datasets both contain information about the phone's orientation in the form of pitch, roll, and yaw. So, we created a rotation matrix based on the pitch, row, and yaw measurements. The rotation matrix tells us how the phone was rotated relatively to its initial position, which is intrinsic to each phone. To transform the acceleration vectors to be in the initial phone position, regardless of orientation, we need to rotate each of the acceleration vectors back to their initial positions, based on the orientation measurements. Thus, we need to use the inverse of the rotation matrix.

For simplicity, the initial rotation matrix (R_q), representing the current phone orientation, and acceleration vectors were represented as quaternions.

$$q = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} = \begin{bmatrix} \cos(\phi/2)\cos(\theta/2)\cos(\psi/2) + \sin(\phi/2)\sin(\theta/2)\sin(\psi/2) \\ \sin(\phi/2)\cos(\theta/2)\cos(\psi/2) - \cos(\phi/2)\sin(\theta/2)\sin(\psi/2) \\ \cos(\phi/2)\sin(\theta/2)\cos(\psi/2) + \sin(\phi/2)\cos(\theta/2)\sin(\psi/2) \\ \cos(\phi/2)\cos(\theta/2)\sin(\psi/2) - \sin(\phi/2)\sin(\theta/2)\cos(\psi/2) \end{bmatrix} \quad (1)$$

where θ, ϕ, and ψ correspond to the Euler angles representing the phone's roll, pitch, and yaw. Similarly, we have a quaternion representing the acceleration vector as measured by the accelerometer,

$$p = a' + b'\mathbf{i} + c'\mathbf{j} + d'\mathbf{k} = [0, a_x, a_y, a_z]^T \quad (2)$$

where (a_x, a_y, a_z) correspond to the tri-axial acceleration components.

The rotation of acceleration by orientation can be done with quaternions as shown in Eqn. (3) below

$$R_q(p) = q \odot p \odot q^*, \quad (3)$$

where $\odot$ represents the Hamilton Product and * represents the conjugate of the quaternion.

The inversion of the rotation matrix (R_q^{-1}) is equivalent to the conjugate of the corresponding quaternion (q^*).

We rotated each of the acceleration vectors back to their initial positions by using the corresponding orientation measurements, as shown in Eqn. (4) below,

$$R_q(p)^{-1} = q^* \odot p \odot q. \quad (4)$$

Once the initial rotation was completed for each acceleration vector, we applied Principle Component Analysis (PCA) on the rotated components within a single feature vector. PCA allows us to find the axes of the data that contain the most variation. We normalize the axes, and then project the data onto the principle axes. Now, most of the variation in the signal will be isolated on the major axis, rather than across multiple axes. This transformation is shown in Fig. 5. The principle axes with the projected acceleration data is shown in Fig. 6.

5.3 Final Dataset Cross Validation

After processing the acceleration data based on the orientation and PCA rotations, we ran our cross dataset evaluation using the extra-trees and k-NN classifiers. Our results are summarized in Tables 5 and 6 below.

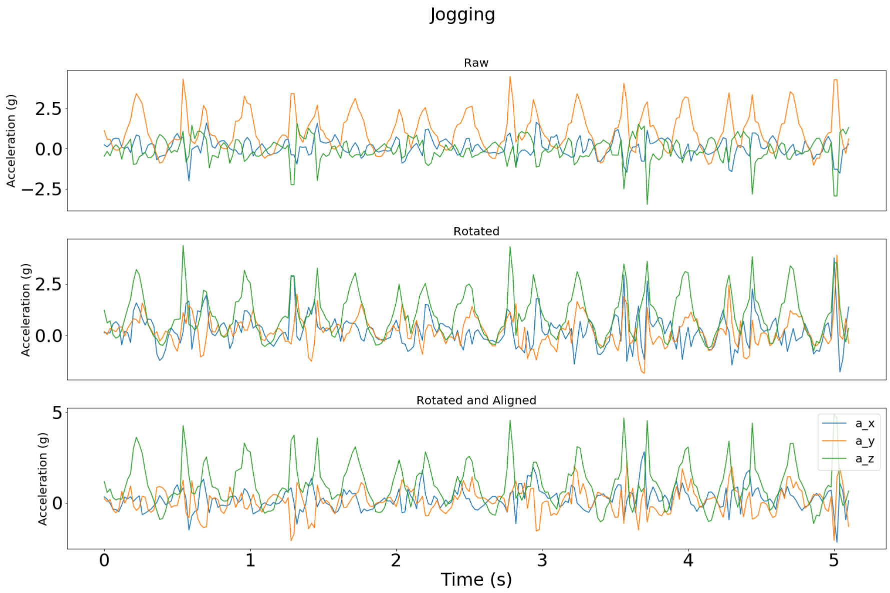


Figure 5: The orientation pre-processing initially utilizes the yaw, pitch, and roll angles to rotate the acceleration data back to the phone's initial position. PCA is then used to find the axis with the most variation, which is then projected onto the z-axis.

Table 5: Cross Dataset Rotated Accuracy, 128 Feature Length

Test Dataset	k-NN	Trees
MotionSense	28.7%	46.2%
MobiAct	34.5%	62.6%

Table 6: Cross Dataset Rotated Accuracy, 256 Feature Length

Test Dataset	k-NN	Trees
MotionSense	27.8%	46.0%
MobiAct	32.2%	52.4%

If we compare these results to the results of the naïve data analysis that are summarized in Tables 3 and 4, then we see that our accuracy actually decreased. There are a few reasons for why this may be the case.

The first reason, we believe, is that there may be an error in our data preprocessing for the MobiAct dataset when we orient the data to the principal components of accelerometer feature vector. Specifically, either the orientation data in unreliable or the description of what the orientation data represents is ambiguous, so by orienting the accelerometer data into the navigational frame, we are actually corrupting the data rather then enhancing it. To test this idea, we tried testing our data on data that has only been rotated to the principal axes of the accelerometer data to the phone's body frame without orienting the accelerometer data by the roll, pitch, and yaw angles. The results of this summarized in Tables 7 and 8 below.

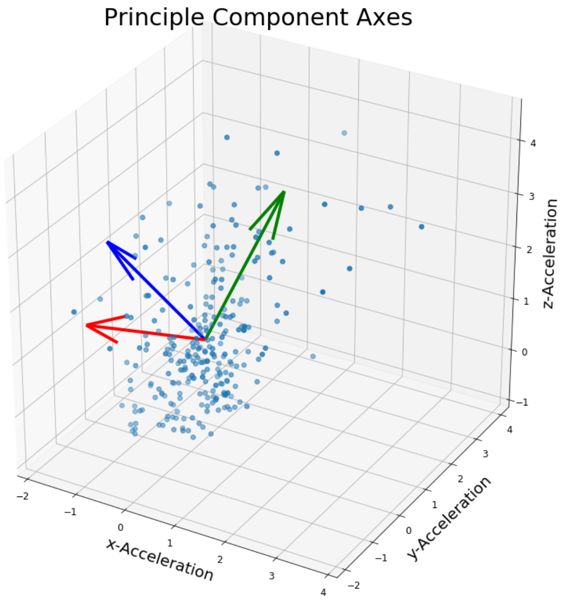


Figure 6: The rotated and aligned acceleration data is plotted with the principle axes it has been projected on.

Table 7: Cross Dataset PCA Accuracy, 128 Feature Length

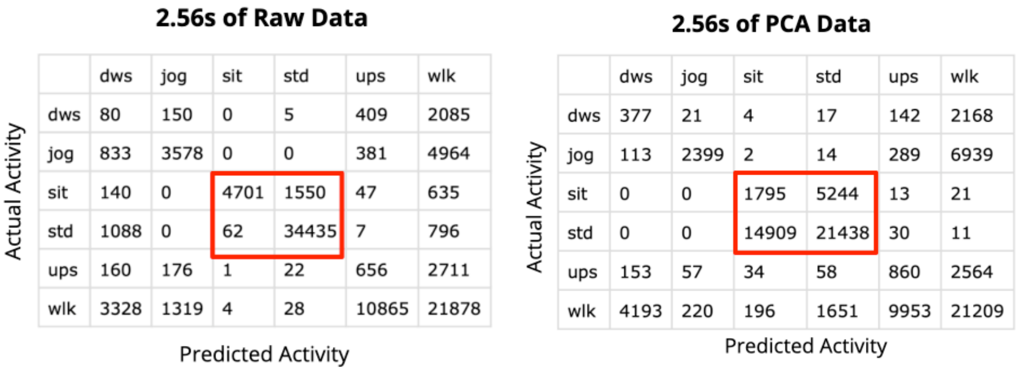
Test Dataset	k-NN	Trees
MOTIONSENSE	25.9%	54.5%
MOBIACT	15.1%	48.1%

Table 8: Cross Dataset PCA Accuracy, 256 Feature Length

Test Dataset	k-NN	Trees
MOTIONSENSE	23.1%	54.7%
MOBIACT	14.6%	53.9%

Notably, these results demonstrate that rotating the data to the accelerometer's principle components also decreases the accuracy of our classifier across datasets. This decrease could be due to a few reasons. First and foremost, for sedentary activities—such as sitting, standing, and laying down—the relative orientation of a smartphone matters. By transforming our data into its principal components, we lose important relative orientation information of the smartphone. When we look at our results more clearly, this is exactly what we find.

In order to isolate the effect of the preprocessing (avoiding differences in the signature of specific activities across datasets), we looked more closely at the performance of the pre-processing on the MOTION SENSE dataset. Figure 7 illustrates that the classifiers of the raw data does a much better job at accurately distinguishing between sitting and standing. However, after transforming the data to its principle axes, we lose this ability as we have implicitly erased the orientation information that the classifier was was depending on.



2.56s of Raw Data

Actual Activity

	dws	jog	sit	std	ups	wlk
dws	80	150	0	5	409	2085
jog	833	3578	0	0	381	4964
sit	140	0	4701	1550	47	635
std	1088	0	62	34435	7	796
ups	160	176	1	22	656	2711
wlk	3328	1319	4	28	10865	21878

Predicted Activity

(a) Raw data.

2.56s of PCA Data

Actual Activity

	dws	jog	sit	std	ups	wlk
dws	377	21	4	17	142	2168
jog	113	2399	2	14	289	6939
sit	0	0	1795	5244	13	21
std	0	0	14909	21438	30	11
ups	153	57	34	58	860	2564
wlk	4193	220	196	1651	9953	21209

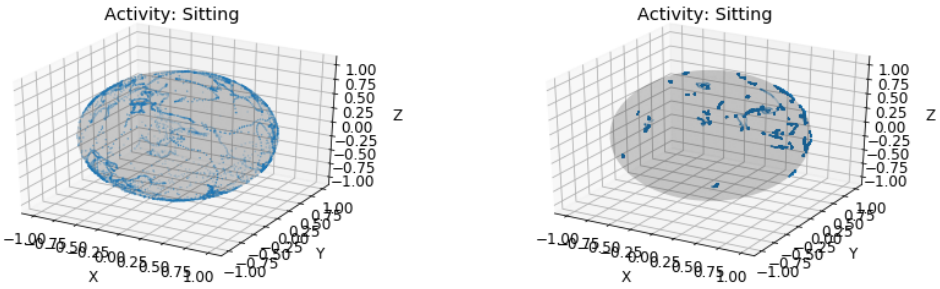
Predicted Activity

(b) Data rotate to Principle Axes.

Figure 7: The confusion matrices for an extra tree classifier trained on the MOTION SENSE dataset, cross validated by user. The red box highlights the major differences in our activities: by rotating the data to the principal components, our classifier is no longer able to distinguish between the sitting and standing activities as both are relatively sedentary and the only major difference is the phone's orientation.

5.3.1 Orientation Visualization

Although our idea of rotating the data based on the orientation angles seemed like a promising approach, our initial results in Sec. 5.3 were far from what we had hoped. We decided to inspect the orientation angles to make sure that they were reasonably similar for the same activities between datasets, as shown in Fig. 8.

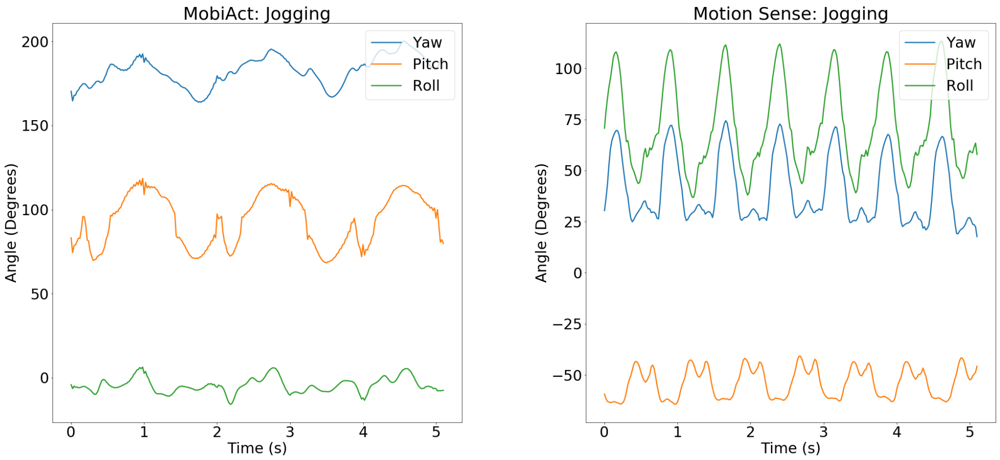


(a) Mobi Act Dataset

(b) Motion Sense Dataset

Figure 8: The images above are based on the orientation data for a random user in each dataset. There is significant difference in the clustering of the orientation data between the two datasets used.

While the Motion Sense dataset had relatively distinct clusters of data, the MobiAct data was much more spread out. We then visualized the orientation angles individually, as shown in Figure 9. This furthered our suspicion that the given angles for the two datasets had different conventions and/or had different ranges of operation. Unfortunately, there was not any information on either of these crucial components.



(a) Mobi Act Dataset

(b) Motion Sense Dataset

Figure 9: The images above are based on the orientation angles for a random user in each dataset.

5.4 Final Verdict on Orientation Processing

Ultimately, we believe that the orientation processing, as we presented it in this section, introduces more error into our dataset than it is worth. The combination of an ambiguous pitch, roll, and yaw measurements in the dataset with the unpredictability of PCA on our clearly non-linear data gives us pause to continue down this path. However, our goal remains the same: to create orientation and time independent features, in order to create a human activity recognition algorithm that is robust to different orientations of the phone.

To do this, we plan on using the curvature and torsion of velocity as coordinate invariant representations of the phone's movement. The question still remains whether we need to transform the phone from a body frame to a navigation frame to do this effectively. For these coordinate invariant features, we only consider data from a 0.06 s period (3 data points), so we believe that the variation in orientation is small enough that adjusting for orientation is not necessary. Indeed, due to the uncertainties in the orientation data, we believe that this could cause more harm than good.

6 Invariant Features

Although we have mostly focused on processing the raw accelerometer data, we have also considered introducing the following ideas to more accurately capture the behavior of the data. Our goal would be to create invariant features from our raw accelerometer data. In many circumstances, it actually makes more sense to analyze the velocity of the phone's body frame—i.e. the integral of the accelerometer data. When we integrate accelerometer data, we get a nice parameterized curve in space as illustrated nicely in Fig. 10.

6.1 Coordinate-Invariant Features

Perhaps we can create features based on torsion and curvature measurements for sub-segments of a single feature vector. We can take each of our accelerometer measurements. Since these are geometric quantities, they do not depend on any reference frame. So, for example, we can plot the magnitude of the torsion of the velocity to get a new feature that is independent of any coordinate system. Similarly, we can look at the magnitude of the torsion vector and get another, more distinct feature from the velocity. Both the torsion and curvature of a specific user is illustrated in Figure 11 below.

The periodic structure of Fig. 11 implies that we should be looking for periodic patterns in our data, as it has important, activity-dependent information contained in it. This leads us to consider another type of feature—one that is time-invariant.

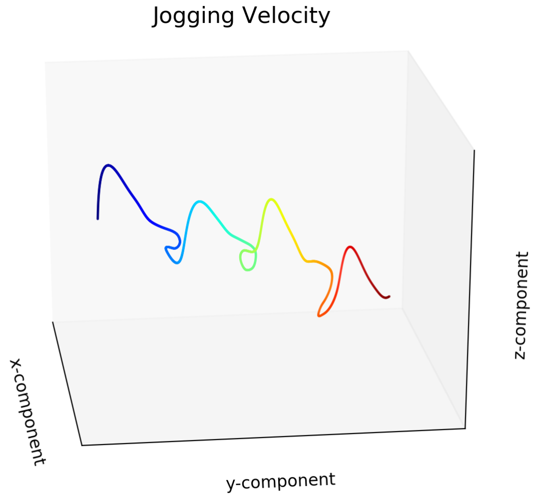
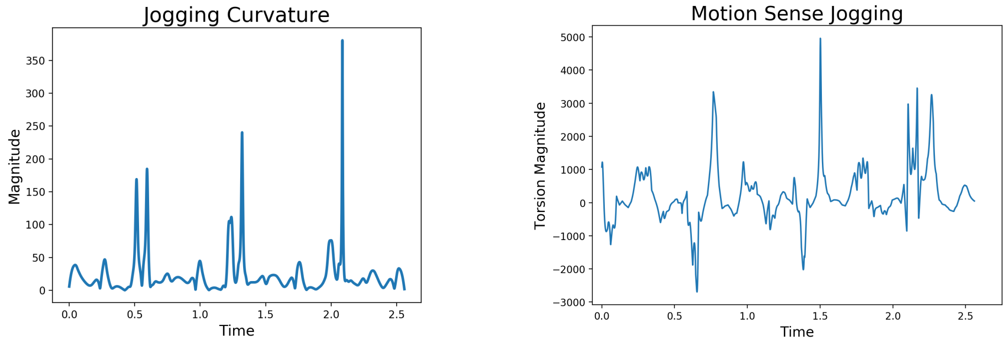


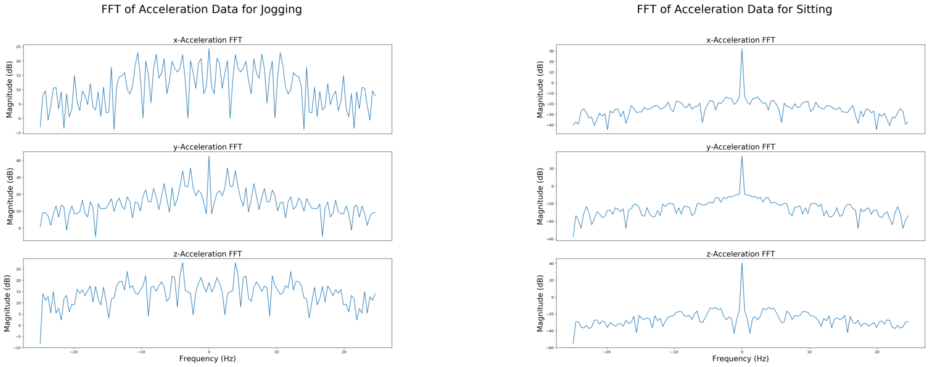
Figure 10: An example of a user's jogging accelerometer data integrated. Notice the inherent drift and quasi-periodic structure. The blue color represents time at $t = 0$ and the red represents time at $t = 2.56\,\mathrm{s}$.



(a) Curvature.

(b) Torsion.

Figure 11: The magnitude of the torsion of velocity parameterized by time. This is a coordinate-invariant which is not dependent on the choice of axes.



(a) FFT of a jogging feature vector.

(b) FFT of a sitting feature vector.

Figure 12: The FFTs of the two aforementioned activities look very different. While sitting is primarily made up a DC component, due the lack of movement, the jogging signal has a variety of frequencies with large magnitudes.

6.2 Time-Invariant Features

As shown in Fig. 3, the processed acceleration data is actually quite periodic, with the frequency of the signal changing for each activity. However, by including the raw accelerometer data in our feature vectors (in line with other researchers in the field), our classifier will be sensitive to where in the period of someones gait a time-window is centered on. By viewing the accelerometer data in the frequency domain instead of in the time domain, we could reduce the sensitivity to where an individual sample is centered. Researchers have previously have utilized discrete Fourier transform based features for accelerometer data [1], although they stop short of directly using a power spectra in lieu of the time series. Figure 12 shows a comparison of an FFT for jogging and sitting.

6.3 Coordinate and Time Invariant Features

As stated in sections 6.1 and 6.2, there is likely to be significant gain from using coordinate and time invariant features. Figure 11 shows that there is some periodicity within the coordinate invariant features of curvature and torsion. Thus, if we were to do the FFT of these features, we would then have features that are time and coordinate invariant.

6.4 Hand-Crafted Features

While coordinate- and time-invariant features are potentially powerful tools to distinguish distinct activities—e.g. running vs. sitting—Figure 8a suggests that similar behaving activities may be difficult to identify (e.g. standing and sitting). Thus, phone orientation is important. Because of this, we decided to create a hand-crafted feature vector to train our extra trees classifier on. This hand-crafted feature was designed such that both the coordinate- and time-invariant features are combined features that are orientation dependent. The feature vector a 257-length feature vector of the following:

1. The power spectra of the fast Fourier transform of the magnitude of curvature of the parameterization of the velocity as a function of time. Curvature, quite simply, is the measure of how much a curve deviates from a straight line. Its mathematical definition is
$$\kappa = \frac{\|\boldsymbol{\gamma}' \times \boldsymbol{\gamma}''\|}{\|\boldsymbol{\gamma}'\|^3}$$
where $\boldsymbol{\gamma} = (x(t), y(t), z(t))$. In our case we are actually looking at the velocity of the phone's body frame. Figure 11a that there is a highly periodic structure of the curvature's magnitude, and this is our reasoning for looking at the curvature in frequency space. In effect, this measures how quickly a phone changes direction in three-dimensional space. Due to the symmetry of FFTs, this accounts for 128 of our 257 dimensions.
2. Similar to above, we are taking real components of the FFT of the magnitude of the torsion of the three-dimensional velocity curve. The torsion is the measure of how quickly a curve is twisting out of the plane of curvature—if it is twisting upwards from the plane of curvature, it is positive and it is negative if it is twisting outwards. The formula for torsion is given by
$$\tau = \frac{\det(\boldsymbol{\gamma}', \boldsymbol{\gamma}'', \boldsymbol{\gamma}''')}{\|\boldsymbol{\gamma}' \times \boldsymbol{\gamma}''\|^2},$$
where $\boldsymbol{\gamma}$ is the velocity parameterization again. This will hopefully account for changes in plane (e.g. walking up vs. down stairs). It accounts for another 128 of our 257 dimensions.
3. Now we will look at the average acceleration vector $|\langle \mathbf{A} \rangle| \equiv (|\langle A_x \rangle|, |\langle A_y \rangle|, |\langle A_z \rangle|)$. This accounts for three of our 257 dimensions and will provide us with orientation data to distinguish sitting and standing as we expect standing will primarily be in the A_y direction, hence the absolute value for an entire time series.
4. The remaining features are the L_1-norm of each component of the time series (three dimensions), the standard deviation of the average of each component (three dimensions), the average magnitude of the acceleration (one dimension), standard deviation of the magnitude of the acceleration magnitude (one dimension), the average of the standard deviation of each component (three dimensions).

This, hopefully, will provide us with enough information to more accurately characterize human activities.

6.5 Custom Data Collection

We are attempting to flesh out an orientation independent representation of accelerometer data. However, the datasets that we are using are both gathered from data with the phone in a user's front pocket, and the MOTION SENSE even tells them exactly how to put it in their pocket. As shown in Table 4, the ExtraTrees classifier performed surprising well on the raw data for the cross dataset classification. If the phones were also orientated in the same direction (screen up vs screen down), then perhaps the cross data performance isn't so surprisingly. The drop from within dataset performance might simply be caused by the difference in the phone sensors. In addition, the MOBIACT dataset was much more noisy than MOTIONSENSE, which could also contribute to the observed performance difference.

To confirm that the phones were orientated in the same direction in the users' pockets, we decided to take our own data. We plan on taking data doing the same six activities, while varying the phone orientation, from screen up to screen down. If both datasets were taken with the phone in the same orientation, our personal dataset should perform similarly to the cross dataset results when our orientations match, but much worse when the orientations are completely opposite.

If our classifier is truly coordinate invariant, we expect there to be no difference in performance between the two different orientation datasets that we collected.

Acknowledgments

We would like to thank Prof. Gu for fruitful discussions on how we could improve and extend our analysis of the data.

References

[1] Anna Ferrari, Daniela Micucci, Marco Mobilio, and Paolo Napoletano. Hand-crafted features vs residual networks for human activities recognition using accelerometer. pages 153–156, 2019.

[2] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. Protecting sensory data against sensitive inferences. pages 1–6, 2018.

[3] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–30, 2019.

[4] Zdeňka Sitová, Jaroslav Šeděnka, Qing Yang, Ge Peng, Gang Zhou, Paolo Gasti, and Kiran S Balagani. Hmog: New behavioral biometric features for continuous authentication of smartphone users. *IEEE Transactions on Information Forensics and Security*, 11(5):877–892, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 2016.

[6] Anna Ferrari, Marco Mobilio, Daniela Micucci, and Paolo Napoletano. On the homogenization of heterogeneous inertial-based databases for human activity recognition. 2642:295–300, 2019.

[7] George Vavoulas, Charikleia Chatzaki, Thodoris Malliotakis, Matthew Pediaditis, and Manolis Tsiknakis. The mobiact dataset: Recognition of activities of daily living using smartphones. In *ICT4AgeingWell*, pages 143–151, 2016.

[8] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge Luis Reyes-Ortiz. A public domain dataset for human activity recognition using smartphones. In *Esann*, 2013.

[9] John R Apel. *Principles of ocean physics*. Academic Press, 1987.

[10] Abdul Kadar Muhammad Masum, Erfanul Hoque Bahadur, Ahmed Shan-A-Alahi, Md Akib Uz Zaman Chowdhury, Mir Reaz Uddin, and Abdullah Al Noman. Human activity recognition using accelerometer, gyroscope and magnetometer sensors: Deep neural network approaches. pages 1–6, 2019.

[11] Colin M. Adams. “The Glory Dayz: An Oral Account of the Recent History of CMS Cross Country”. A very formal interview by Eli J. Weissler.