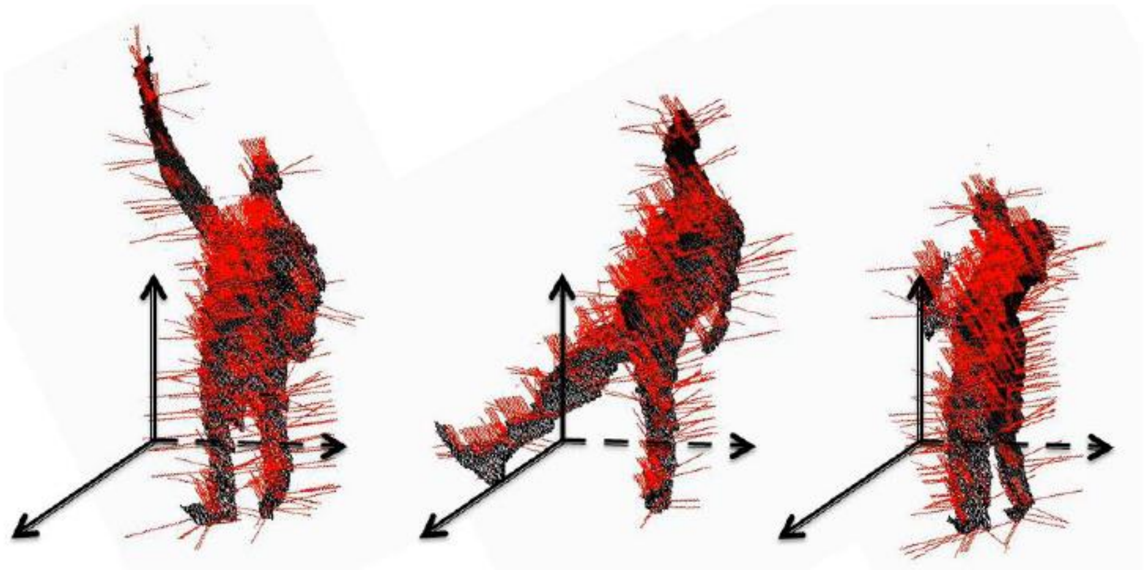
Human Activity Recognition



Mikit Kanakia

Introduction & Problem Statement:

Two years back in 2014, the global wearable market device market size was valued at USD 3.9 billion. With advancement in Internet of Things (IoT) and groundbreaking research in the field of wearable devices, this rate is expected to grow at 28% over the forecast period [1]. As a result, the implications in future can be comprehended to be tremendous. US citizens are becoming more aware about health related issues and bringing about a change in their lifestyle. According to a report by PWC, 1 in 5 Americans owns a wearable and 1 in 10 of them wears them daily. By 2020, it is projected that a total of 285.3 million units will be active [2].

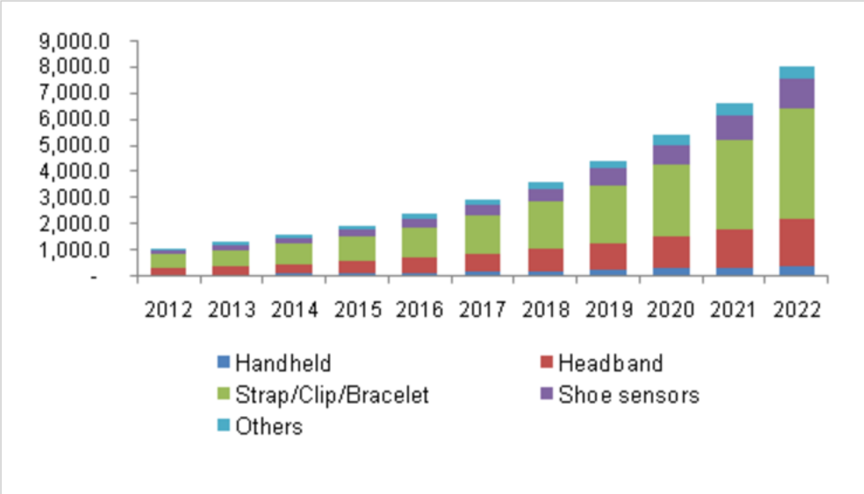


Fig. 1 North America wearable devices market share, by site, 2012-2022 (USD Million)

Big giants like Nike, FitBit, Apple, Google, Under Armour etc. are entering the market with IoT based wearables and intuitive mobile apps to monitor the health activities of individuals. Users are sharing data with insurance companies to reduce premiums or with doctors to help treatments. Additionally, Human Activity Recognition (HAR) has emerged as a key research area in the past few years and is gaining increasing attention by the pervasive computing research community for the development of context-aware systems. Apart from common applications in supporting weight loss programs, HAR can find implications in many important areas like elderly monitoring, life log systems for monitoring energy expenditure and digital assistants for weight lifting exercises [3].

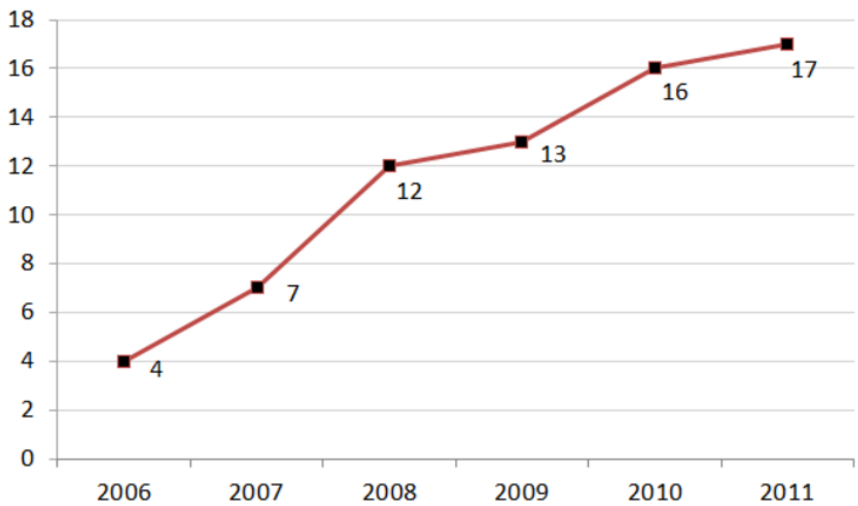


Fig. 2 HAR: IEEE publications (2006-2011) based on wearable accelerometers' data

Taking into consideration the importance of this issue, we aim at providing solutions to construct a wearable device that can accurately perform specific goals as mentioned below:

1. To accurately detect the activity of an individual as sitting, sitting-down, standing, standing-up and walking from the given accelerometer data.
2. With the help of additional features that exploit the temporal nature of the data, reliably detect the current activity and also find out the time taken to switch to another activity.
3. Fabricate additional attributes/parameters to improve the change-point detection.
4. For a single accelerometer, determine which location results in the best classifier in terms of activity and change-point detection.
5. With the help of additional research and/or surveys, find out the desirable position that an individual prefers while wearing a heath device (ergonomics of the device for design comfort).

The Data Set & Data Cleaning:

The dataset comprised of 5 classes (sitting-down, standing-up, standing, walking, and sitting) collected on 8 hours of activities of 4 healthy subjects. Initially it contained 3 columns and 165634 rows. The columns contained shifted data and inappropriate attribute values due to this. The data was cleaned and the new dataset eventually comprised of 19 attributes and the same number of rows. The attributes such as height and BMI were mathematically corrected as it had missing decimal values. Box plots were plotted to analyze the distribution of the variables. It was observed that there was some bad data appended with timestamps that were appearing as outliers. Such data was eliminated and the clean data set was ready for further analysis.

Risks & Contingencies Implemented To Mitigate Risks:

1. **Insufficient data**
   Data might prove to be less representative as only four individuals are monitored in the period of eight hours. In addition, the four people's behaviour or habits may cause some bias. To mitigate this risk, attributes without the human characteristics like height, weight, age and BMI were analyzed.
2. **Outliers**
   Outliers in the given attributes were identified and eliminated for a better distribution of data.
3. **Missing values**
   There were no missing values in the provided data set.
4. **Instantaneous nature of data**
   The data points in the data set are instant, however the temporal information between several data points are to be taken into account. To mitigate this risk, a segmentation analysis by aggregating the data points to a sliding window which can provide some information about the temporal nature was carried out.
5. **High Amount to Dimensionality ratio**
   Since there is a high A/D ratio in the data set, a complex model was trained and additional features were fabricated to take care of the data bias.
6. **Risk of high computation costs**
   Since dealing with a large amount of data is accompanied with higher computation time and costs, we plan to propose models which are computationally less expensive but at the same time also do not compromise on the classification accuracy.

Exploratory Analysis:

An exploratory analysis was carried out on the cleaned dataset. The class distribution is as shown in the figure below. There is more data for classes - sitting, standing and walking. The transition classes - sitting down and standing up have less data points as can be understood since the transition from one activity to another are not contained for a longer period in a given time span as compared to the other human activities.

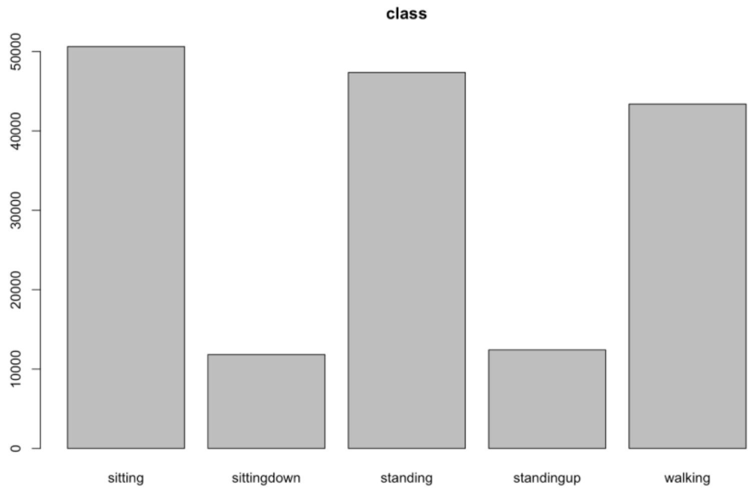
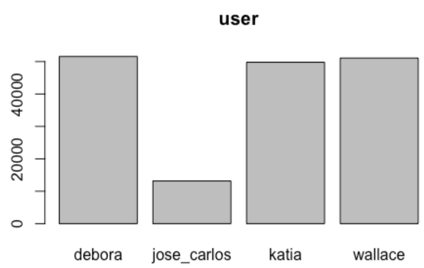


Fig. 3 Distribution of classes

The user data collected from the 4 devices is as shown below. There were more data points for women as compared to men and Jose Carlos had the least entries which was understood as his age was 75 years.



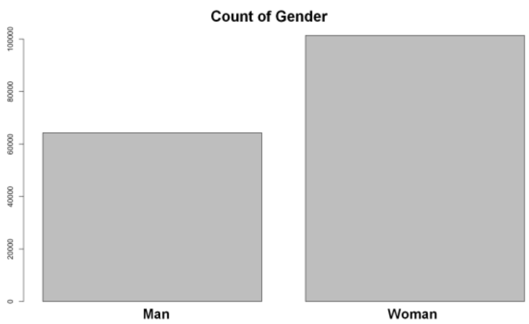


Fig. 4 User Data Distribution

Fig. 5 Gender Distribution

The time series plots for each accelerometer were plotted to understand the variations in human activities as shown in figure 6. For representation purposes, the time series of first accelerometer data is shown. There is a lot of variation in the axes when the users are walking. There is very less variation when the users are standing or sitting as there is a momentary stoppage of motion. Sitting-down and standing-up capture short bursts of variations. All these are in line with the human intuition of activity trends.

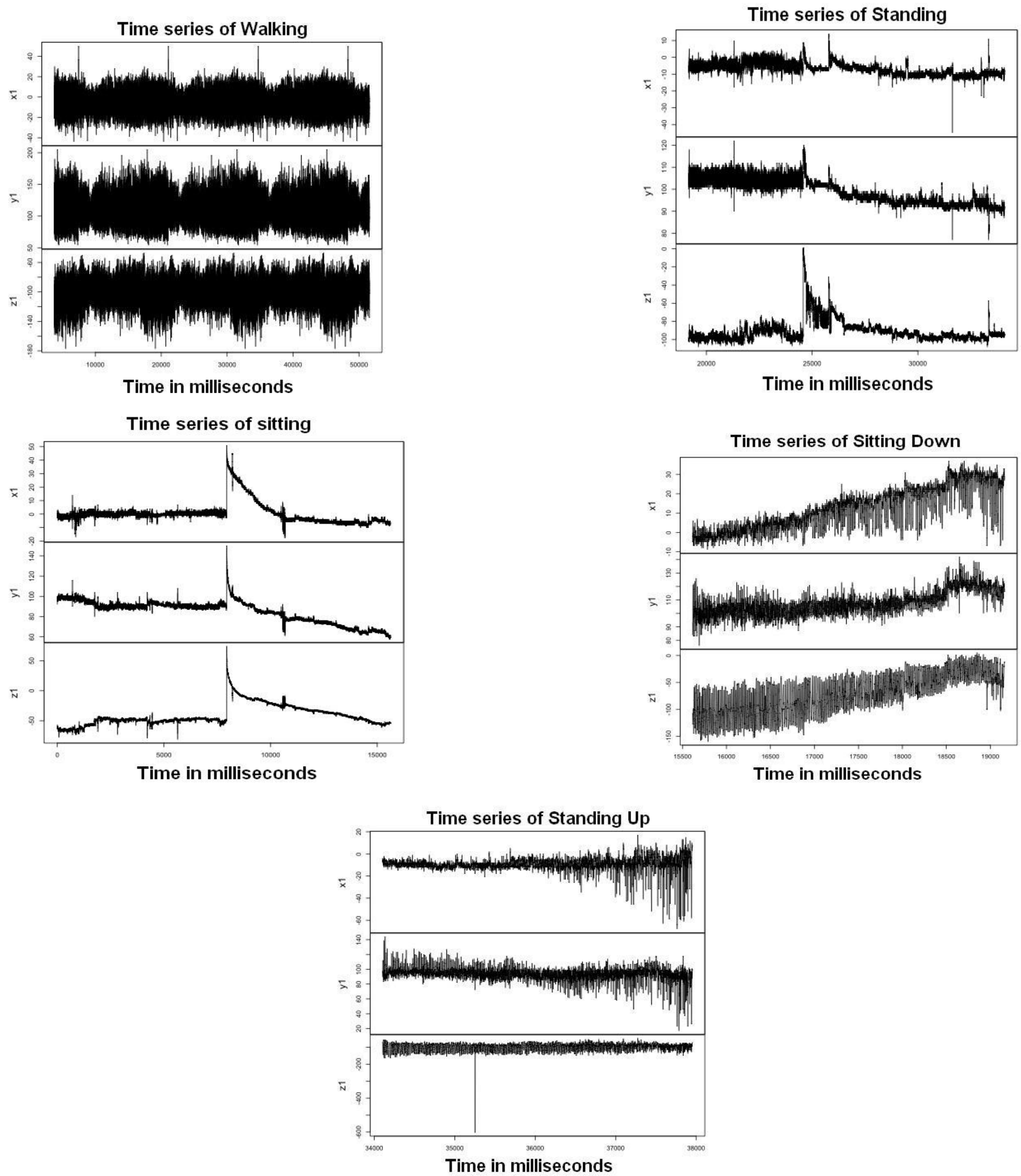


Fig. 6 Time series plots for 1st accelerometer data

Initial Classification on Raw Data Set:

1. Since there are 5 target classes (multinomial distribution), the Naive Bayes and K - Nearest Neighbours classifiers were tested on the raw dataset to predict human activity. The performance was compared to the default classifier.
2. The dataset was partitioned into train and test by training it on 3 users and testing it on 1 user. It was ensured that the dataset was not shuffled and sampled and the future data was not used to predict the past since it is based on time series.
3. In order to find the optimal K-value, values from 1 to 20 were tested and it was observed that at K = 17, the data was avoided being underfit and overfit. This was the optimal K-value and it since the the query point computes distances to all data points, the process to find the optimal

K was computationally (expensive) time consuming (approximately 4 hours) but the trade off was a better accuracy.

4. The random forest classifier was implemented and figure below shows the model performances based on accuracy.

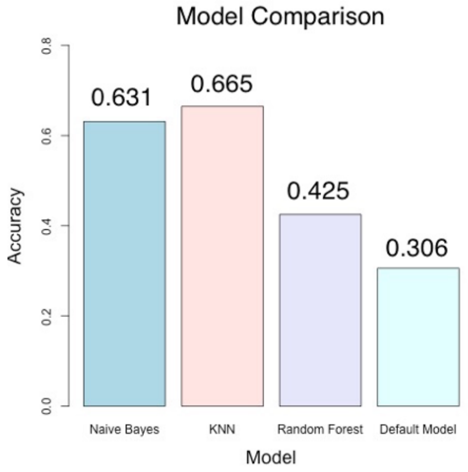


Fig. 7 Model comparisons on raw data based on accuracy

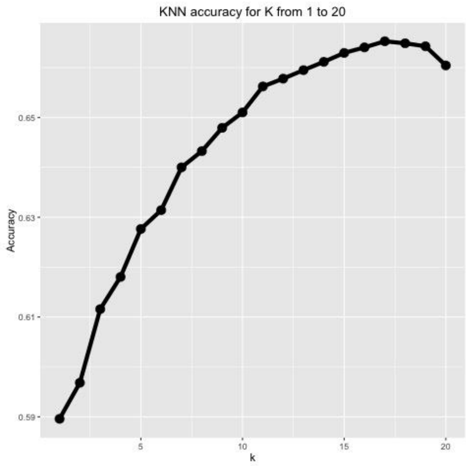


Fig. 8 Selecting optimal K-value

pred \ class	sitting	sittingdown	standing	standingup	walking
sitting	33793	94	0	634	0
sittingdown	6814	1940	974	3954	2118
standing	0	1200	36258	1162	5857
standingup	6670	6684	1306	4155	1414
walking	3354	1909	8832	2510	34001

Fig. 9 Confusion matrix for K-NN model

5. The K-NN model was tested on the dataset with and independent of the user features like height, weight, age, BMI etc. and it was observed that the user independent dataset gave slightly more accuracy and future analysis was carried out on this dataset.

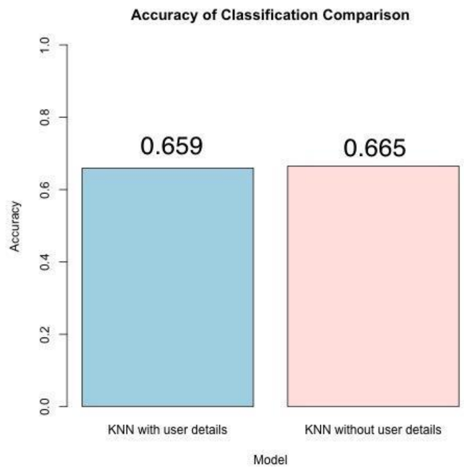


Fig. 10 K-NN comparisons on user dependent and user independent data

Data Preparation:

1. In order to capture the 3-Dimensional motion, three variables Pitch, Normal and Roll were fabricated to capture rotations around x, y and z axis for the user independent data set.
2. The data was rolled up to capture the temporal nature of the time series data. Four data sets were prepared:
   a. Derived mean and standard deviation for all the variables using six consecutive rows with 5 rows overlapping window.

b. Derived mean and standard deviation for all the variables using twelve consecutive rows with 11 rows overlapping window.
c. Derived mean and standard deviation for all the variables using 6 consecutive rows without any overlapping window.
d. Derived mean and standard deviation for all the variables using 12 consecutive rows without any overlapping window.

The attributes were now increased from 19 (original dataset) to 56. Since the initial dataset had a high A/D ratio with 165634 rows and 19 attributes, a complex model is now prepared with strong predictors to take care of the data bias.

Principal Component Analysis (PCA):

Since the input variables are all numeric (x, y, z axes and fabricated parameters like pitch, roll and normal), we selected PCA for dimensionality reduction to save computational time during classification (decrease the cost of learning). This will also help in finding efficient combinations of the x, y & z axes with their pitch, roll & normal rather than a greedy search approach like stepwise forward/backward elimination which give sub-optimal results. Before PCA, the dataset comprised of 56 attributes.

1. First the attributes were tested for near zero variance to make sure that there is some variance reflected by each of the attributes.
2. PCA for dimensionality reduction was carried out on all the 4 datasets mentioned above.
3. The top N components that reflected 80% variance of the entire dataset were retained and the new datasets were prepared for classification.
4. Based on the result of PCA and rolled up data, the non-overlapping dataset with a sliding window of size 6 achieved the best performance.

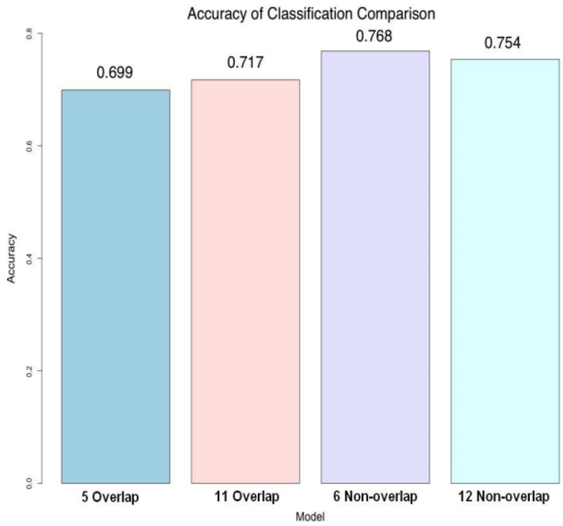


Fig. 11 Post PCA Accuracy for overlapping and non-overlapping data sets

```
               har.class
pred            sitting sittingdown standing standingup walking
  sitting          7096          24       22          6       0
  sittingdown       113         215      254        748     489
  standing          760         208     7319        190      49
  standingup        452        1406      239        554     663
  walking            15         116       59        571    6021
[1] 0.7686034
```

Fig. 12 Confusion matrix for 6 Non-overlap model

It is observed that the non-overlapping sliding window of size 6 increased the accuracy by 10% to ~77%. PCA helped in dimensionality reduction by reducing the dataset to 8 principal components which captured 80% variance of the entire original dataset. This helped in reducing computational time during classification. Additionally, non-overlapping sliding windows for 18 and 33 window size were created to see if a higher time period could help in better detecting patterns in human activities but the accuracy reduced to 75%. Hence, the non-overlapping sliding window of size 6 was the optimal window for carrying out classification.

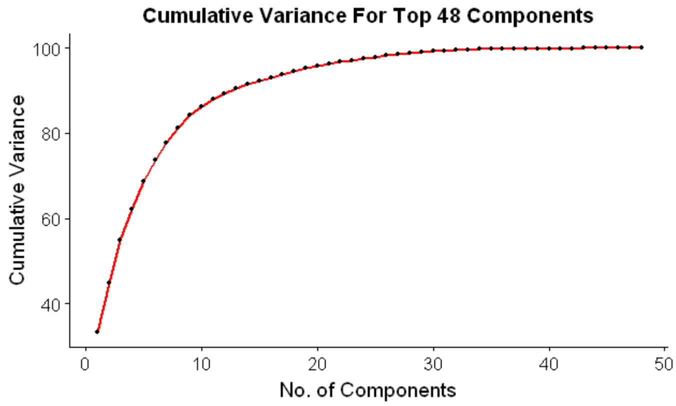


Fig. 13 Cumulative Variance Plot for Principal Components

It can be seen from the above plot that the top 8 components reflected 80% of variance for the dataset.

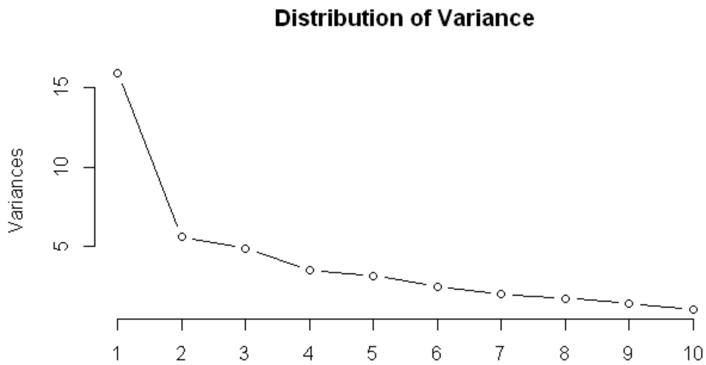


Fig. 14 Scree Plot for distribution of variance contained in subsequent principal components sorted by their eigenvalues.

Change Point Detection:

All the fabricated attributes (pitch, roll & normal for accelerometer) were tested to see which serves as a better change point detection parameter. The pitch and roll did not contribute to change point detection but the normal (distance) of the accelerometer helped significantly in identifying change points. The normal is a function of the square root of sum of the squares of all three axes (x, y and z). The 'ChangePoint' package in R was used to calculate the change points and the maximum change points were set to 20 in an 8 hour span. This was set after calculating the change points in data set which came to around 16. The change points method was set to 'Bin Segmentation' rather than 'PELT' which considers all data points as change points in order to avoid over-fitting. An AIC penalty function was also included to fight against over-fitting. As shown below, the graph for normal of first accelerometer gives a better view of change point as compared to other accelerometer positions. Although this accelerometer position gave the second highest accuracy (calculated later below), we would suggest to trade off on the accuracy (1.7%) and go ahead with the 'waist' as the best position for accurate change point detection and classification results.

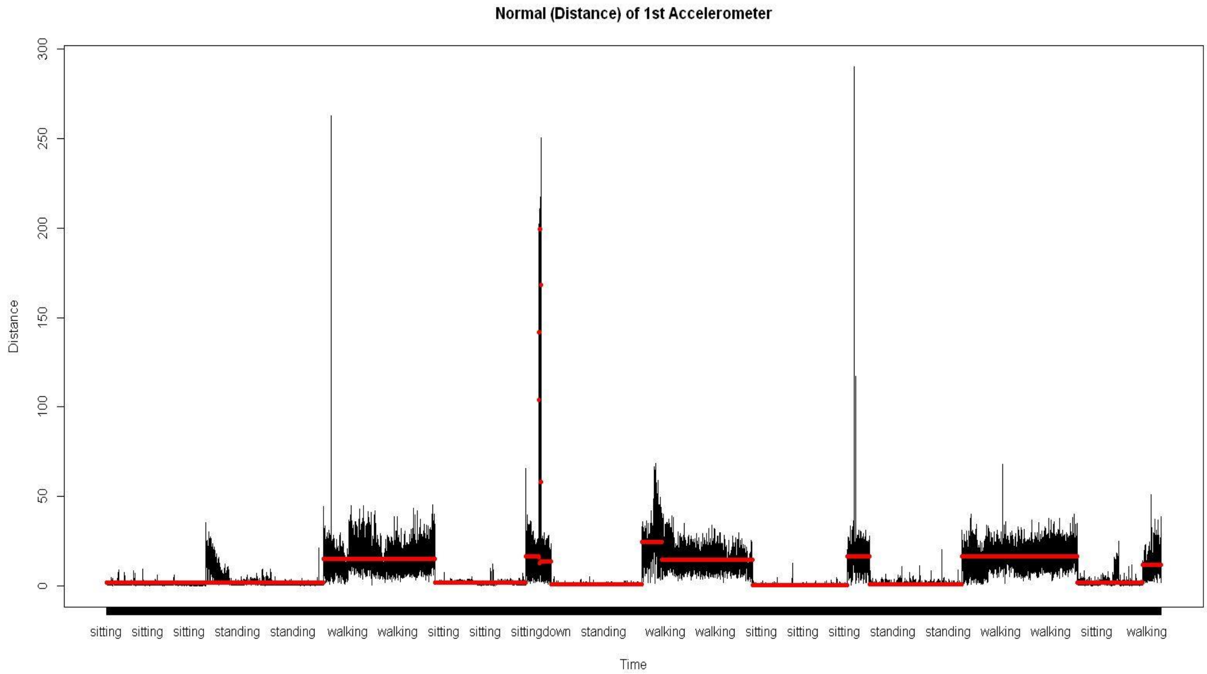


Fig. 15 Change Point Detection using the Normal (Distance) of Accelerometer.

If we just plotted the axes of the data without the new parameters, the change points were very inconclusive and did not provide any meaningful insights as shown below:

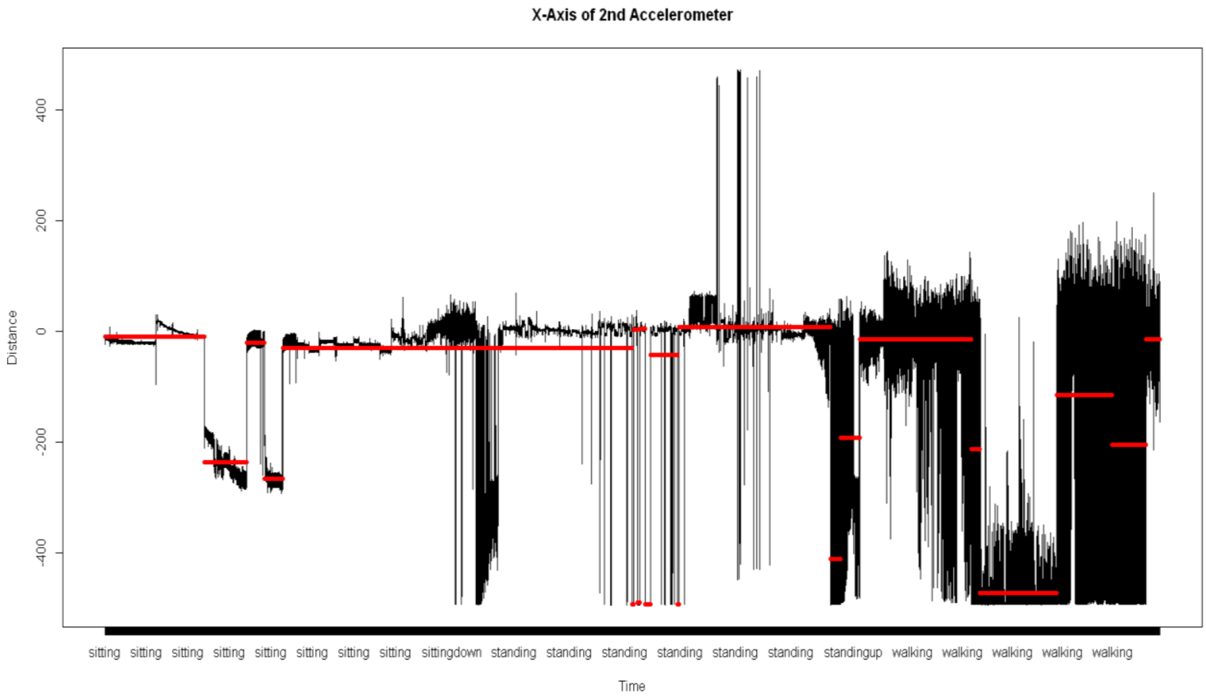


Fig. 16 Change Point Detection without additional features.

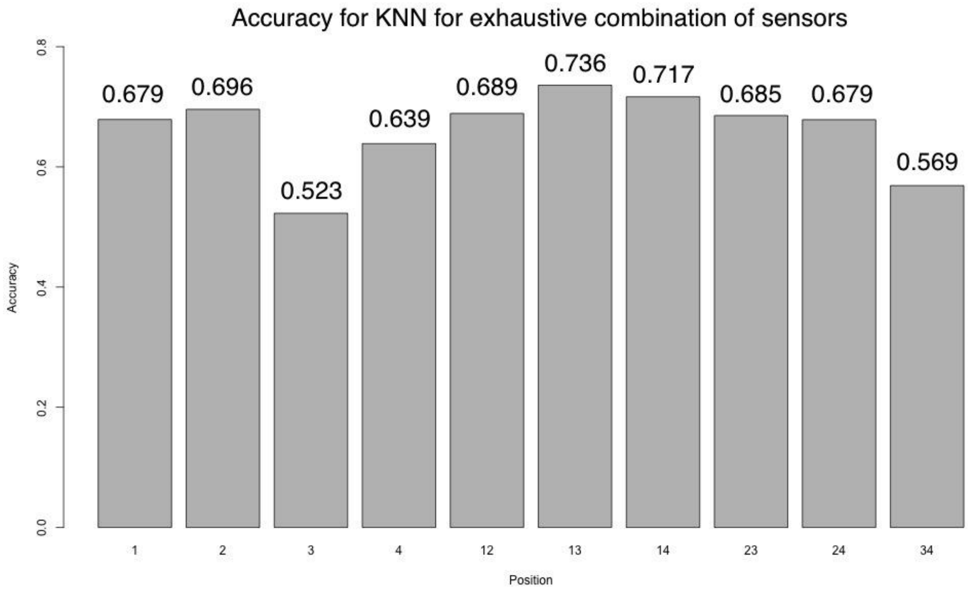
Latency and Accuracy Calculation for Change Point:

Out of 20 change points, for accelerometer 1, 11 change points were predicted with average latency of 1.8 seconds. Accelerometer 1 and 3 when worn in combination can detect 4 additional change points which were missed out by wearing just accelerometer on position 1. This is supported by our findings below which show that a combination of accelerometer 1 and 3 give the highest accuracy. Latency is calculated as difference in time interval between the change point predicted versus actual change point.

Accelerometer No	1	2	3	4
Predicted Change Point(out of 20)	11	8	7	9
Latency in Seconds	1.8	3.4875	0.642857	13.2

Best individual position to wear a device:

Based on classification accuracy, we can observe that the best position to wear a device is position 2 or the left thigh. Further, combination of wearing 2 devices was explored as well and it was observed that a combination of devices worn on position 1 and 3 i.e. waist and right arm gave us the best classification accuracy. But as mentioned above, after trading off on the accuracy for a better change point detection, we can conclude that accelerometer 1 or 'waist' position is desirable to be worn for better results.



Final Business Recommendations:
All the problem statement tasks were successfully dealt with.

Task 1: The activity of an individual can be correctly predicted from the given accelerometer data with 77% accuracy.

We recommend to collect data for more than just 4 users and for a longer span of time (couple of days) to better train a model and improve classification accuracy. Right now, the main classes like sitting, standing and walking are being classified better (as seen in the confusion matrix) but the transition classes need more data for training which can be improved by following this recommendation and collecting more transition activity over few days.

K-NN is computationally expensive but we have to keep in mind the sensitivity and future scope of the IoT + Health industry as projected above. With this in view, it is advisable to invest more on a computationally expensive and statistically better model for classification of human activities. We recommend to train and implement deep learning concepts to capture the transition phase precisely and accurately such as Hidden Markov Chain Model. We also recommend to use ensemble modeling techniques to improve the model accuracy.

Task 2: Additional features helped in exploiting the temporal nature of the data.
The roll, pitch and normal of the accelerometer helped in better explaining the temporal nature of data. In fact, the normal of accelerometer helped successfully in identifying change points for different human activities as shown above. We recommend to include 'normal' of accelerometer in future analysis for classification and change point pattern detection with more accuracy and less latency. This also gels with the fact that the Naive Bayes model performed less better as compared to the K-NN model. The dependencies between the x, y and z axes were accounted for and the normal distance of the accelerometer was useful in predictions.

Task 3: Change point was reliably detected.
As the change point is a 3 - dimensional action, additional features such as normal, pitch and roll were derived to improve the change point detection. The normal (distance) of accelerometer is the key attribute to capture these variations and should be used for future analysis. Successful detection of change point with the lowest latency of 1.8 seconds was achieved.

Task 4: Single best accelerometer should be worn on 'waist'.
For successful detection of change point with the lowest latency, use the waist (or 1st) accelerometer. As we know that the waist accelerometer will provide the best classification and change point detections, the company should focus on developing a wearable device which is comfortable to the user. The material and weight of the device should be kept in mind so that it does not prove to be troublesome during activity.
Additionally, the right arm (or 3rd) accelerometer can be combined for better and improved analysis of change point detection of all 5 human activities. It is intuitive that the waist accelerometer will capture standing, standing up, sitting and sitting down variations whereas the right arm can contribute for walking actions. Equip the future test subjects with these 2 devices for data gathering.

Additional Recommendations:

1. Provide the users with additional benefits like monthly detailed health reports with visual interfaces and intuitive results.
2. Partner with health insurance companies and hospitals by sharing the user data for a fee. This will help the user to tie up with the insurance providers for customized premiums or with doctors for better diagnosis of predicted health issues. Profit margins can be obtained on both of these scenarios.
3. Partner with sports companies like Adidas, Reebok, Puma etc. which have not entered the market space for IoT based health devices and develop wearables for them. A first mover advantage as a device manufacturer can be obtained with huge distribution margins.
4. Comfort is of utmost importance when it comes to taking care of customers. Since the most important devices are to be worn on the waist and right arm, go ahead with low cost but high resistant materials like 'polyolefins' which are water, sweat and chemical resistant and lightweight while constructing the wearable devices [4].

References:

Cover pic - http://crcv.ucf.edu/projects/HON4D/HON4D_4.png
[1] http://www.grandviewresearch.com/industry-analysis/wearable-medical-devices-market
[2]http://www.mobihealthnews.com/37543/pwc-1-in-5-americans-owns-a-wearable-1-in-10-wears-them-daily
[3] http://groupware.les.inf.puc-rio.br/har%20-%20ixzz2aUaBROdz
IEEE statistics for HAR publications - http://groupware.les.inf.puc-rio.br/har%20-%20ixzz2aUaBROdz
[4]http://www.plasticstoday.com/medical/impact-materials-selection-on-medical-wearables-technology/49723462746167/page/0/1