

Feature Extraction Process for Sensor Data

1 Introduction

This document provides a detailed description of the feature extraction process used in the analysis of sensor data collected during experimental sessions. The focus is on accurately representing the mathematical operations and transformations applied to the raw data to generate features that are used in subsequent analysis. This supplement is intended to be a comprehensive reference for reproducing the feature extraction process.

2 Sensor Data and Preprocessing

2.1 Euclidean Norm of Acceleration

The Euclidean Norm, or the magnitude of acceleration, is computed for each sensor by combining the three axes of acceleration. This represents the overall acceleration experienced by the sensor.

$$\|a(t)\| = \sqrt{a_x(t)^2 + a_y(t)^2 + a_z(t)^2} \quad (1)$$

2.2 Filtering

The sensor data undergoes filtering to isolate relevant components. Two types of filters are applied: low-pass (LP) and high-pass (HP).

2.2.1 Low-Pass Filter

The low-pass filter is designed to capture the gravitational component of the acceleration, removing higher-frequency noise. The cutoff frequency is set at 2 Hz.

The low-pass filter was implemented using a butter function from the SciPy library, setting the order of the filter to 1.

2.2.2 High-Pass Filter

A high-pass filter is applied to the raw acceleration data to remove the low-frequency components (e.g., gravity), preserving the dynamic movements of the subject. The cutoff frequency is set at 2 Hz.

The high-pass filter was implemented using a butter function from the SciPy library, setting the order of the filter to 1.

Both filters were used on the raw acceleration data as well as the Euclidean Norm of acceleration signals.

2.3 Pitch and Roll Calculation

Pitch and roll angles were computed from the raw and filtered 3-axis accelerometer data, representing the orientation of the sensor relative to the Earth's surface.

2.3.1 Pitch

$$\text{Pitch} = \arctan \left(\frac{a_x}{\sqrt{a_y^2 + a_z^2}} \right) \quad (2)$$

2.3.2 Roll

$$\text{Roll} = \arctan \left(\frac{a_y}{a_z} \right) \quad (3)$$

Both were implemented using an arctan2 function from the NumPy library.

3 Feature Calculation

3.1 Statistical Features

For each 2-second window of data, the following statistical features were computed:

$$\text{Mean} = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

$$\text{Skewness} = \frac{1}{N\sigma^3} \sum_{i=1}^N (x_i - \bar{x})^3$$

$$\text{Kurtosis} = \frac{1}{N\sigma^4} \sum_{i=1}^N (x_i - \bar{x})^4 - 3$$

Median = middle value of the sorted data

Minimum = $\min(x_i)$

Maximum = $\max(x_i)$

25th Quantile = $\text{Quantile}(x, 0.25)$

75th Quantile = $\text{Quantile}(x, 0.75)$,

where x_i are the data points, \bar{x} is the mean, N is the number of data points and σ is the standard deviation.

All of those features were obtained using functions from the NumPy library.

3.2 Frequency Domain Features

3.2.1 Fourier Transform

The Fourier Transform was used to convert the time-domain signal into the frequency domain, allowing for the extraction of frequency-related features. The magnitude vector \hat{x} was defined as an absolute values vector of Fourier transforms:

$$\hat{x} = \left\{ \left| \hat{f}(\xi) \right| \mid \xi \in \left[0, +\frac{N}{2} \right] \right\}, \quad (4)$$

where N is the number of data points.

For each vector \hat{x} the following values were calculated:

$$\text{Energy } E = \sum \hat{x}^2$$

$$\text{Entropy } H = - \sum \frac{p \cdot \log_2(p)}{\log_2\left(\frac{N}{2}\right)}, \text{ where } p = \frac{\hat{x}}{\sum \hat{x}}$$

$$\text{Centroid } c = \xi \cdot p, \text{ where } \xi = \left\{ \xi \mid \xi \in \left[0, +\frac{N}{2}\right] \right\}$$

$$\text{Bandwidth } b = \delta \cdot p, \text{ where } \delta = \xi - c$$

$$\text{Maximal frequency} = \operatorname{argmax} \left(\hat{f}(\xi) \right).$$

All of the aforementioned features were computed using the NumPy library.

3.3 Summary Features

To capture the overall motion and coordination of the whole body, sums and sums of absolute values were computed across sensor locations and across axes. Summary features were calculated by aggregating sensor data:

$$\text{Sum} = \sum_i x_i$$

$$\text{Absolute Sum} = \sum_i |x_i|$$

The mean of each sum was taken as a final feature.

3.3.1 Sums Across Sensors

The mean sums were computed for all axes (or magnitudes in the case of Euclidean Norms) of the same signal type across all sensor locations, for all types of signals.

3.3.2 Sums Across Axes

For all of the locations, mean sums of all axes of the same signal type were computed, for all types of signals.

3.4 Difference Features

To capture the relative motion and coordination between different body parts, differences and differences of absolute values were computed across sensor locations and across axes. The differences between pairs of sensor readings were computed as:

$$\text{Difference}_{ij}(t) = x_i(t) - x_j(t)$$

$$\text{Absolute Difference}_{ij}(t) = |x_i(t)| - |x_j(t)|$$

The mean of each difference was taken as a final feature.

3.4.1 Differences Across Sensors

Differences across sensor locations were computed for each unique pair of axes (or magnitudes in the case of Euclidean Norms) of the same signal type across all sensor locations, for all types of signals.

3.4.2 Differences Across Axes

For all of the locations, differences across axes were computed for each unique pair of axes (3 combinations in case of 3-axis signals) of the same signal type, for all types of signals. They were not computed for signals with just one axis – magnitude signals.

3.5 Correlation Features

To capture the relative motion and coordination between different body parts, correlations and correlations of absolute values were computed across sensor locations and across axes. The Pearson correlation coefficient between two sensor signals $x_i(t)$ and $x_j(t)$ was computed over a time window with N data points:

$$\rho_{ij} = \frac{\sum_{t=1}^N (x_i(t) - \bar{x}_i)(x_j(t) - \bar{x}_j)}{\sqrt{\sum_{t=1}^N (x_i(t) - \bar{x}_i)^2 \sum_{t=1}^N (x_j(t) - \bar{x}_j)^2}} \quad (5)$$

where \bar{x}_i and \bar{x}_j are the mean values of the signals over the window T .

3.5.1 Correlations Across Sensors

Correlations across sensor locations were computed for each unique pair of axes (or magnitudes in the case of Euclidean Norms) of the same signal type across all sensor locations, for all types of signals.

3.5.2 Correlations Across Axes

For all of the locations, correlations across axes were computed for each unique pair of axes (3 combinations in the case of 3-axis signals) of the same signal type, for all types of signals. They were not computed for signals with just one axis – magnitude signals.

4 Feature Count and Contribution – for a Trunk & Legs Model

Each step in the feature extraction process contributes to the overall feature set. Below is a breakdown of the number of features generated at each step:

4.1 Preprocessing Data Streams

We began with the original 36 data streams (3 sensor types \times 3 axes \times 4 locations), and applied the following preprocessing steps:

- Computed roll and pitch from the acceleration data, resulting in 8 new streams (2 streams \times 4 locations).
- Applied low-pass and high-pass filtering to the acceleration data, yielding 24 filtered streams (2 filters \times 3 axes \times 4 locations).
- Computed the Euclidean Norms of the acceleration data and applied the same filtering process, generating 12 streams (4 locations \times 3 acceleration types).

This resulted in a total of 80 input data streams, which were then passed to the feature extraction step.

4.2 Feature Extraction

When mentioning data types for 3-dimensional signals, we refer to Acc, AccLP, AccHP, Mag, and Gyro. For 1-dimensional signals, we refer to AccNorm, AccNormLP, AccNormHP, Roll, and Pitch.

The following features were extracted:

- **Statistical Features:** 9 features, including mean, variance, skewness, kurtosis, maximal value, minimal value, median, 25th quantile, and 75th quantile (9 features \times 80 streams = 720 features).
- **Frequency Features:** 5 features, including energy, entropy, maximal frequency, centroid, and bandwidth (5 features \times 80 streams = 400 features).
- **Differences and Correlations:**
 - Across axes: Differences and correlations of original and absolute values for 3-dimensional data streams (2 measures \times 2 types of values \times 3 combinations of axes \times 4 locations \times 5 data types = 240 features).
 - Across locations: Differences and correlations of original and absolute values for all data streams (2 measures \times 2 types of values \times 6 combinations of locations \times (3 axes \times 5 data types + 1 axis for Norms \times 5 data types) = 480 features).
- **Sums:**
 - Across axes: Sums of original and absolute values for all data streams (2 types of values \times 4 locations \times 5 data types = 40 features).
 - Across locations: Sums of original and absolute values for all data streams (2 types of values \times (3 axes \times 5 data types + 1 axis for Norms \times 5 data types) = 40 features).

In total, 1920 features were extracted.