Human Gait Cycle Classification Improvements Using Median and Root Mean Square Filters Based on EMG Signals

M F Ghalyan1, M A Alher1 and M J Jweeg2
1 Department of Mechanical Engineering, University of Kerbala, Iraq.
2 Al-Farahidi University, Baghdad, Iraq.
E-mail: eng.master95@gmail.com

Abstract. Human walking is an important and intuitive operation in daily life, but it is different for people missing this ability. Recognizing human gait cycle portions during walking is very useful for understanding the biomechanics of the muscles, pre-disease’s diagnosis, and designing lower limb prosthetics. The normal gait cycle is divided into stance and swing phases. In this work, benchmarking of several classification techniques is performed, based on electromyography (EMG) data collected from seven lower limb muscles, with the gait cycle phase used as a target vector to label the EMG data. The dataset is split into a training set used to build a statistical model using a specified classification technique, and a test set for purpose of testing a pre-generated statistical model. EMG signals are unfortunately normally corrupted with noise, which drastically reduces their classification performance. A median filter and root mean square (RMS) filter were thus applied to the raw EMG signals, and the performance of the classification techniques calculated in all cases for the raw EMG, EMG with median filter, and EMG with RMS filters for comparison purposes. The use of filters offered good enhancement in the classification process. The median filtered EMG signal performed best and gave higher accuracy than the raw EMG and RMS filtered EMG signals due to the technique’s efficacy in removing outliers. This work also offers a simple explanation of each classifier algorithm used, and a graphical feature selection was also applied to all seven muscles of interest to identify the muscle with the most influence on the gait cycle. The Rectus Femoris muscle shows the best activity separation between swing and stance phases, working mainly during the swing phase; this higher activity during the swing phase thus caused it to be identified as the muscle with the most influence over the normal gait cycle.

Keywords: Normal gait cycle, electromyography, classification, median filter, root mean square filter, feature selection.

1 Introduction
The number of amputees globally is increasing for many reasons, including, diabetic complications, congenital deformations, and car accidents, as well as resources for amputations being more readily available. Several researchers have thus devoted their work to designing, analyzing, and suggest new shapes for prostheses, depending on the type and level of amputations. Most research of this type in Iraq has been on manufacturing prostheses to meet the needs of rehabilitation centers, especially with regard to using new materials in prostheses sockets and orthotic device manufacturing; however, some work on the use of EMG signals in motion control has also emerged [1-5].

Electromyography (EMG) is an experimental application concerned with the improvement, recording, and inspection of myoelectric signals generated by the variation of physiological muscle
fibres’ membrane which is excited by a single or multi-motor unit [6]. Such EMG signals are used in various applications and approaches, including medical research, rehabilitation, and sport science, and they have also been identified as offering a potential source of control for artificial limbs, especially with regard to lower limb prostheses. Their use in the classification of human motion may also play a vital role in the advancement of prosthesis and orthosis work.

Raw EMG signals are usually corrupted with noise, however, as shown in Figure 1. This noise mostly emerges from the activity of nearby muscles, creating “cross-talk” and ECG spikes, or the relative movement of electrodes with respect to the muscle under consideration; external noise sources are also an issue.

![Figure 1. Raw EMG Signal.](image)

Noisy signals may lead to undesirable output in a prosthesis, electromechanical delay, or poor classification and regression, and thus EMG signals should be refined by applying physical or mathematical filtering techniques. This research attempts to achieve a better signal by using two filtering techniques, the **Median filter** and **Root Mean Square (RMS) filter**, and making a comparison between them based on observation of various classification algorithms’ performance.

Classification is the process of dividing an unsorted data set that is not yet labeled based on prior examples from a different already labeled data set through a process called training. Classification produces a discrete output (labels), unlike regression which offers continuous output. Several researchers have developed techniques to classify human gait cycle phases: some have tracked the motion of joint positions to classify gait phases [7], while Joshi et al. [8] used EMG signals with Bayesian Information Criteria (BIC) to classify 8 eight phases as an application for Exoskeleton orthotic device. Panahandeh et al. [9] used a microelectromechanical-systems inertial measurement unit (IMU) mounted on the subject’s chest and applied a continuous hidden Markov model to classify the human gait phases into stance and swing phases based on several movement types such as walking, running, going upstairs, and going downstairs. Ziegier et al. [10] used EMG signal data with a support vector machine algorithm to classify the two main gait phases (swing and stance), proposing a new method for normalising the EMG signal called weighted signal difference (WSD).

The current work sought to improve the classification of gait cycle phases into swing and stance phase using EMG signals with the use of two filters (Median filter and Root mean square filter); a comparison was then made between the performance of each classification algorithm used with each type of the filter. This allowed the identification of the most suitable filter-classification technique for use with EMG signals in terms of the normal gait cycle.

2 Problem Description

Figure 3 illustrates both stance and swing phases respectively. While Figure 2 shows the activity of the soleus muscle alongside the gastrocnemius muscle representing the prime plantar-flexor for both phases. As shown, it is difficult to distinguish between the two phases using raw data due to the presence of noise. To classify the phases, a binary classification model must be applied, as shown in equation (1)
where $x(t)$ is a specific muscle activity (EMG), and thus $y_{\text{phase}}(t) \in \mathbb{R}^n$, $x(t) \in \mathbb{R}^n$, where $n$ is the number of muscles under observation. In this paper, seven lower limb muscles are considered: the soleus, tibialis anterior, gastrocnemius lateralis, vastus lateralis, rectus femoris, biceps femoris, and gluteus maximus.

Figure 2. Stance phase (A) and Swing phase (B) for a single cycle.

The main objective of this paper is to enhance the classification process discussed in equation (1) using two well known filtering techniques and to compare the effects of filtering techniques and classification algorithms to develop better information about working with this type of signal.

3. Gait cycle Models
The gait cycle refers to the advancement of a body based on the movement of the lower limbs. The normal gait cycle in humans is classified into stance and swing phases, and the cycle is the duration of movement of a single limb between an event and the next consecutive event of the same type (e.g. heel strike to heel strike). The stance phase takes place when the limb under consideration is in contact with the ground, and this occupies 60% of the gait cycle, while the swing phase takes place while the limb is off the ground, which occupies 40% of the gait cycle [7]. As shown in Figure 3, the gait cycle can be further classified into sub-phases such as heel strike or initial contact (IC), loading response (LR), mid-stance (MST), terminal stance (TST), and pre-swing (PSW); all of these sub-phases occur during the stance phase, while toe-off (TO), mid-swing (MSw), and terminal swing (TSw) occur in the swing phase [11]. The main concern of this paper is to classify the two main phases, however.
All datasets used in this paper were downloaded from the HuMoD database [12], an open database concerned with the inspection and recording of human motion dynamics that offers well-arranged and documented datasets mainly concerned with the lower limbs. The database includes three-dimensional motion tracking data gathered using sophisticated cameras and markers adhered to a subject moving on a treadmill with a force plate underneath it to record the ground reaction force at 1,000 Hz. It also records all contact events of the lower limb such that if the limb is in contact with the ground it records 1, and otherwise it records 0, which offers an effective means of labelling the data for classification purposes. Seven electromyography devices are also mounted to each leg muscle to record muscle activity while subjects perform required movements, such as straight walking, at several speeds, straight running at several speeds, sideways walking, kicking a soft football, and so on. The EMG data is supplied as raw data and filtered data, and the latter is obtained using a root mean square filter with a window size of 100. The EMG data is obtained from seven muscles namely the Soleus muscle, Tibialis anterior muscle, Gastrocnemius Lateralis muscle, Vastus Lateralis muscle, Rectus Femoris muscle, Biceps Femoris muscle, and Gluteus Maximus muscle. The data used in this paper is that for straight walking at 1.0 m/s, based on a female subject (age 27, 161 cm tall, and 57.3 kg in mass). The dataset is in. mat format. One of the problems with this data set, however, is that the EMG recording frequency (frame rate) is at 2,000 Hz, while the force plate works at 1,000 Hz, which makes data pre-processing necessary to achieve equal reading rates for both sets of observations. Python was used for this pre-processing and for the ensuing classification in this paper. The unequal frequency of the EMG and contact events was handled by creating a new zero vector twice the size of the contact events vector; the data was then fed not these zero vectors with even indices (Python starts with 0 indexing), causing each in-between empty odd cell to its prior cell.

| Table 1. Details of subject |
|-----------------------------|
| Gender | Female |
| Age    | 27     |
| Height | 161 cm |
| Weight | 57.3 kg|
| Origin | Central Europe |
| Clothing | Underpants, sports bra |
| Date   | April 2014 |

4. **Signal Filtering Techniques**

Raw EMG signals deliver valuable data about specific muscles, but it is difficult to apply the required analysis to raw signals for use in rehabilitation or biomechanical engineering unless the signal is rectified.
4.1 Median Filter
Due to the random nature of the noise introduced to EMG signals, it is difficult to achieve good results simply by applying linear filters. The median filter is a non-linear filter type most often used for image filtering to deal with scattered light from different environmental sources introduced to an image. It utilizes a **moving filter window**, where you apply the filter for a certain range around the point under consideration (in images, pixels are considered as the points in question). The median is the middle value in a set, for example, if we have a set of \( s = \{1, 2, 3, 4, 5\} \), then the median of \( s \) is 3. The window size can be formulated as follows:

\[
N = 2k + 1
\]  

(2)

\( N \) should be always odd to ensure that the point under consideration exists as a real number in the middle. Let’s consider a signal of 100 simple points, to apply a median for a sample of 20 points with a window size of 7, \( N = 7 \), and by applying equation (2), \( k = 3 \).

\[
\begin{align*}
0 & \quad 1 & \quad 2 & \quad 3 & \quad \cdots & \quad 17 & \quad 18 & \quad 19 & \quad 20 & \quad 21 & \quad 22 & \quad 23 & \quad \cdots & \quad 100 \\
\end{align*}
\]

![Image of the moving filter window](image.png)

In Python, this process can be applied using single class in a **scipy** library with the **medfilt** method.

The mean filter was thus used to spread overshooting over all the specified windows, while the median filter was used to eliminate this overshooting [13].

4.2 Root Mean Square Filter
For the same window size of the median filter, as in equation (2), the root mean square can be formulated as follows

\[
RMS = \sqrt{\frac{1}{N} \sum_{i=n-k}^{n+k} x_i^2}
\]  

(3)

where \( n = 0 \ldots \) final sample, and \( N \) is the window size.

Unfortunately, the Python signal library does not natively include an RMS filter for the raw EMG signal as shown in Figure. A Python function was thus created for this work, as shown in Figure [14].
5. **Gait Classification algorithms**

The pattern recognition field offers many classification techniques; in this paper, four well-known classification algorithms were used as summarised below. Further details about more general statistical modeling and classification can be found in machine learning literature [15].

### 5.1 Support Vector Machine (SVM)

For a set of training data $D = \{(x_1, y_1) \ldots (x_n, y_n)\}$, SVM tries to find the optimal line separating the data within different labels by maximizing the distance between support vectors, the vectors separating the various data labels. SVM can be formulated as

$$
\min_{w, \varepsilon_i} J(w, \varepsilon_i) = \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \varepsilon_i
$$

subject to

$$
y_i \varphi(w^T x_i) \geq 1 - \varepsilon_i
$$

---

**Figure 4.** Raw EMG with RMS of EMG signal.

**Figure 5.** RMS python function.
where \( w \) and \( e_i \) are SVM parameters, and \( c \) is a constant vector. These equations can be solved using the Lagrange classifier [15].

5.2 **K-Nearest Neighbour (K-NN)**
This approach finds the specified nearest point and compares the number of classes there, then labels the unknown point under consideration as the highest repeatable class within a certain distance. K-NN can be formulated as

\[
\hat{y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i
\]  

(6)

where \( \hat{y}(x) \) is the K-NN predicted label, \( N_k \) is a vector of the neighbouring points, and the \( N \) vector is of a size equal to \( k \). Several approaches can then be applied to find the distances from the unknown point to its neighbouring points, as follows:

**Euclidean distance**
\[
d(x_i, x_j) = \sqrt{\sum_{i=1, i \neq j}^k (x_i - x_j)^2}
\]

(7)

**Manhattan distance**
\[
d(x_i, x_j) = \sum_{i=1, i \neq j}^k |x_i - x_j|
\]

(8)

**Minkowski distance**
\[
d(x_i, x_j) = \left( \sum_{i=1, i \neq j}^k (x_i - x_j)^p \right)^{\frac{1}{p}}
\]

(9)

5.3 **Logistic Regression**
For a data set as shown in Figure (A), where the data is labelled either 0 for a certain range of \( x \) or 1 for the reset range, a linear model can applied; however, for a binary classification problem, the results should be either 1 or 0, and thus the upper and lower parts (marked with blue circles in the figure) can be trimmed, as shown in Figure (B). Applying a sigmoid function to the linear model can result in the figure shown in Figure 7.

![Figure 6. Linear model](image-url)
\[ y = b_0 + b_1 x \]  \hspace{1cm} \text{Linear model} \tag{10}

\[ p = \frac{1}{1 + e^{-y}} \]  \hspace{1cm} \text{Sigmoid function} \tag{11}

\[ \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \]  \hspace{1cm} \text{Logistic regression} \tag{12}

where \( p = \) probability of the prediction, allowing the setting of a threshold (such as 0.5) where, if \( p \) has passed this point, \( y \) will equal 1, and otherwise, it equals zero. Shalev-Shwartz et al. [16] offer further information about other classification techniques.

### 6. Results and Discussion

The filters discussed above were applied to facilitate classification and comparison. Data was fed into the training functions of the classification algorithms for seven left limb muscles (soleus, tibialis anterior, gastrocnemius lateralis, vastus lateralis, rectus femoris, biceps femoris, and gluteus maximus) and their equivalent phases for a 60-gait cycle. The Scikit Learn (sklearn) Python library was used to perform classification using Spyder IDE on a laptop with Windows 10 running on a core i7 processor with a clock speed of 2.40 GH and 32 GB of ram as well as an Nvidia Quadro graphics card with 4 GB of dedicated memory. HuMoD data for “female straight walking” at 1.0 m/s was used, in particular, the data between 20 s and 80 s, selected as the force plate device started recording at 20 s and stopped at 80 s from the start of the experiment. For each classification case, performance was calculated using a confusion matrix composed of both true and false predicted samples from the test set. Each classification algorithm was assigned to data without filtering and then to data with the median and RMS filters separately.

**Table 2. Classification technique performance for Raw EMG, EMG with Median filter, and EMG with RMS filter.**

| Scheme          | Raw Data | Median filter | RMS filter |
|-----------------|----------|---------------|------------|
| SVM             | 65.8     | 95.3          | 91.7       |
| Logistic Regression | 63.3     | 63.3          | 72.9       |
| K-NN            | 69.2     | 99.8          | 99.3       |
| Decision Tree   | 62.8     | 98.4          | 66.3       |
| Random Forest   | 67.6     | 97.2          | 67.6       |

Classification accuracy increased with the application of filters, and in all cases, the K-nearest neighbour technique showed the best performance, as despite the data’s noisy nature, since data were clustered for each phase. The median filtered does a better job than the RMS filter based on its ability to remove outliers, however. The success percentage ratio is increased markedly with the application...
of filters; for example, the support vector machine approach saw a percentage ratio increment of 144% for the median filter and 139% for the RMS filter.

Feature selection refers to the process of selecting the single most influencing variable within a classification model. In this case, this refers to the muscle that has the greatest influence on gait cycle classification, which must be the muscle with the greatest difference in activity value for each phase. Feature selection is useful as it makes the model easier to apply and reduces the training time; it can also be used to remove redundant and irrelevant data. Several mathematical techniques can be applied to specify this variable, and in this paper, a simple graphical method was applied. This graphical method involves plotting each muscle activity value as equal on the x and y axes then colourizing each phase separately to identify the muscle with the most obvious separation between phases. The results for the seven lower limb muscle plots for stance and swing phases are shown in Figure 6.
Figure 8. Muscle activity over a cycle for each of seven muscles under consideration

The symbol L in the titles of previous figures represents the left lower limb, and the muscle abbreviation are as shown in Table 3

| Abbreviation | Full muscle name          |
|--------------|---------------------------|
| SOL          | Soleus muscle             |
| TIA          | Tibialis Anterior muscle  |
| GLS          | Gastrocnemius Lateralis muscle |
| VSL          | Vastus Lateralis muscle   |
| RCF          | Rectus Femoris muscle     |
| BCF          | Biceps Femoris muscle     |
| GLX          | Gluteus Maximus muscle    |

Observation of the plots and the maximum and minimum activation values for each muscle suggests that the rectus femoris muscle has the best separation of phase activity. The rectus femoris muscle also has a high level of influence on knee extension, working mainly in the swing phase, giving it a far larger level of activity in the swing phase than the stance phase, and making these phases more easily discernible.

7. Conclusion
In this paper, the gait classification problem was investigated using various classification techniques based on electromyographical (EMG) signals, used to represent a muscle activity based on the voltage induced in muscle fiber during contraction. Gait cycle classification is valuable in many applications, including prosthesis development. Classification was performed for the two main gait cycle phases, stance phase, and swing phase. EMGs from seven lower limb muscles (soleus, tibialis anterior, gastrocnemius lateralis, vastus lateralis, rectus femoris, biceps femoris, and gluteus maximus) were used as a training dataset to create a statistical model, with the contact of the limb with the ground (based on force plate measurements) was used as a target vector to determine whether the subject was in stance phase or swing phase. In terms of splitting the data set, a 3/20 split was used for testing data and training data, with all data collected from the HuMoD Database.
EMG signals are known for their noisy nature; two types of filters were thus tested for improvement to classification accuracy, the root mean square filter (RMS) and the median filter. A comparison was then made between the two filters based on classification accuracy. Five types of classification techniques were used: support vector machine (SVM), logistic regression, K nearest neighbour (KNN), decision tree, and random forest. All classification techniques’ accuracy was enhanced after implementing filters, showing that noise was removed effectively. The median filter showed the best ability in terms of optimizing the classification process, which was not as originally predicted. This may be due to the effective removal of outliers by the median filter. The K-NN classification technique had the best classification ability under all filters, however, although the data is not linearly distributed and but values were clustered together.

Feature selection is the process of seeking the most influential variable within a statistical model, in this case, the muscle with the most effect on the classification process. A graphical method was used to identify the rectus femoris as the muscle with the largest impact on gait cycle identification. Gait cycle classification for each sub-phase could be done in future investigations, though this would require data labeled for each sub-phase. Applying a new classification technique in this way may, however, produce better classification performance.

References

[1] Jweeg MJ, Ameen SH. Experimental and Theoretical Investigations of Dorsiflexion Angle and Life of an Ankle-Foot-Orthosis Made From (Perlon/Carbon fiber/Acrylic) and Polypropylene Materials. Journal of the Japanese Society for Experimental Mechanics. 2011;11(Special_Issue):s305-s10.

[2] Ismail MT, Jweeg MJ, Resan KK. Study of Creep-Fatigue Interaction in the Prosthetic Socket below Knee. Innovative Systems Design and Engineering. 2013;4(5):35-42.

[3] Kadhim AA, Al-Waily M, Ali Z, Jweeg MJ, Resan KK. Improvement fatigue life and strength of isotropic hyper composite materials by reinforcement with different powder materials. International Journal of Mechanical & Mechatronics Engineering IJIME-IJENS. 2018;18(02).

[4] Jweeg MJ, Ahumdany AA, Jawad AM. Dynamic Stresses and Deformations Investigation of the Below Knee Prosthesis using CT-Scan Modeling. International Journal of Mechanical & Mechatronics Engineering IJIME-IJENS. 2019;19(01).

[5] Abbas EN, Jweeg MJ, Al-Waily M. Fatigue Characterization of Laminated Composites used in Prosthetic Sockets Manufacturing. Journal of Mechanical Engineering Research and Developments, 2020, Vol. 43, No. 5, pp. 384-399.

[6] Konrad P. The abc of emg. A practical introduction to kinesiological electromyography. 2005;1(2005):30-5.

[7] Taborri J, Palermo E, Rossi S, Cappa P. Gait partitioning methods: A systematic review. Sensors, 2016;16(1):66.

[8] Joshi CD, Lahiri U, Thakor NV, editors. Classification of gait phases from lower limb EMG: Application to exoskeleton orthosis. 2013 IEEE Point-of-Care Healthcare Technologies (PHT); 2013: IEEE.

[9] Panahandeh G, Mohammadiha N, Leijon A, Händel P. Continuous hidden Markov model for pedestrian activity classification and gait analysis. IEEE Transactions on Instrumentation and Measurement. 2013;62(5):1073-83.

[10] Ziegler J, Gattringer H, Mueller A, editors. Classification of gait phases based on bilateral emg data using support vector machines. 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob); 2018: IEEE.

[11] Perry J, Davids JR. Gait analysis: normal and pathological function. Journal of Pediatric Orthopaedics. 1992;12(6):815.

[12] Wojtusch J, von Stryk O, editors. Humod-a versatile and open database for the investigation, modeling and simulation of human motion dynamics on actuation level. 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids); 2015: IEEE.
[13] Poularikas AD. Understanding Digital Signal Processing with MATLAB® and Solutions: CRC Press; 2017.
[14] Farfán FD, Politti JC, Felice CJ. Evaluation of EMG processing techniques using information theory. Biomedical engineering online. 2010;9(1):72.
[15] Bishop CM. Pattern recognition and machine learning: springer; 2006.
[16] Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms: Cambridge university press; 2014.