CSE 572: Data Mining

Ashwin Jose Poruthukaran - 1215204692
Athira Nambiath Asokan -1213202560
Avinash Anilkumar Pai - 1215043193
Harish Anand - 1215098885

February 6, 2019

Abstract

Over the years, technological advancements led data analysis from a manual and time consuming process to automated and easy process. The potential of unearthing relevant patterns is susceptible to the complexity of the data collected.
This assignment aims at building a computing system that can understand human activities. We will be identifying the eating activities amongst a mix of eating and non eating activities. Through the assignment we will be comparing the features selected through various extraction methods with the feature selection using Principal Component Analysis.

Keywords
PCA, FFT, RMS, Data Mining

1 Introduction

Our aim is to identify eating activities from non-eating activities. Data will be collected from Myo armband. Data from IMU and EMG sensors is collected along with the ground truth in the form of frame numbers where an eating action starts and ends. The assignment is divided into 3 phases. Phase 1 consists of data cleaning and organization, phase 2 consists of data extraction, and phase 3 for feature selection.

1.1 Terminology

- *PCA* : Principal Component Analysis is an orthogonal transformation which converts a set of correlated variables to a set of uncorrelated variables.
- *FFT* : A fast Fourier transform is an algorithm that computes the discrete Fourier transform of a sequence, or its inverse.
- *RMS* : The root mean square is defined as the square root of the arithmetic mean of the squares of a set of numbers.

1.2 Assumptions

- For the sake of this assignment we have only considered fork data for all users.
- For the sake of simplicity the final data matrix has been populated with individual rows for each user activity type rather than for instances of each activity.
- While exploring feature extraction methods only 4 users were considered to identify patterns.

2 Proposed Solution

2.1 Phase 1: Data cleaning and organization

2.1.1 Dataset

The data is collected from two sources: A) wristband data: with i) accelerometer, ii) gyroscope, iii) orientation, and iv) EMG sensors. The sampling rate is 50 Hz for all the sensors. B) Video recording of the person performing eating actions used for establishing ground truth. Ground truth is in the form of frame numbers with starting and ending of eating actions. The video data is taken at 30 frames per second.

2.1.2 Identifying eating and non-eating activities

The groundTruth file contains frame number pairs corresponding to each eating activity. The last frame pair is compared with the timestamp in the EMG and IMU files for the respective user and the duration of each activity upwards are synchronised accordingly. An extra column is added to the dataset to classify eating activities as 1 and non-eating activities as 0.

Once all data has been classified as eating and non eating activity, one eating and one non-eating activity range is chosen for analysis of features for feature extraction.

The organised dataset now consists of both IMU and EMG Sensor Data for 1 eating and 1 non-eating activity each with its corresponding class label added as an extra column. The values for each column were further scaled down to the range of -1 to 1 so as to normalize the data.

2.2 Phase 2: Feature Extraction

For the purpose of feature extraction, 5 different extraction methods were chosen.

- *Mean:* Average of sampled values from sensor.
- *Variance:* Variance of sampled values from sensor.
- *RMS:* Root mean square value of sampled values from sensor.
- *Entropy:* A statistical measure of randomness of sampled values from sensor.
- *Fast Fourier Transform:* Fast Fourier transform is performed on the sampled sensor values and the power spectral density is calculated with respect to each frequency.

Each of the above feature extraction methods were applied on all 18 sensors one by one and for 4 different users. The trends and patterns between users for each method was studied. One out of 5 above-mentioned methods were chosen for each of the 18 sensors based on the best split among features for eating and non-eating activities.

2.2.1 Mean

The means of eating and non-eating activities were compared with each other for each one of the sensors. The comparisons were performed for 4 different users and mean was chosen as the feature extraction method for the IMU6(Accelerometer Y) sensor. As observed in figures 1,2,3 and 4, the average mean of IMU6 - Accelerometer Y sensor values for eating activities is greater than that of non eating activities. This makes sense since moving the spoon or fork over the y axis is a part of eating activities and hence is bound to distinguishable from non-eating activities.

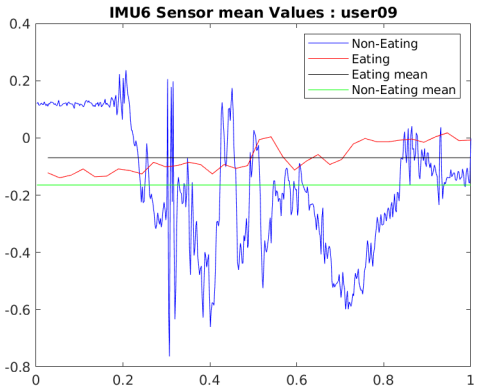

Figure 1: Accelerometer Y - User 09

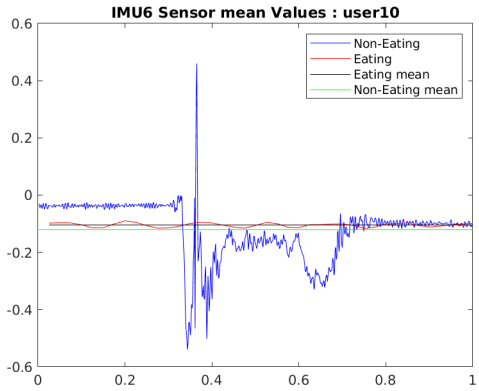

Figure 2: Accelerometer Y - User 10

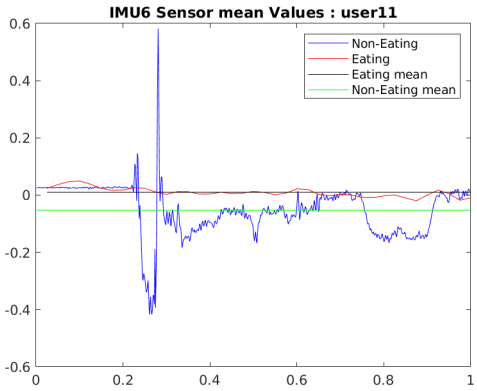

Figure 3: Accelerometer Y - User 11

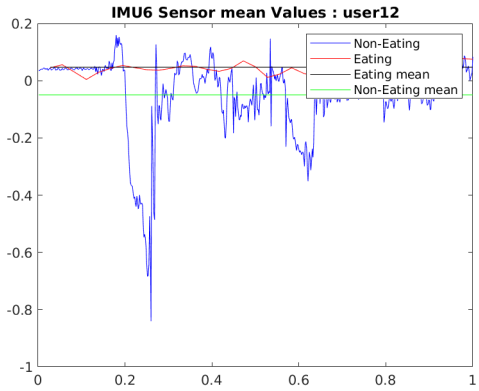

Figure 4: Accelerometer Y - User 12

2.2.2 RMS

The Root mean square values were compared for eating and non-eating activities. Figures 5,6,7 and 8 show the comparison of RMS values for 4 different users for IMU10 - Gyroscope Z sensor. It is observed that there is a significant difference in the RMS values for eating and non-eating activities irrespective of users.
RMS has been used for the extraction of features from the following sensors

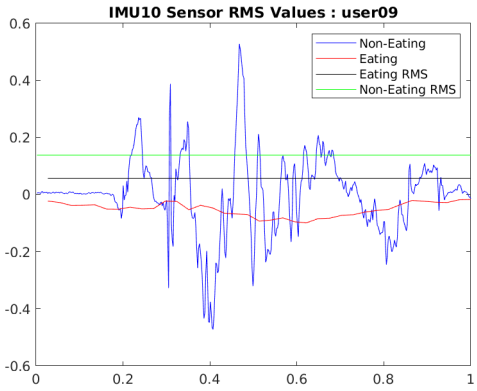

Figure 5: Gyroscope Z - User 09

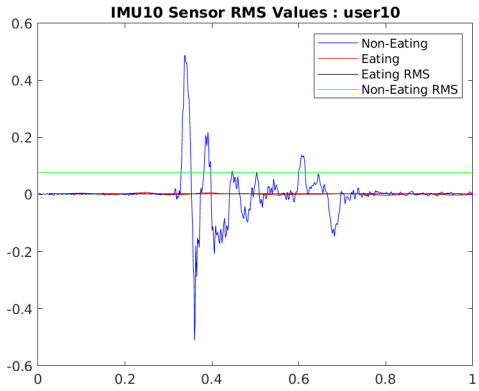

Figure 6: Gyroscope Z - User 10

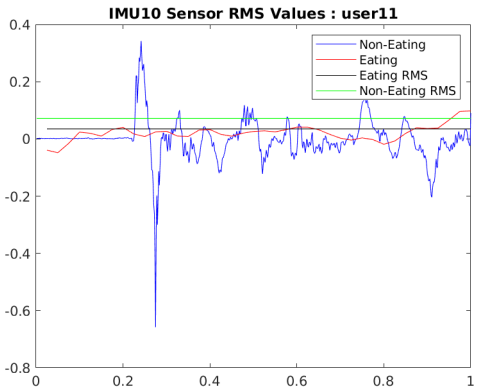

Figure 7: Gyroscope Z - User 11

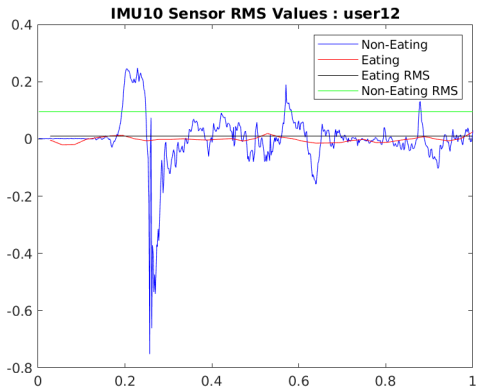

Figure 8: Gyroscope Z - User 12

- EMG3
- EMG4
- EMG5
- IMU9 - Gyroscope Y
- IMU10 - Gyroscope Z

2.2.3 Entropy

The Entropy values were compared for eating and non-eating activities. Figures 9,10,11 and 12 show the comparison of Entropy values for 4 different users for IMU5 - Accelerometer X sensor. It is observed that there is a significant difference in the entropy values for eating and non-eating activities irrespective of users. The entropy of non-eating activities seem to be higher than that of eating activities for Accelerometer X Values.
Entropy has been used for the extraction of features from the following sensors

- IMU5 - Accelerometer X
- IMU7 - Accelerometer Z

- IMU8 - Gyroscope X

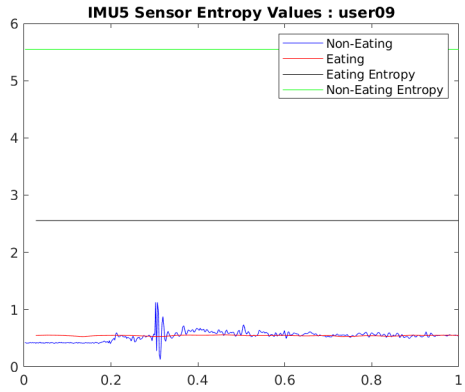

Figure 9: Accelerometer X - User 09

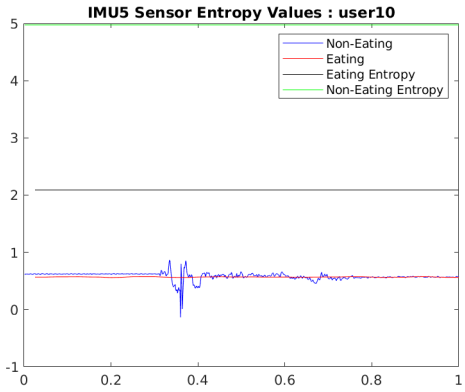

Figure 10: Accelerometer X - User 10

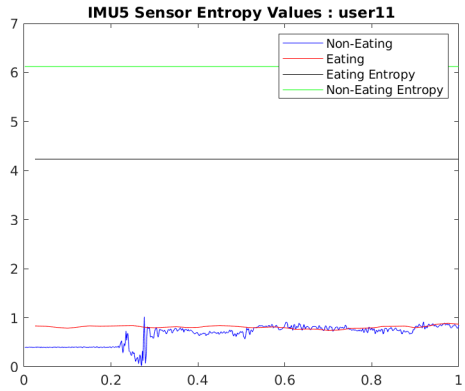

Figure 11: Accelerometer X - User 11

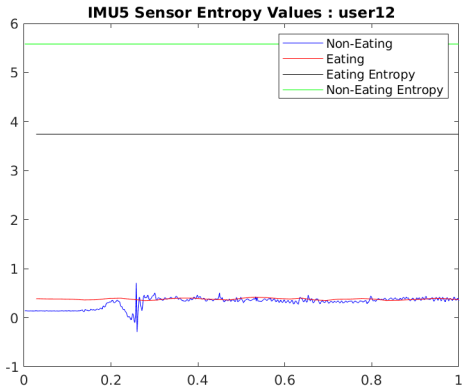

Figure 12: Accelerometer X - User 12

2.2.4 Variance

The variance values were compared for eating and non-eating activities. Figures 13,14,15 and 16 show the comparison of variance values for 4 different users for IMU3 - Orientation Z sensor. It is observed that there is a significant difference in the variance values for eating and non-eating activities irrespective of users. The variance of non-eating activities seem to be higher than that of eating activities for Orientation Z Values.
Variance has been used for the extraction of features from the following sensors

- IMU1 - Orientation X
- IMU2 - Orientation Y
- IMU3 - Orientation Z
- IMU4 - Orientation W

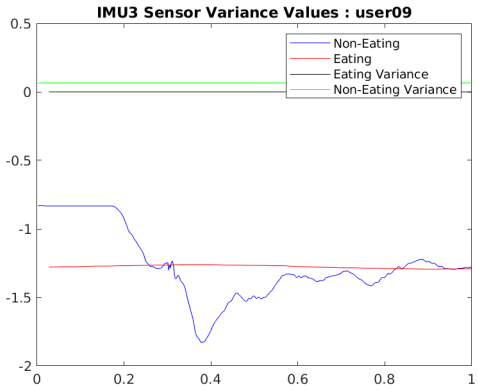

Figure 13: Accelerometer X - User 09

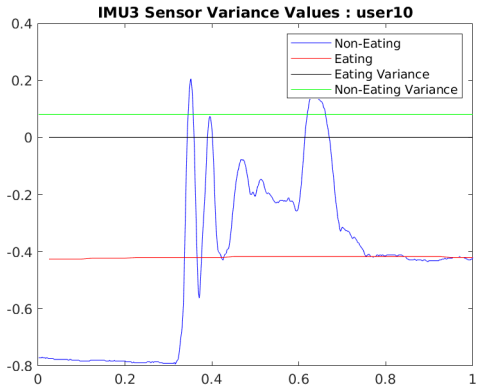

Figure 14: Accelerometer X - User 10

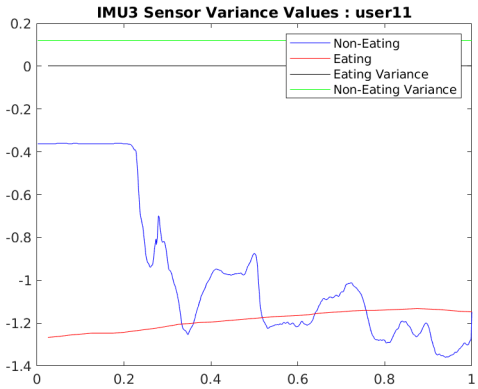

Figure 15: Accelerometer X - User 11

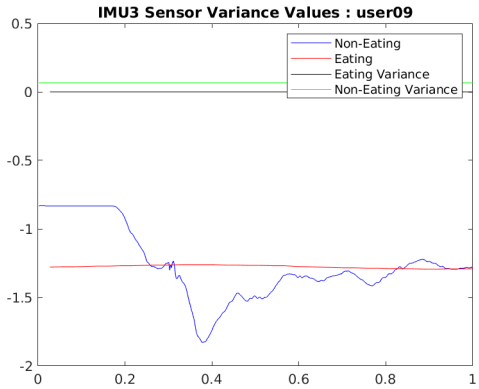

Figure 16: Accelerometer X - User 12

2.2.5 Fast Fourier Transform

The sensor data values were transformed using fast fourier transform and and power spectral density was calculated based on the complex conjugate part of the signals. Figures 17,18,19,20 show how the power spectral density varies with respect to frequency for eating and non-eating activities for each user. The figures show result of the transformation on EMG1 sensor values. It can be observed that the peaks at frequency = 8Hz for eating is larger than non-eating activity.
This transformation has been used for the extraction of features from the following sensors

- EMG1
- EMG2
- EMG6
- EMG7
- EMG8

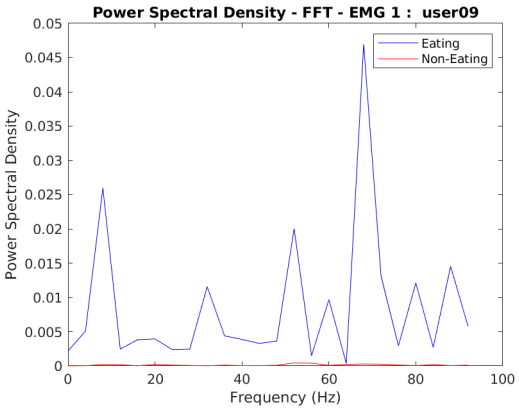

Figure 17: EMG1 - User 09

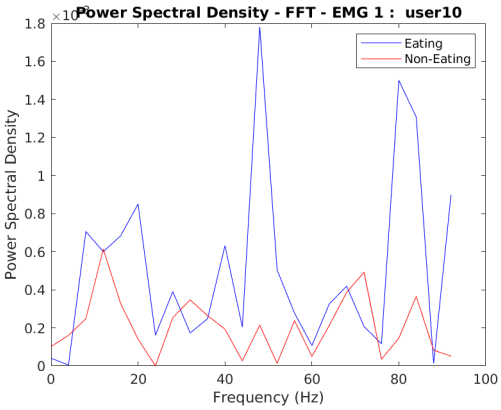

Figure 18: EMG1 - User 10

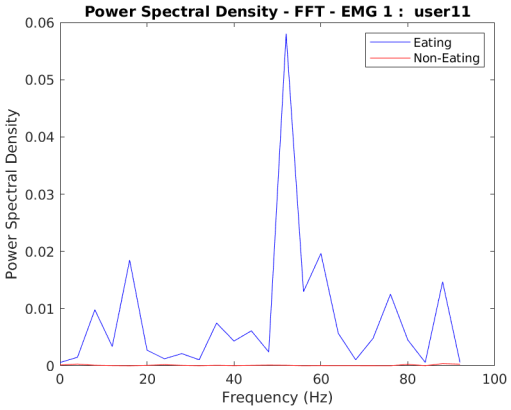

Figure 19: EMG1 - User 11

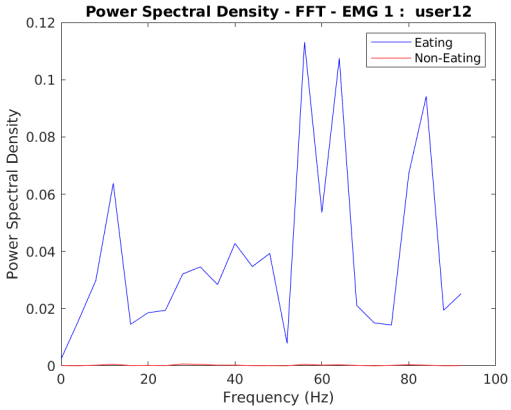

Figure 20: EMG1 - User 12

2.3 Phase 3: Feature Selection

2.3.1 Subtask 1 : Arranging the feature matrix

In order to apply Principal Component Analysis and perform dimensionality reduction, the data-set needed to arranged in a specific format. After the initially obtained senor data was scaled down, each of the sensor values were transformed using the chosen feature extraction methods to extract the best features required for classification. Once the feature extraction transformations were performed, the data was then collated into a 60x18 matrix. Each of the rows represent the eating/non-eating activity corresponding to the user. For ease of calculation all eating activities of one user is combined into a single row. Similarly all non-eating activities of the user are also combined into a single row. Hence, for all 30 users a total of 60 rows are added to the final matrix. Each of the columns of the matrix represent the data from each of the sensors.

2.3.2 Subtask 2: Execution of PCA

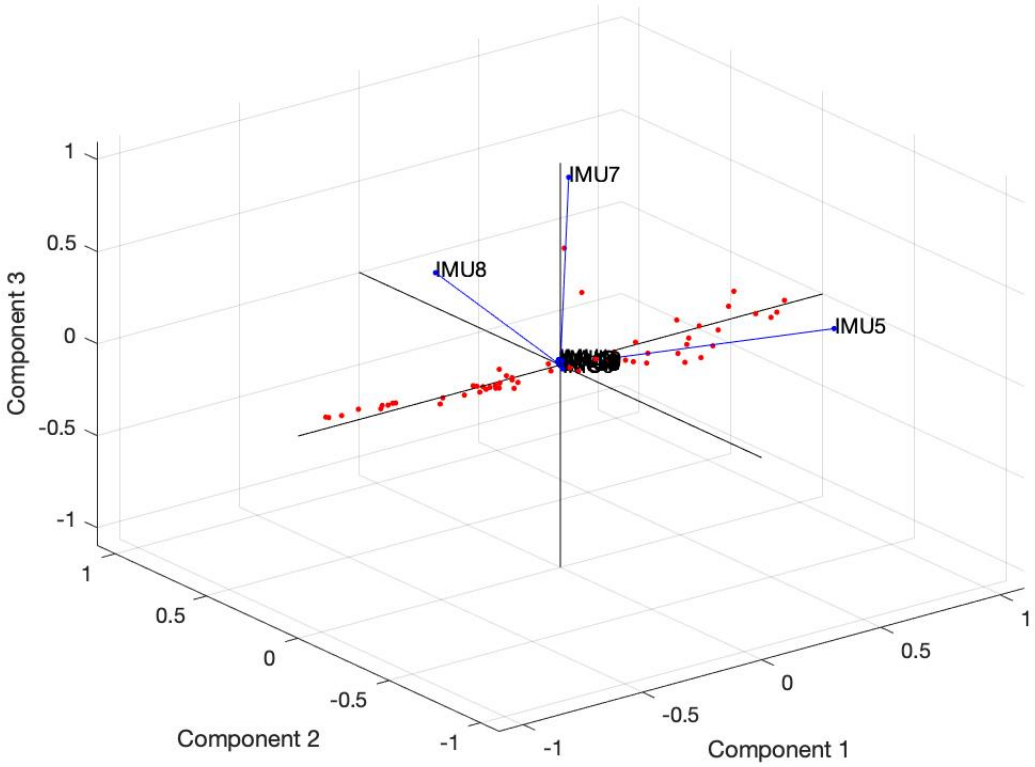

Figure 21: Eigen Vectors with all features labelled

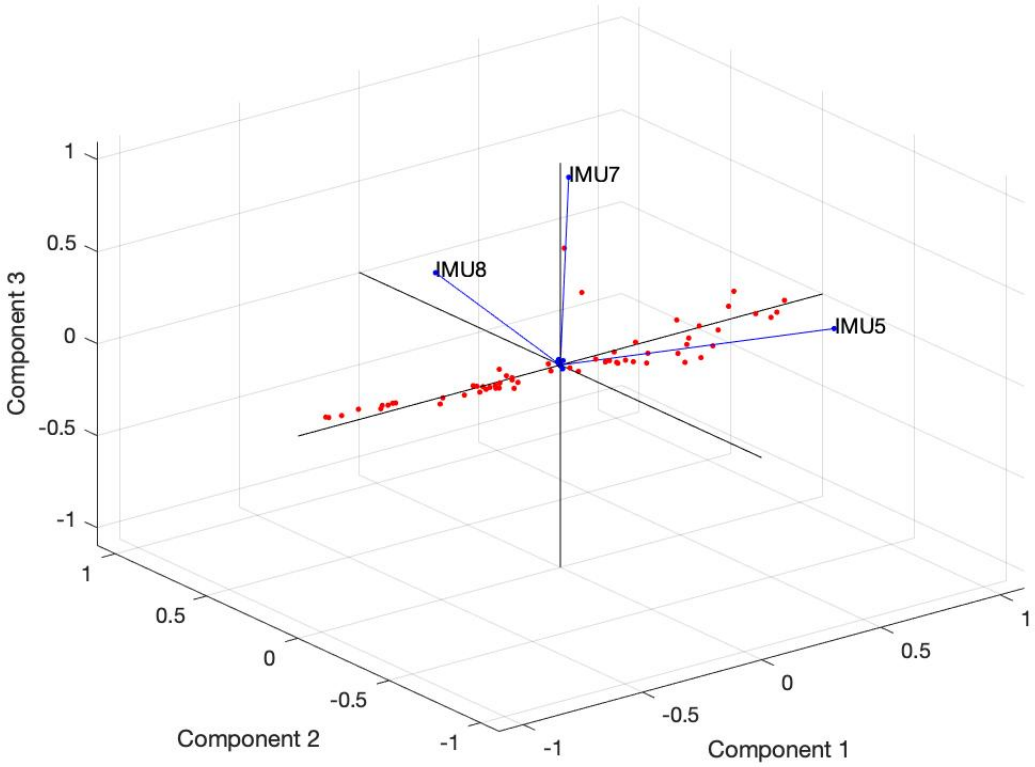

Figure 22: Eigen Vectors of IMU7, IMU8, and IMU5 are shown

2.3.3 Subtask 3: Eigen Vectors

The following are the 3 Eigen Vectors corresponding to the largest 3 Eigen Values. Upon close observation it can be seent that the largest values in the 1st Eigen Vector are the 13th, 16th and 15th values. These correspond to the values of the sensors IMU5, IMU7 and IMU8 repectively. This also implies that these sensors contribute to a large portion of the overall variance of the dataset.

Sl.	Eigen Vector 1	Eigen Vector 2	Eigen Vector 3
1.	1.09919373744800e-05	0.000150852145137170	-0.000161320327103266
2.	2.82111772413884e-05	0.000386240461666798	-0.000432608283681885
3.	0.00412302488048477	0.0135083698911024	-0.00453249090557029
4.	0.00277010110459693	0.00339876833913501	-0.00391672370055118
5.	0.00219299475891747	0.00250950024710709	-0.00632987846083100
6.	4.42607062042201e-07	6.31125620342717e-06	-1.79680433273039e-06
7.	1.41150430708721e-05	0.000220547429176925	0.000324412743596315
8.	9.37504365855631e-07	1.30502937078656e-05	-1.01267541557749e-05
9.	0.0122174401181973	0.0254553805273053	0.0139307148441998
10.	0.00352698848367091	0.0143334762558397	0.0103561970781069
11.	0.00712980229701775	-0.00609786964491388	0.0242716860693414
12.	0.0131721584705908	0.00773169492478378	-0.0222696227782088
13.	0.973229931009042	-0.225472751798616	-0.0409198654861715
14.	-0.0164590539016570	-0.0336634412290162	0.000258151590291385
15.	0.0449840155739089	0.0129382258430978	0.998153344585269
16.	0.223764122331930	0.972632625500517	-0.0229603692073501
17.	0.00466526289265603	0.0147334386856659	-0.000197584826040311
18.	0.00462208687968886	0.0218901965567936	0.00505415720962421

2.3.4 Subtask 4: PCA Results

Since the PCA function in MATLAB takes the complete feature matrix as input, the above mentioned matrix obtained from subtask 1 is passed to the function. The result obtained contains the principal components and their contributions to overall variance. It is observed that first 3 principal components correspond to 99% of the overall variance. The variance contributions are as follows:

Sensor	Overall Variance%
IMU5 - Accelerometer X	91.99%
IMU8 - Gyroscope X	7.34%
IMU7 - Accelerometer Y	0.63%
Total	**99.96%**

2.3.5 Subtask 5: Helpfulness of PCA

From the above table it is clear that the most deciding features in this feature set were 3 sensors(IMU5, IMU8, IMU7). These 3 sensors' data was transformed using the entropy function before applying PCA and seemed to show a clear distinction for eating and non-eating activities even before PCA as shown in figures 9-12 . This only goes on to prove that the selection of features was accurate.

References

[1] Matlab Documentation. www.mathworks.com/help/matlab/