Oct 10, 2018

Assignment 2 Report

Data Mining

CSE 572

Fall 2018

Submitted to:

Prof. Ayan Banerjee

Ira A. Fulton School of Engineering

Arizona State University

Submitted by:

Srinivas Vallabhaneni

Akhil Madamala

Divya Madhuri Reddy Gurram

Vaishali Reddy Kankanala

Gourab Mitra

October 10, 2018

Table of Contents

1.Introduction

In this project, we are attempting to develop a computing system that can **understand** human activities. There are a few components to understanding human activities: a) identify known activities, b) segment sequence of activities, and c) identify unknown activities.

The answers to the above are found by employing five data analysis methods and comparing their results:

1)Discrete Wave Transform

2)Fast Fourier Transform

3)Mode

4)Root Mean Squared

5)Standard Deviation

Team members

1)Srinivas Vallabhaneni

2)Akhil Madamala

3)Divya Madhuri Reddy Gurram

4)Vaishali Reddy Kankanala

5)Gourab Mitra

3. Assignment 1

The Assignment 1 Consists of 3 tasks as follows

The assignment 1 is for collecting the sensor data. All the team members should go to Impact lab for collecting the data. Any one of the team member should volunteer for gesturing in front of the screen wearing 2 wristbands, one on right and one on left hand. The gestures are identified by the sensors using the hand movement of the person wearing the wristbands. The gestures are captured using 4 sensors Gyroscope, Accelerator, EMG Sensor, Orientation. According to movement of the hand, each sensor captures the data, for example, if the gesture consists of rotations, orientation sensor data is useful to identify the gesture.

(a) Data collection

Currently, we will be using Myo sensors for this project. However, for this phase of the project we are using only sample data that has been provided. Two students in the group would later wear one Myo wristband on their dominant arm for 1 day. The person would perform one of the activities he or she wants. The person is requested to note down the time at which he or she starts the particular activity and the time he or she completes it. Optionally, we will choose one more activity that you are interested in and log the time stamps for such activity.

b) Phase 2 feature extraction

For each type of feature extracted do the following things,

A) write an explanation on how the feature is extracted.

B) Write an intuition on why you use such a feature.

C) Write a matlab code to extract that feature from each time series stored in the csv files created in task 1.

D) Generate plots each corresponding to each activity. For multiple actions of the same type you can choose to overlap the plots. This will give you a better idea of potential patterns in the features.

E) Discuss whether your initial intuition about the features that you selected holds true or not.

A) We selected and implemented five existing feature extraction methods as follows:

1)Discrete Wave Transform

2)Fast Fourier Transform

3)Mode

4)Root Mean Squared

5)Standard Deviation

We employed the following Feature Extraction method to obtain our desired values:

1) We have considered the datasets of each activity and taken their instances together namely Cooking1, Cooking2, drive1, drive2, eat1, eat2, EatFood1, EatFood2, EatFood3, EatFood4, keyboard1, keyboard2, NoMovement1, NoMovement2, Playing the Guitar.
2) From the above activities we grouped together activities to make seven specific group of activities.
3) Henceforth, we took attribute values of different activities and plotted the magnitude values of each type of sensor data to find which activity exhibited most deviation in values.
4) Once the above is ascertained, we figured out that this particular feature for the activity is to be considered for extraction.

B) We use the features using the following analysis methods for the following:

1)Discrete Wave Transform

The discrete wavelet transform (DWT) gives information about the frequency (actually, basis) components as well as being able to indicate what time these components occur at.

2)Fast Fourier Transform

The fast Fourier transform (FFT) is a computationally efficient method of generating a Fourier transform. The main advantage of an FFT is speed, which it gets by decreasing the number of calculations needed to analyze a waveform

3)Mode

It is easy to understand and simple to calculate.It is not affected by extremely large or small values.It can be located just by inspection in ungrouped data and discrete frequency distribution.

4)Root Mean Squared

The root mean square is a measure of the magnitude of a set of numbers. It gives a sense for the typical size of the numbers.

5)Standard Deviation

The standard deviation is a good measure of variation. It is based on every item of the distribution. You can do algebraic operation and is less affected by fluctuations of sampling than most other measures of dispersion.

C) Matlab code will be attached in Blackboard along with the report.

D)

We also employed another Feature Extraction by intuition method to evaluate the same:

1) Now again we consider the datasets of each activity and take their instances together namely Cooking1, Cooking2, drive1, drive2, eat1, eat2, Eatfood1, EatFood2, EatFood3, EatFood4, keyboard1, keyboard2, NoMovement1, NoMovement2, Playing the Guitar.
2) From the above activities we grouped together activities to make seven specific group of activities.
3) Henceforth, for each of the activities we appended the values of datasets of different sensors into a single matrix. Thus, we are able to obtain dataset values for all the sensor values for each activity in a single matrix. Thus we are able to obtain a matrix with dimension [7*21].
4) Then we take each of these matrices and perform different data analysis methods on each individual columns of individual matrices namely:

   1)Discrete Wave Transform
   2)Fast Fourier Transform
   3)Mode
   4)Root Mean Squared
   5)Standard Deviation
5) Then we consider the 5 most significant values in the resultant matrix with the help of variance.

1)Discrete Wave Transform

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0.211691	0.127365	0.086812	0.311868	0.285156	0.149256	0.098783	0.117852	0.28482	0.373047	0.120663	0.569685	3.300741	0.809524	0.209524	3.638095	2.114286	0.314286	0.6	0.12381	0.066667

2)Fast Fourier Transform

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	0.050768	0.071991	0.037244	124.4153	114.1615	349.0899	0.026421	0.021678	0.050668	0.044656	0.101398	0.119155	0.701832	7.8625	18.41687	60.91691	21.70208	4.317629	6.480334	4.813805	9.914155

3)Mode

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
18834294	41417284	34127382	4.8E+10	3.92E+10	1.28E+11	10928186	14617349	32464068	43854750	87445826	25269889	4.07E+08	8.97E+10	2.11E+11	4.85E+11	1.76E+11	9.97E+10	6.09E+10	5.21E+10	5.68E+10

4)Root Mean Squared

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0.211691	0.127365	0.086812	0.311868	0.285156	0.149256	0.098783	0.117852	0.28482	0.373047	0.120663	0.569685	3.300741	0.809524	0.209524	3.638095	2.114286	0.314286	0.6	0.12381	0.066667

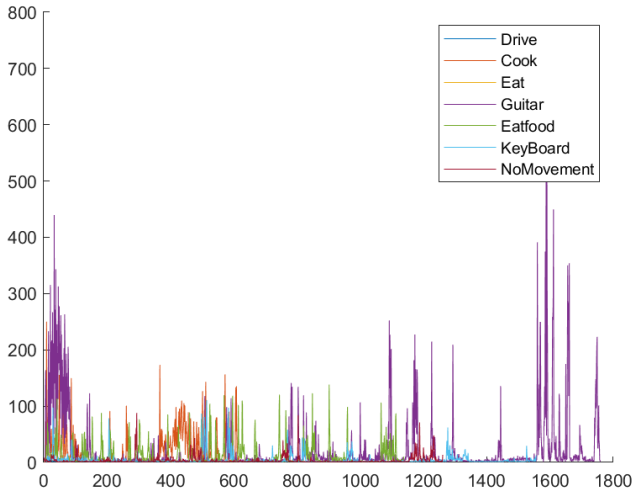
5)Standard Deviation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
	0.050768	0.071991	0.037244	124.4153	114.1615	349.0899	0.026421	0.021678	0.050668	0.044656	0.101398	0.119155	0.701832	7.8625	18.41687	60.91691	21.70208	4.317629	6.480334	4.813805	9.914155

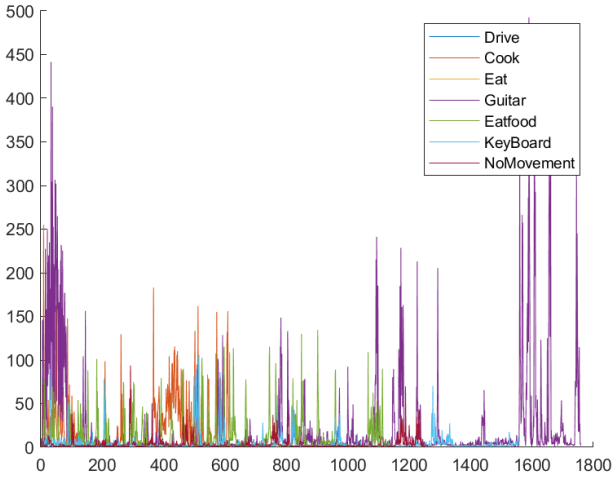
For Discrete Wave Transform(DWT) the following graphs were obtained:

By performing Discrete Wave Transform, Fast Fourier Transform, Mode, Root Mean Squared and Standard Deviation for specific features we found wide variations of values when expressed graphically. Some features were dominant over others and the same was plotted in matlab and provided in the report.

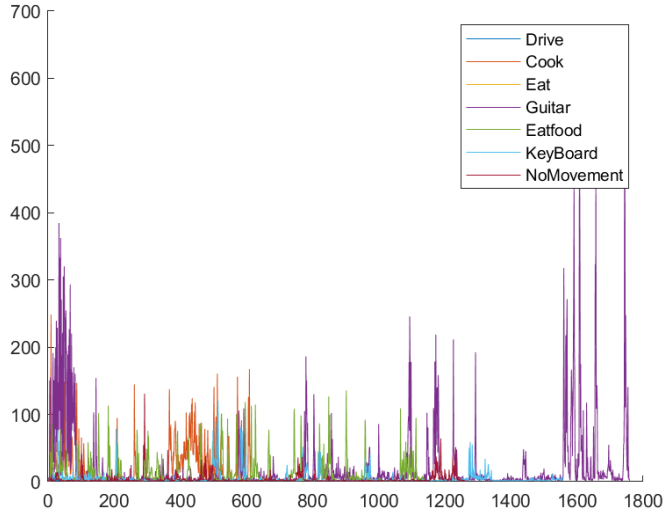
a.gyroz



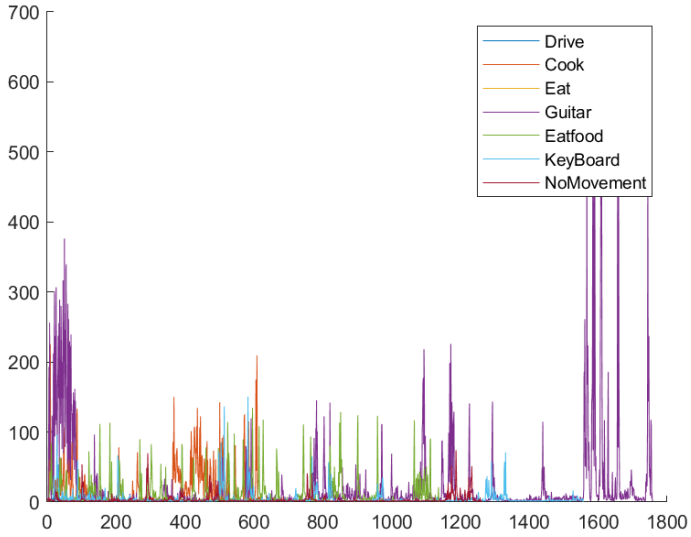
b.gyrox



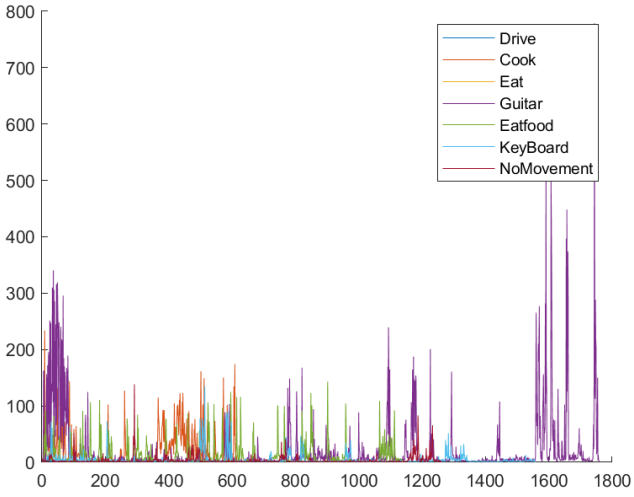
c.gyroy



d.emg2

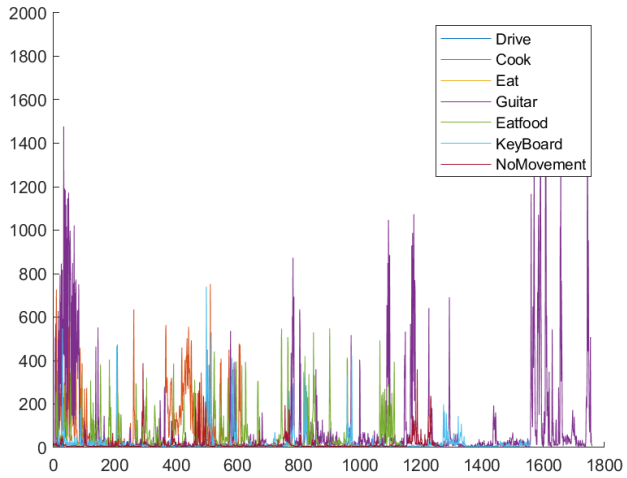

Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement
700
600
500
400
300
200
100
0
0
200
400
600
800
1000
1200
1400
1600
1800

e.emg3

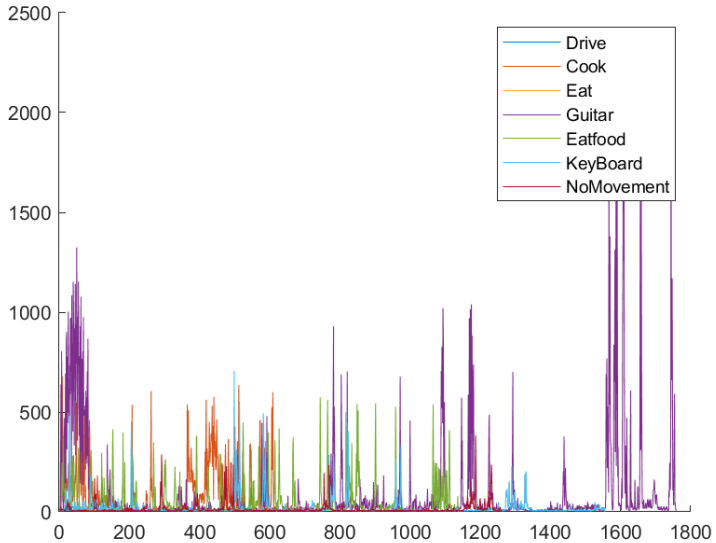

Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement
800
700
600
500
400
300
200
100
0
0
200
400
600
800
1000
1200
1400
1600
1800

For Fast Fourier Transform(FFT) the following graphs were obtained:

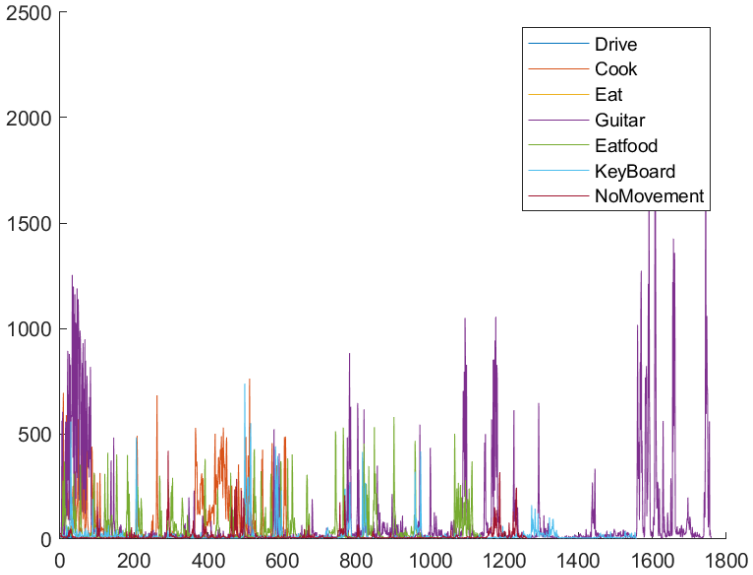
a.gyroz



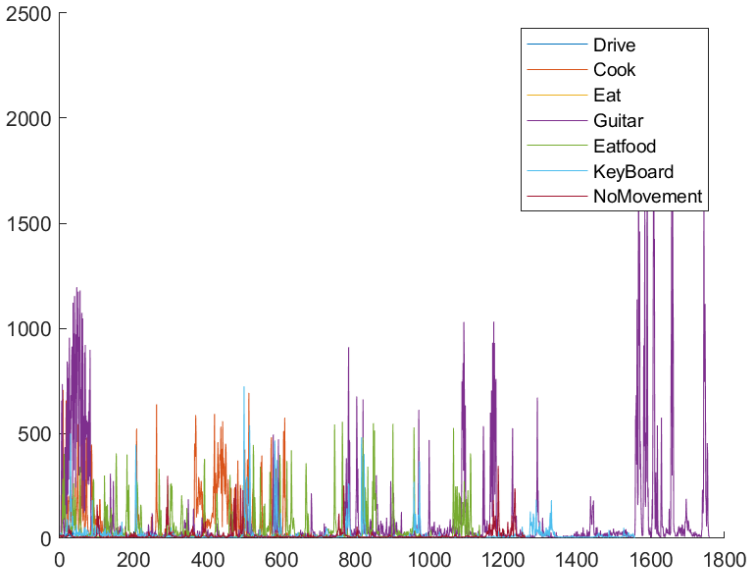
b.emg2



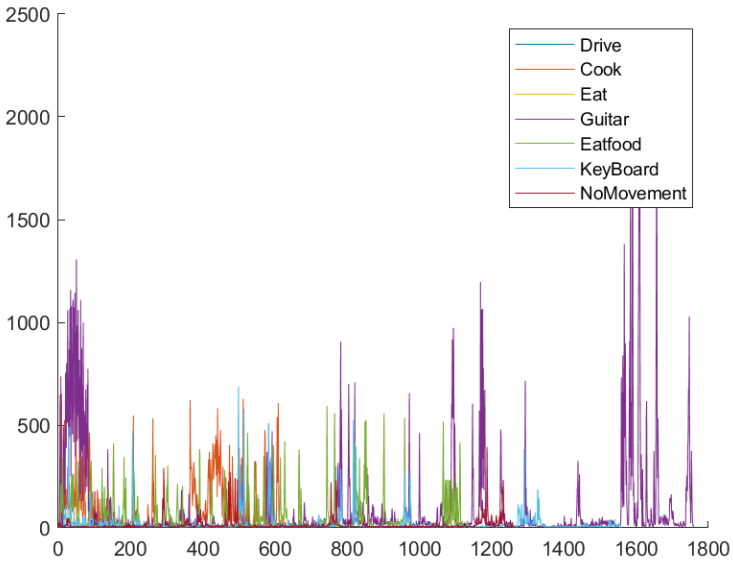
c.emg3


2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

d.emg4

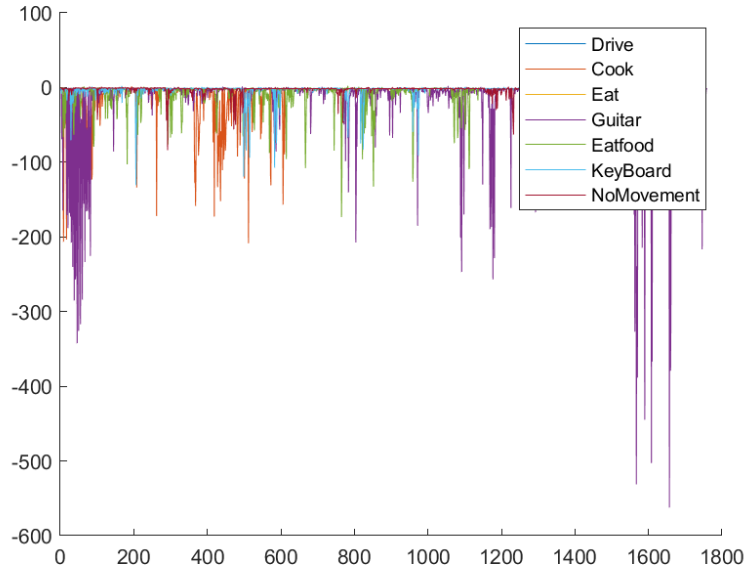

2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

e.emg5

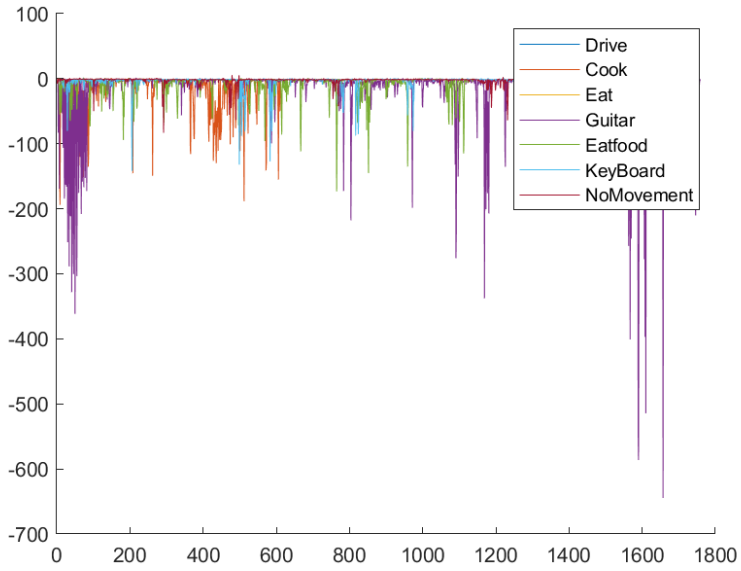


For Mode the following graphs were obtained:

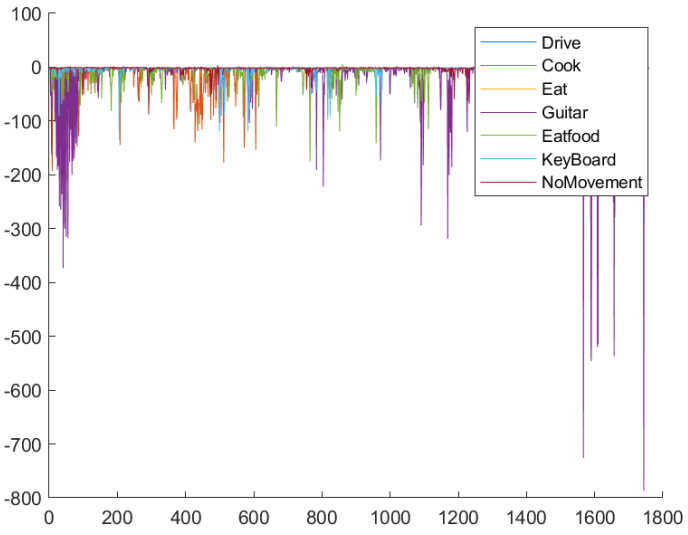
a.oriz



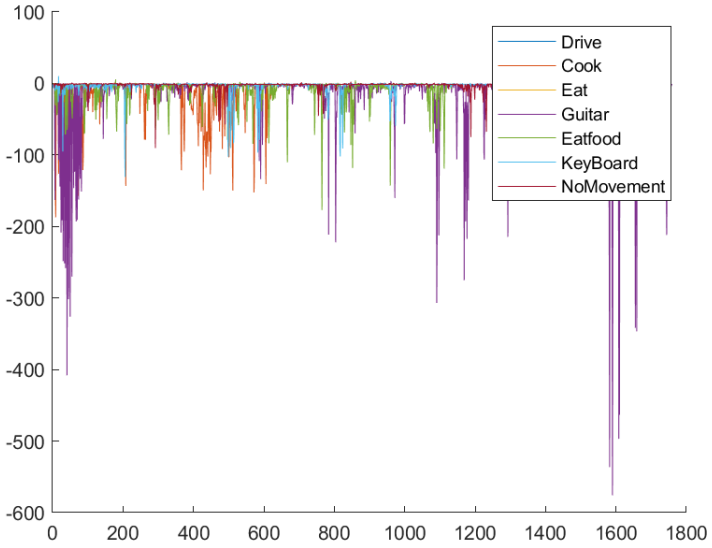
b.emg1


100
0
-100
-200
-300
-400
-500
-600
-700
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

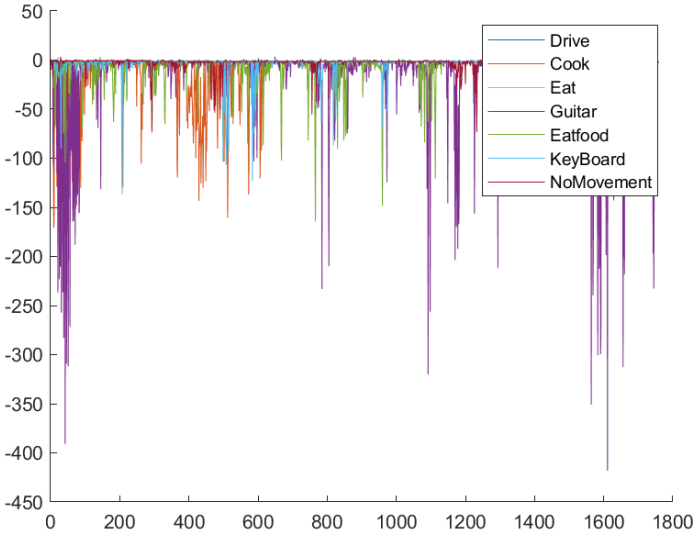
c.emg3


100
0
-100
-200
-300
-400
-500
-600
-700
-800
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

d.emg4

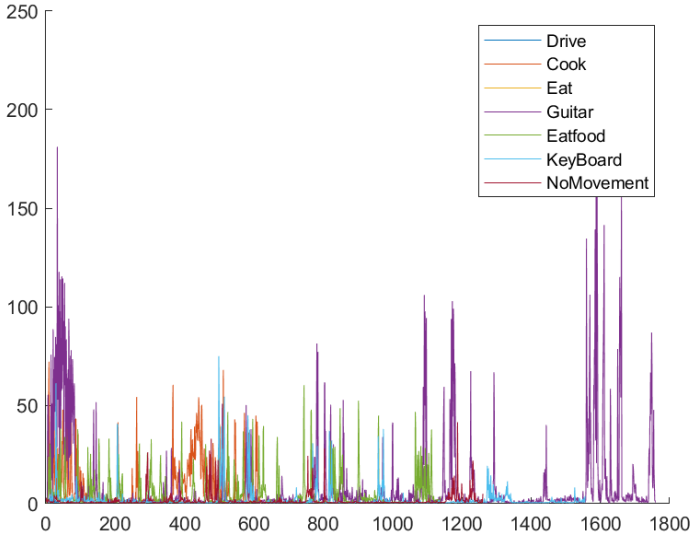

100
0
-100
-200
-300
-400
-500
-600
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

e.emg6

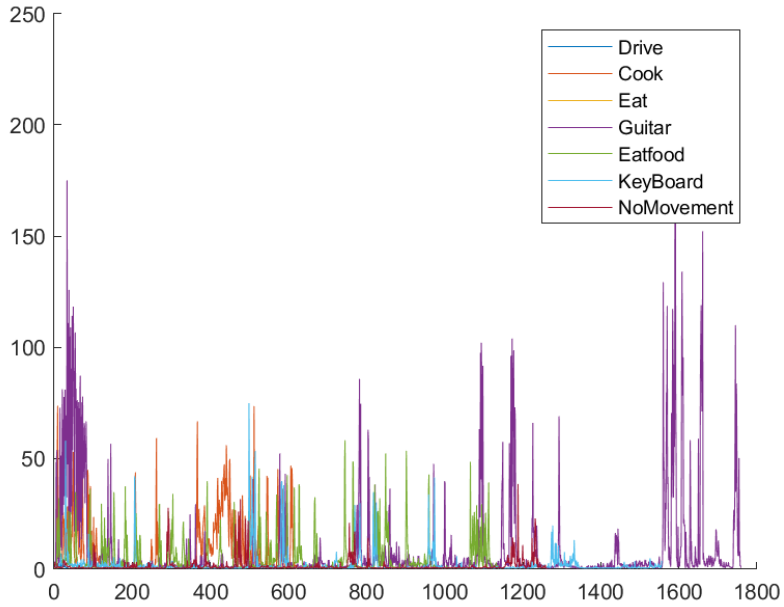

50
0
-50
-100
-150
-200
-250
-300
-350
-400
-450
0
200
400
600
800
1000
1200
1400
1600
1800
Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement

For Root Mean Squared the following graphs were obtained:

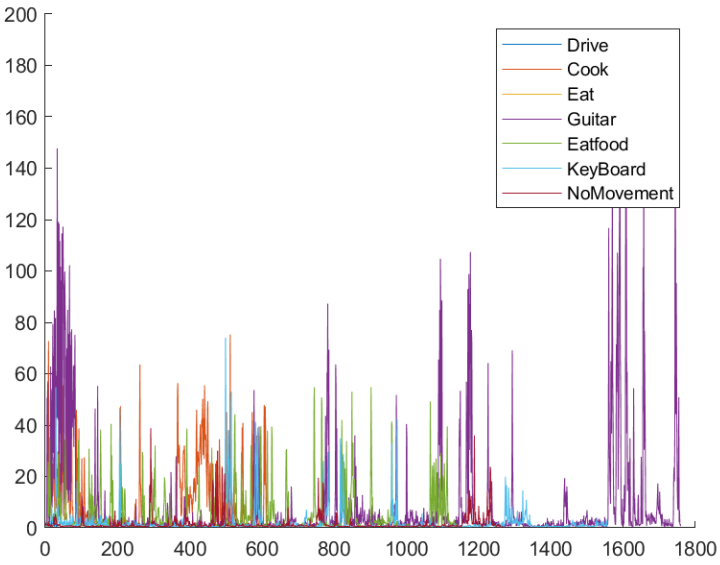
a.gyrox



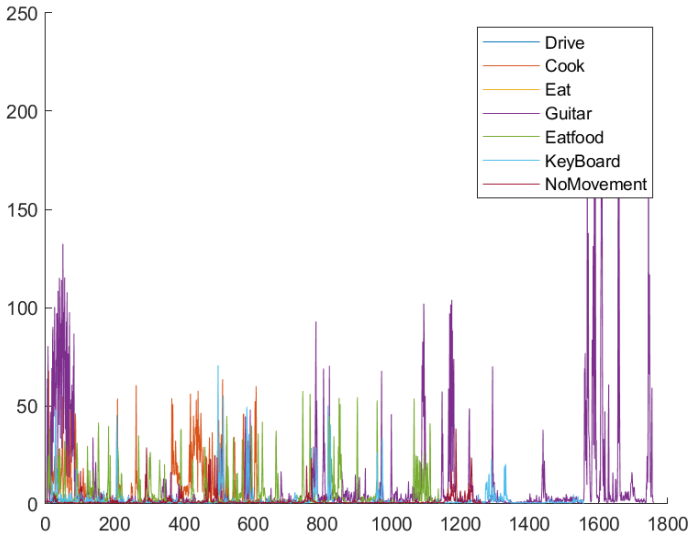
b.gyroy



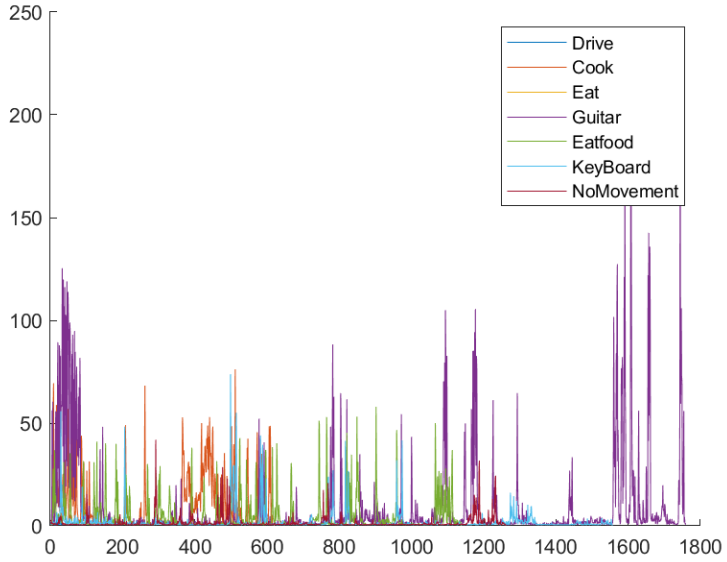
c.gyroz



d.emg3

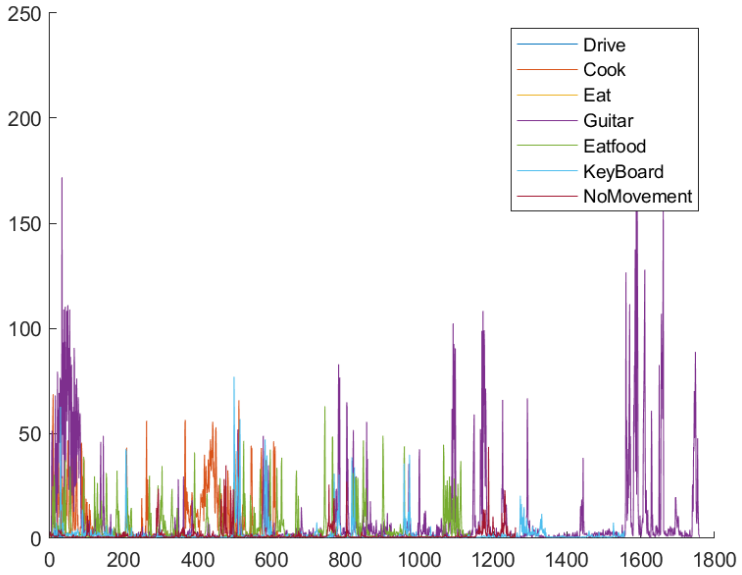


e.emg4

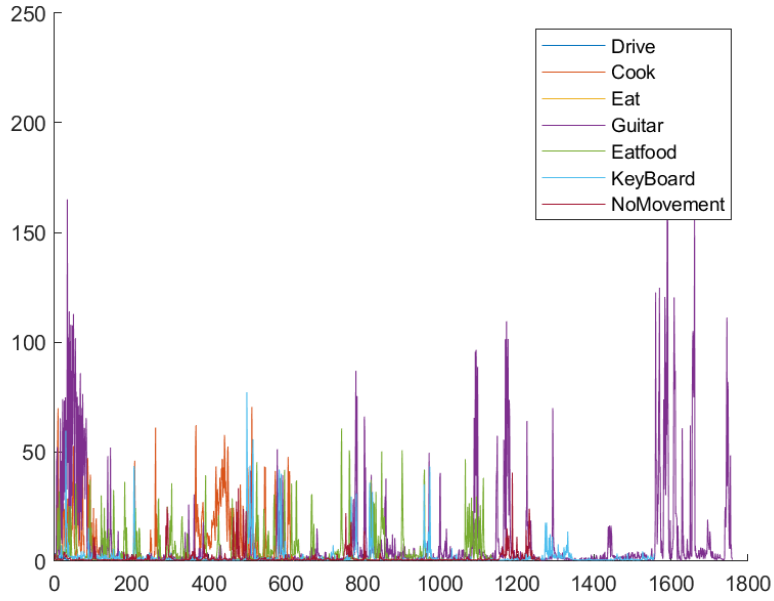


For Standard Deviation the following graphs were obtained:

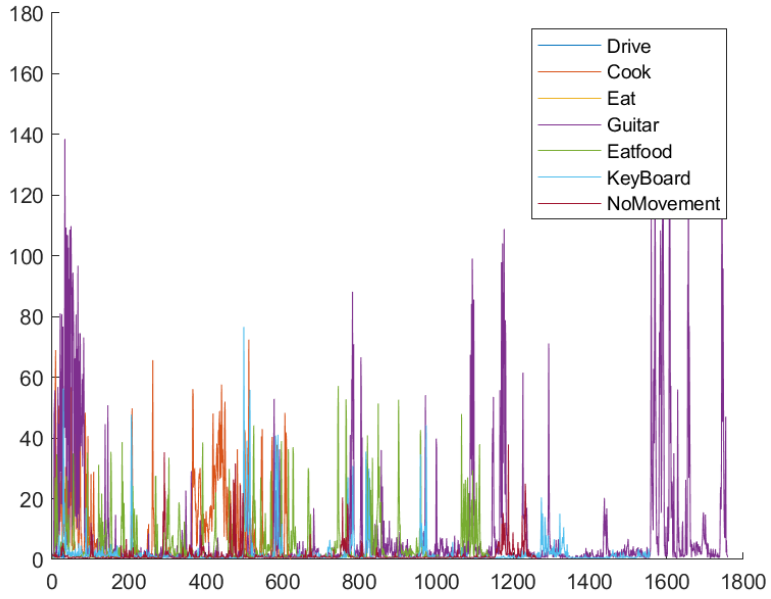
a.gyrox


Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
1600
1800

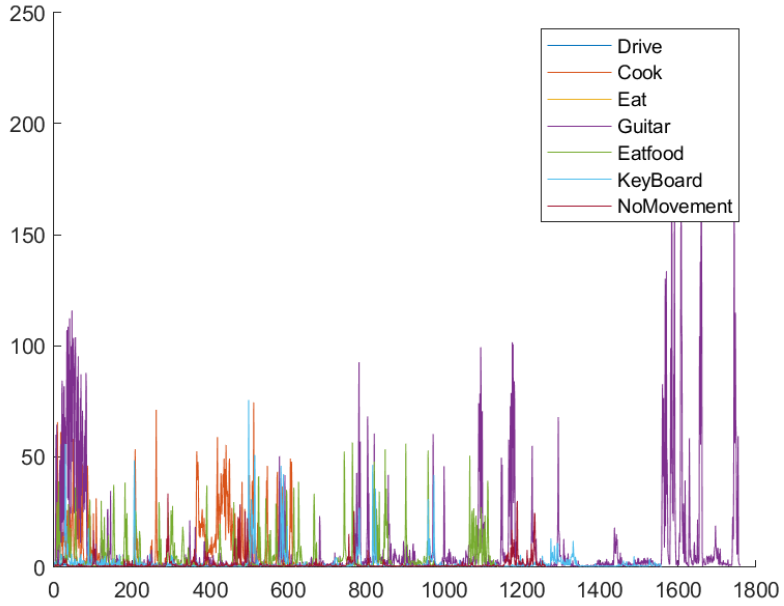
b.gyroy


Drive
Cook
Eat
Guitar
Eatfood
KeyBoard
NoMovement
250
200
150
100
50
0
0
200
400
600
800
1000
1200
1400
1600
1800

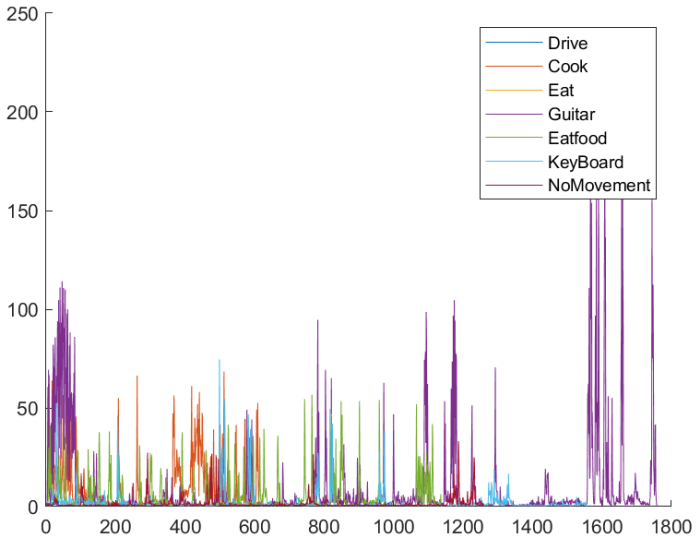
c.gyroz



d.emg3



e.emg6



E) Our initial intuition about the features holds true because we found that only for specific features we found wide variations of values when expressed graphically. Some features were dominant over others and the same was plotted in matlab and provided in the report.

Phase 3: Feature Selection

This step involves reduction of the feature space and keeping only those features which show maximum distance between the two classes. We will use Principal Component Analysis technique discussed in class for this purpose. The PCA code is already available in Matlab, hence there is no need to PANIC! Just use it.

Subtask 1: Arranging the feature matrix

You know PCA only takes one matrix. How will you arrange all sensors and their corresponding features into a single matrix such that the eigenvectors of the covariance matrix directly makes sense to your data set? This means that if the PCA results gives you a eigen vector then the new feature matrix can be obtained by simply multiplying the eigen vector with the old feature matrix. (You might need ten matrices corresponding to the ten classes)

Write your logic of feature matrix arrangement.

For creation of the Feature Matrix we have used the following logic and performed the below steps:

1) We have considered the datasets of each activity and taken their instances together namely Cooking1, Cooking2, drive1, drive2, eat1, eat2,Eatfood1, EatFood2, EatFood3, EatFood4, keyboard1, keyboard2, NoMovement1, NoMovement2, Playing the Guitar.
2) Henceforth, for each of the activities we appended the values of datasets of different sensors calculated by FFT, DWT, STD Mode and RMS into corresponding single matrices for every action as stated below.
3) Then we take each of these matrices and perform different data analysis methods on each individual columns of individual matrices namely:

   1)Discrete Wave Transform
   2)Fast Fourier Transform
   3)Mode
   4)Root Mean Squared
   5)Standard Deviation

   As a result we are able to obtain 7 separate matrices corresponding to the 7 activities with a dimension [2*105].

   Then we apply covariance function to formulate the feature matrix of each individual activity.
4) From the equation :
   (a, b, c)= PCA(Cov matrix)

   we get the Eigen valus in the values of c which has a matrix dimension of 1*105
5) Henceforth, we select the Ten most dominant values of the Eigen values thus obtained.
6) After obtaining the top 10 eigen values we need to obtain the corresponding eigen vectors of those top eigen vectors
   We found the eigen vectors using the following statementt
   The equation for the above stands as follows:

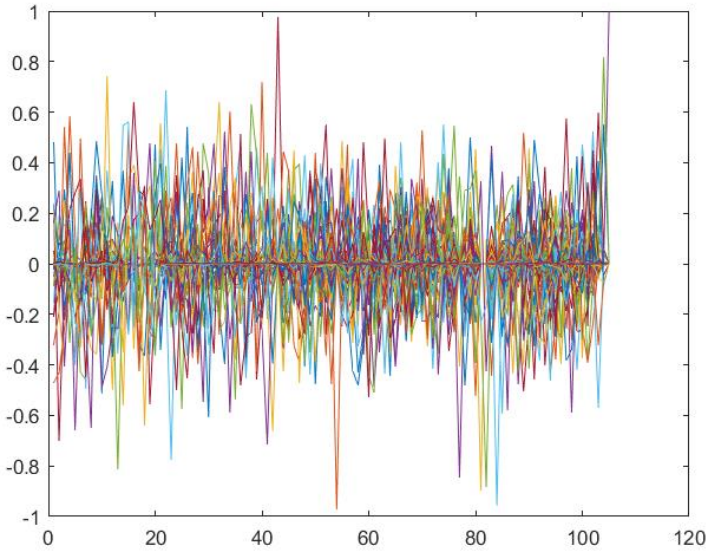
   [w,v]= eig(Cov matrix)
   we compute the 10 most significant Eigen vectors and now the eigen vevtors are multiplied with covariance matrix to obtain 10 most significant Principal Components.

Subtask 2: Execution of PCA

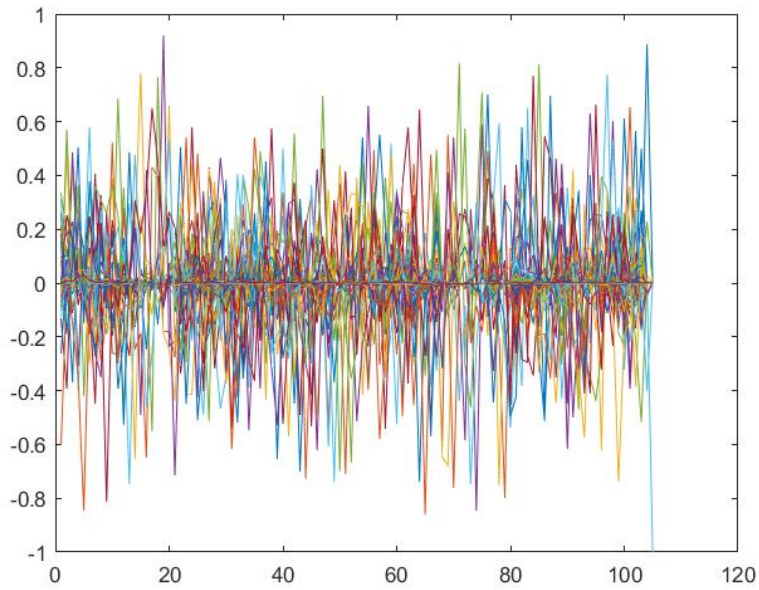
Use Matlab's PCA function to run PCA on your feature matrix. Show all the eigen vectors in a plot.

Eigen vector plotting are as follows:

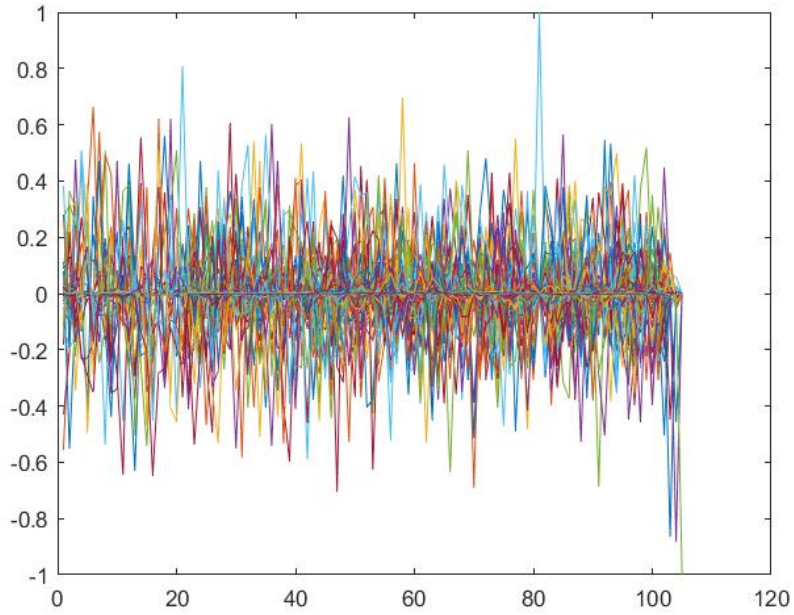
Cook:


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
20
40
60
80
100
120

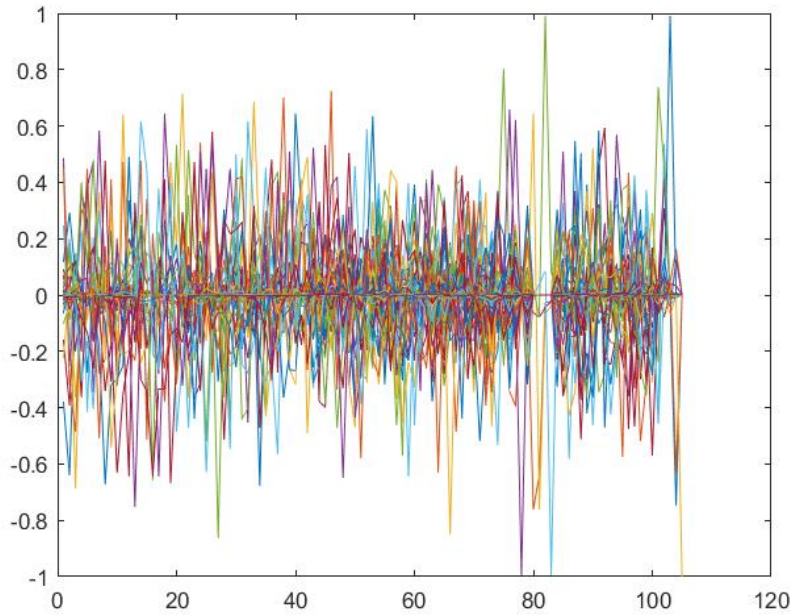
Drive:


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0
20
40
60
80
100
120

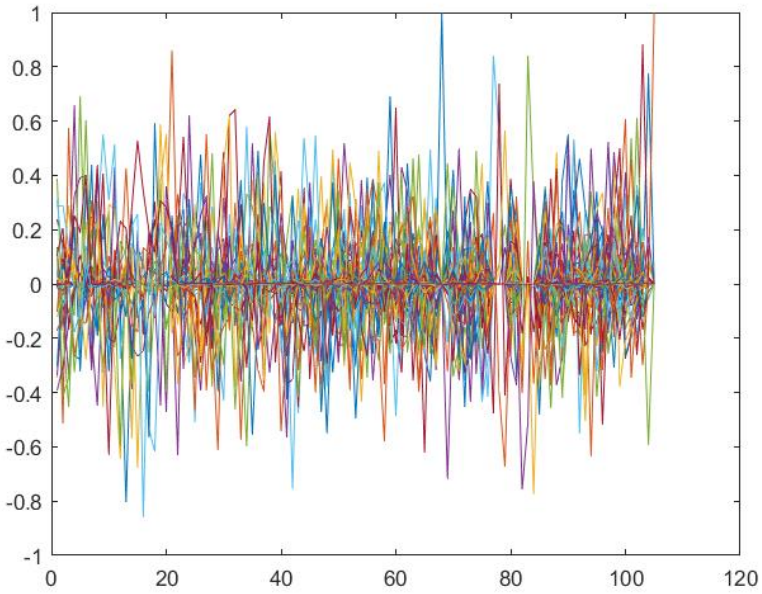
Eat



Eatfood



Keyboard



Subtask 3: Make sense of the PCA eigen vectors

Write an explanation on the reason why the eigen vectors turned out the way they did.

1. Latent matrix obtained contains the eigenvalues of the covariance matrix. The values in the latent will be in decreasing order of variance. Similarly, The columns in Coefficient are in order of decreasing component variance. Since the first 10 columns are prominent we consider coefficient matrix with dimensions 105*1.

Subtask 4:

Results Of PCA:

The FinalFeatureMatrix (105*1) is multiplied with the reduced coefficient matrix obtained in step 4 which gives us a NewFinalFeatureMatrix of dimensions (105*1),where the 10 columns are the prominent features obtained by comparing the Eigenvalues.

Th initial input matrices from which the FinalFeatureMatrix was obtained are as follows:

Drive

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	138.8716	353.4932	289.9268	1299.483	1540.723	491.8476	162.8765	127.0669	359.6997	212.3505	411.7148	137.6616	901.4859	14109.4	8921.447	17498.28	35446.86	21610.19	6428.792	12462.4	9790.015	6.463401	16.47859
2	75.05721	192.3586	157.6909	727.5378	750.6227	251.3948	88.64525	69.27093	196.1809	115.634	224.195	74.95757	491.7912	6500.777	4377.236	9042.654	19261.34	10221.54	3070.965	5865.057	4642.583	4.763209	12.21047

Eat

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
	3959.865	11804.93	5972.187	323284.7	434419.3	521847.8	6193.634	4446.467	7112.008	8372.451	16394.64	4131.802	19389.59	211671.9	300105.1	980778.3	421503.3	400994.9	189569	291092.3	390850	34.11122	102.0569
	168.9644	6.429841	11.41525	326.0232	147.8862	111.9748	35.32343	109.6508	33.66561	118.9715	79.29792	251.8943	19.87234	1365.604	1407.476	26923.97	1983.408	2548.873	6091.005	2511.554	1312.965	13.26337	0.503512
	414.2391	141.2136	359.5784	3535.015	10733.89	4898.492	78.6426	248.1402	246.8461	440.3813	298.0174	498.8675	786.6267	13881.99	8101.747	26579.32	7747.462	10805.39	8616.592	7494.769	31753.28	17.4515	5.953018

Guitar

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
6600.963	6525.454	21851.81	847984	704882.9	1309263	2319.325	3631.031	7864.769	21050.06	7495.843	6165.863	18162.66	1040734	1809854	2327775	1448648	987355.4	1013720	933218.3	902378.7	42.246	42.62949

Subtask 5:

Argue whether doing PCA was helpful or not. May be compare the plots generated from subtask d of task 2 and subtask 4 of Task 3.

PCA was very helpful in reducing the dimensions of dataset. With such a huge dataset containing many dimensions, PCA helped us to find eigen vector matrix which in turn helped us to find the prominent features among the entire dataset.