A Systematic Method for Outlier Detection in Human Gait Data

Bradley Hobbs and Panagiotis Artemiadis*, *IEEE Senior Member*

***Abstract*— When it comes to observing and measuring human gait data for further analysis, determining whether the observed behavior is within the normal range of variability, or should be considered abnormal, is very challenging. Moreover, usually gait data are multivariate including motion capture, electromyography, force measurements, etc., each source having its own unique causes of irregularities and anomalies. This paper introduces a unique algorithm for outlier detection in periodic gait data using multiple sources and multiple procedures to improve the overall accuracy. The proposed algorithm's performance is evaluated using realistic synthetic gait data to gauge its accuracy to a truly objective known solution. It is shown that the proposed method is able to detect 91.2% of the true outliers in an extensive synthetic dataset, while only producing false positives at a rate of 0.1%, outperforming other procedures usually utilized in gait data outlier detection. The proposed method is a systematic way of removing outliers from gait data, with direct applications to human biomechanics, rehabilitation and robotics, and can be applied to other scientific fields dealing with periodic data.**

I. Introduction

Biological life will always present a natural variability, inherent to both the environment's physics and to the past decisions made by such creature. When it comes to observing human behaviors, determining whether the observed behavior is within the normal range of variability, or should be considered abnormal is very challenging. Furthermore, when studying a human's reaction to a specific intervention, determining whether the observed behavior is consciously induced, or is a product of sensor noise or other artifacts has been a major problem in many science fields. The above necessitate the need for anomalies, or outliers, to be detected and removed from the recorded data in a systematic and non-biased way [1].

Outliers are defined in many ways, depending on the application or field of study [2, 3]. The formal definition for the purpose of this work is: "*An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*" [1]. Focusing on human biomechanics data, outliers can originate from many different sources: ElectroMyoGraphic (EMG) data can be corrupted with sensor noise that originates from sensor detachment from the skin, change of skin impedance etc [4, 5]. Motion capture data usually include outliers originating from skin detachment [6], clothing interference [7], mislabeled markers, or comfort-related intentional sources such as postural adjustments [8]. It is often the case that without proper outlier removal, many study conclusions can be altered. It is then established, that the goal is to remove outliers while preserving the natural biological variation [9].

Many methods have been proposed to remove outliers from data that exist on one dimension [10, 11], but there are limited methods for outlier detection in large, complex and multivariate data [12, 13]. Outlier detection methods for either case can be grouped based on the categories of techniques that are employed [14], but few have proposed combining outlier detection methods to increase performance relative to single methods [15].

In this work, methods that apply to gait-specific data are of interest, and observing the data from a different lens can be useful in determining outliers. Gait data usually include motion capture, EMG signals, and ground reaction force recordings. Therefore, there is a need for determining outliers in multivariate data that have inherent but different variability among each other. Walking rhythm anomalies have been detected from video data using a Fast Fourier Transform (FFT) to observe the data from a frequency perspective [16]. Principal Component Analysis (PCA) has been used to aid in outlier detection [17] with good performance against a human control [18]. A simple method using the integral of the signal can also be used to select outliers in gait data [19], taking into account multiple kinematic variables. However, outlier detection procedures are usually completed manually without much afterthought, while considering multiple variable sources as input is often overlooked. Therefore, there is a need for a systematic and thoughtful outlier detection method for gait-related data.

This paper introduces a technique to implement outlier detection by combining and building upon many known and established methods that are known to be useful for gait-specific data, but can be generalized to any periodic data. The generalized technique performs initial tentative outlier labelling using each method individually, with tuned parameters for gait specific data, then performs a decision algorithm employing significance testing to determine a quantitative threshold for the tentatively labelled data to be considered outliers, taking into consideration all sources that are input into the algorithm, that commonly include EMG or kinematic data. This technique outperforms singular methods in outlier detection. The proposed algorithm's performance is evaluated with realistic, but artificially created gait data to verify against a truly objective known solution. We show that this algorithm outperforms the individual methods chosen, in

*This material is based upon work supported by the National Science Foundation under Grants No. #2020009, #2015786, #2025797, and #2018905.

Bradley Hobbs and Panagiotis Artemiadis are with the Mechanical Engineering Department, at the University of Delaware, Newark, DE 19716, USA. bwh@udel.edu, partem@udel.edu

*Corrresponding author: partem@udel.edu

minimizing both false negatives and false positives. The proposed method is a systematic way of removing outliers from gait data, with direct applications to human biomechanics, rehabilitation and robotics, while being generalizable enough to be applied to other scientific fields utilizing periodic data.

II. METHODS

This paper introduces an outlier detection algorithm (ODA) consisting of a combination of a flagging procedure and a post-processing decision making method for removing outliers in human gait data that could consist of EMG recordings and kinematic data. First, we present the data pre-processing steps required before algorithm input. Then, the flagging process and subsequent procedures are presented, followed by the post-processing decision process[1].

A. Input Data Description

The input data sources that are accepted to use with this method are any periodic waveforms that would normally be acquired during gait studies. Such data could for example consist of muscle EMG signals, joint kinematics, force plate measurements, or inertial measurement unit data such as accelerometer, gyroscope or magnetometer. Let $n_s \in \mathbb{Z}^+$ be the number of sensors we record data from. All data should be temporally synchronized and then partitioned such that all sources contain $n_c \in \mathbb{Z}^+$ periods of data, or gait cycles, all starting from a periodic event that is the same across all sources. For example, a periodic event that is usually used in gait data is the heel-strike of one foot, with many methods available to determine it robustly from raw kinematic data [20, 21]. Assume that all sensors have either the same sampling frequency $f_s \in \mathbb{R}^+$, or data from them can be resampled to have the same final sampling frequency. Then a $n_c \times n_p$ matrix $\boldsymbol{A^{(i)}}$, can be defined to include all n_c gait cycles of the data, with $n_p \in \mathbb{Z}^+$ samples for each gait cycle, coming from each sensor $i = 1, 2, \cdots, n_s$. If we vertically concatenate the data matrices $\boldsymbol{A^{(i)}}$ from each sensor i, then we arrive at the matrix

$$\boldsymbol{A} = \left[\boldsymbol{A^{(1)}}\ \boldsymbol{A^{(2)}}\ \cdots\ \boldsymbol{A^{(n_s)}}\right]^T \quad (1)$$

which describes the $n_c n_s \times n_p$ input data to the ODA. It must be noted that all standard pre-processing of the data should be done before including them in the algorithm data input. For example, the linear envelop method for EMG data [22] should be applied before any processing of the data by the proposed algorithm. Moreover, if the data originates from a trial with at least one independent condition from a baseline or control, data belonging to each condition should be separately considered in the proposed ODA. For the remaining of the paper, all data inputs will correspond to one condition, unless otherwise noted.

The proposed ODA includes two main steps: first a flagging step is implemented that includes procedures to flag each gait cycle as a possible outlier, for each sensor source. Then, a decision procedure follows to ultimately determine whether a gait cycle is an outlier or not.

[1]The code implementation of the entire proposed ODA method can be found at `github.com/HORC-Lab/ICORR_2022`

B. Flagging Procedures

The goal of the flagging step is to assign a binary flag $f_j^{(i)} \in \{0, 1\}$ to each gait cycle or row $j = 1, 2, \cdots, n_c$ of the input data submatrices $\boldsymbol{A^{(i)}}$, $i = 1, 2, \cdots, n_s$. Those flags are assigned for each sensor separately, and then combined in a flagging vector later. A value of 1 for the binary flag $f_j^{(i)}$ corresponds to a *potential* outlier, while 0 denotes the opposite, i.e. normal data. This binary status is determined through five independent and distinct procedures in order to increase the probability of accurate classification. Each procedure examines the data in a unique way, looking for specific characteristics in the data. A flagging vector $\boldsymbol{H^{(i)}_{\{1\}}}$, $\boldsymbol{H^{(i)}_{\{2\}}}$, $\boldsymbol{H^{(i)}_{\{3\}}}$, $\boldsymbol{H^{(i)}_{\{4\}}}$, $\boldsymbol{H^{(i)}_{\{5\}}}$ will be defined by each of the five distinct procedures respectively, and the final $f_j^{(i)}$ for each sensor and gait cycle will be decided based on those. The five procedures chosen are: Shape-based, Feature-based, Time-based, Amplitude-based, and Statistics-based.

1) Shape-Based Procedure: The shape of each gait cycle is defined by the acute visual shape of the waveform. The procedure finds potential outliers based on acute shape differences is a piece-wise variant of median absolute deviation technique (pMAD). This procedure is applied to each of the data submatrices $\boldsymbol{A^{(i)}}$, $i = 1, 2, \cdots, n_s$. Each one of the n_c gait cycles corresponding to the n_c rows of $\boldsymbol{A^{(i)}}$ has n_p data points (columns of $\boldsymbol{A^{(i)}}$). If $a_{j,l}^{(i)}$ is the (j, l) element of the matrix $\boldsymbol{A^{(i)}}$, then for each element the method computes the median absolute deviation (MAD) measure matrix $\boldsymbol{M^{(i)}}$, which (j, l) elements are given by: $m_{j,l}^{(i)} = \| a_{j,l}^{(i)} - \boldsymbol{\tilde{a}_l} \|$ where, $l = 1, 2, \cdots, n_p$, $j = 1, 2, \cdots, n_c$, $\boldsymbol{a_l}$ is the l-th column of $\boldsymbol{A^{(i)}}$, $\boldsymbol{\tilde{a}_l}$ is the median of the data in $\boldsymbol{a_l}$, and $median\,(\,)$ is the function that gives the median value of the values in a vector. Then, for each column l, the standard deviation of the MAD measures is computed and defined as:

$$\begin{aligned}\boldsymbol{\sigma^{(i)}} &= \left[\sigma_1^{(i)}\ \sigma_2^{(i)}\ \cdots\ \sigma_{n_p}^{(i)}\right] \\ &= \left[std\left(\boldsymbol{m_1^{(i)}}\right)\ std\left(\boldsymbol{m_2^{(i)}}\right)\ \cdots\ std\left(\boldsymbol{m_{n_p}^{(i)}}\right)\right] \end{aligned} \quad (2)$$

where $\boldsymbol{m_l^{(i)}}$, $l = 1, 2, \cdots, n_p$ are the columns of $\boldsymbol{M^{(i)}}$, and $std\,(\,)$ is the function that computes the standard deviation of the values in a vector. Finally, a binary matrix $\boldsymbol{B^{(i)}}$ of the same size as $\boldsymbol{M^{(i)}}$ is created, where each element $b_{j,l}^{(i)}$ is 0 if the $m_{j,l}^{(i)}$ is inside the range of 3 standard deviations from the median value $\overline{m_l}^{(i)}$ of each column l, i.e. if $m_{j,l}^{(i)} \in \left[-3\sigma_l^{(i)} + \overline{m_l}^{(i)}, 3\sigma_l^{(i)} + \overline{m_l}^{(i)}\right]$, $l = 1, 2, \cdots, n_p$, and 1 otherwise. The elements of the final flagging vector for this method $\boldsymbol{H}^{(i)}_{\{1\}} = \left[h^{(i)}_{\{1\},1}\ h^{(i)}_{\{1\},2}\ \cdots\ h^{(i)}_{\{1\},n_c}\right]^T$, are given by:

$$h^{(i)}_{\{1\},j} = \begin{cases} 0 & , if\ \frac{\sum_{l=1}^{n_p} b_{j,l}^{(i)}}{n_p} \leq t_1 \\ 1 & , otherwise \end{cases} \quad (3)$$

where $j = 1, 2, \cdots, n_c$, and t_1 is a threshold parameter for this method. The goal of this parameter is to set the minimum number of ones in each row of the binary matrix $\boldsymbol{B}^{(i)}$, above which, the gait cycle or period is flagged as a potential outlier. This parameter was set to 4% for our data to balance potential false positives and false negatives.

2) Feature-Based Procedure: The features of each period are defined by the distinctive peaks and valleys of the overall waveform. This procedure uses dimensionality reduction via standard principal component analysis (PCA) to determine potential outliers based on those features that do not appear in most gait cycles.

This procedure is applied to each of the data submatrices $\boldsymbol{A}^{(i)}$, $i = 1, 2, \cdots, n_s$. For a submatrix $\boldsymbol{A}^{(i)}$, first each column is transformed to have a zero mean by using:

$$\boldsymbol{A'}^{(i)} = \boldsymbol{A}^{(i)} - \left[\mu_1 \; \mu_2 \; \cdots \; \mu_{n_p}\right] \tag{4}$$

where μ_l is a vector $\in \mathbb{R}^{n_c}$ whose all elements are equal to the mean of the column l of the matrix $\boldsymbol{A}^{(i)}$, $l = 1, 2, \cdots, n_p$. Then the PCA method is applied and the matrix $\boldsymbol{W}$ is computed as the $n_p \times n_p$ matrix of weights whose columns are the eigenvectors of $\boldsymbol{A'}^{(i)} \boldsymbol{A'}^{(i)^T}$. More details on the PCA implementation can be found in [23]. By selecting only the first two principal components, we can compute a low-dimensional representation of the original data as: $\boldsymbol{A}_{\boldsymbol{L}}^{(i)} = \boldsymbol{A'}^{(i)} \boldsymbol{W}_{\boldsymbol{L}}$ where $\boldsymbol{W}_{\boldsymbol{L}}$ is a $n_p \times 2$ matrix including only the 2 columns (eigenvectors) associated with the 2 largest eigenvalues of $\boldsymbol{W}$, and $\boldsymbol{A}_{\boldsymbol{L}}^{(i)}$ is a $n_c \times 2$ low-dimensional representation of the data. The next step of the procedure is identical to the Shape-Based procedure mentioned above using the MAD, however in this case instead of the data in $\boldsymbol{A}^{(i)}$, the low-dimensional representation $\boldsymbol{A}_{\boldsymbol{L}}^{(i)}$ is used.

Following the same process as above (see Sec. II.B.1), the final flagging vector for this method $\boldsymbol{H}_{\{2\}}^{(i)} = \left[h_{\{2\},1}^{(i)} \; h_{\{2\},2}^{(i)} \; \cdots \; h_{\{2\},n_c}^{(i)}\right]^T$ is defined. The new threshold value t_2 included in (3) for this procedure is set to 50%, which will flag a gait cycle as a potential outlier if at least one out of the two dimensions of the data are outside the corresponding 3 standard deviations range.

It must be noted that the choice of the number of principal components to keep for the low-dimensional representation is a free choice, however for biomechanical periodic data, two principal components seem to suffice as they are able to explain about 70% of the original data variability.

3) Time-Based Procedure: The time-based changes of each gait cycle waveform can be due to common patterns such as lengthened or shortened walking stride and can be manipulated or lost when all gait cycles are resampled at the same number of data points; a common practice in gait analysis. However, those subtle changes can be found if the data is converted in the frequency domain after being resampled. This procedure determines the underlying frequencies and amplitudes associated with each gait cycle through the discrete Fast-Fourier Transform (FFT).

Similarly to above, this procedure is applied to each of the data submatrices $\boldsymbol{A}^{(i)}$, $i = 1, 2, \cdots, n_s$. For each column of $\boldsymbol{A}^{(i)}$, the discrete FFT is calculated and let $\boldsymbol{Y}^{(i)}$ be the resultant spectrum matrix. The next step of the procedure is identical to the Shape-Based procedure mentioned above using the MAD, however in this case instead of the data in $\boldsymbol{A}^{(i)}$, the spectrum matrix $\boldsymbol{Y}^{(i)}$ is used. Therefore, following the same process as above, the final flagging vector for this method $\boldsymbol{H}_{\{3\}}^{(i)} = \left[h_{\{3\},1}^{(i)} \; h_{\{3\},2}^{(i)} \; \cdots \; h_{\{3\},n_c}^{(i)}\right]^T$ is defined. The new threshold value for this procedure defined as t_4, similar to the one included in (3), is set to 10%. This is based on the frequency components of gait data, which are usually much lower (1-3 Hz) compared to the usually high acquisition frequency (usually 100 Hz).

4) Amplitude-Based Procedure: This procedure aims to detect outlier gait cycles with irregular peaks or valleys. Similarly to above, this procedure is applied to each of the data submatrices $\boldsymbol{A}^{(i)}$, $i = 1, 2, \cdots, n_s$. For each row of $\boldsymbol{A}^{(i)}$, which corresponds to each gait cycle, the area under the amplitude curve when plotted with respect to time is calculated using the standard trapezoidal integration (TI). This results to a vector of integral values $\boldsymbol{I}^{(i)} = \left[I_1^{(i)} \; I_2^{(i)} \; \cdots \; I_{n_c}^{(i)}\right]^T \in \mathbb{R}^{n_c}$, where each element corresponds to the integral value for each gait cycle (row in $\boldsymbol{A}^{(i)}$) given by:

$$I_j^{(i)} = \sum_{l=1}^{n_p - 1} \frac{a_{j,l}^{(i)} + a_{j,l+1}^{(i)}}{2} \tag{5}$$

where $j = 1, 2, \cdots, n_c$. Then, a similar approach to that of the MAD is followed. The median value of the $\boldsymbol{I}^{(i)}$ vector is calculated, and each of its elements are tested whether they belong in the 3 standard deviations range from the median value or not. From that, the final flagging vector for this method $\boldsymbol{H}_{\{4\}}^{(i)} = \left[h_{\{4\},1}^{(i)} \; h_{\{4\},2}^{(i)} \; \cdots \; h_{\{4\},n_c}^{(i)}\right]^T$, is calculated by:

$$h_{\{4\},j}^{(i)} = \begin{cases} 0 & , if \; I_j^{(i)} \in \left[-3\epsilon^{(i)} + \overline{I}^{(i)}, 3\epsilon^{(i)} + \overline{I}^{(i)}\right] \\ 1 & , otherwise \end{cases} \tag{6}$$

where $\overline{I}^{(i)}$ is the median value of the vector $\boldsymbol{I}^{(i)}$ and $\epsilon^{(i)}$ is its standard deviation.

5) Statistics-Based Procedure: The fifth and final procedure for flagging periods as potential outliers is a multi-dimensional variation of the iterative Generalized Extreme Studentized Deviate (GESD) method. Similarly to above, this procedure is applied to each of the data submatrices $\boldsymbol{A}^{(i)}$, $i = 1, 2, \cdots, n_s$. This method assumes that there is one outlier gait cycle (row) in $\boldsymbol{A}^{(i)}$, finds that row that maximizes the test statistic value R_g defined in [2], and removes that gait cycle from the data. This is done iteratively for $g - 1$ iterations. The final number of outliers is found by finding the highest number of iterations that give $R_g > \lambda_g$, where λ_g is the critical value corresponding to R_g [2]. Let $\mathbb{G}$ be the set of gait cycles that are considered outliers by this method, then the final flagging vector for this method $\boldsymbol{H}_{\{5\}}^{(i)} = \left[h_{\{5\},1}^{(i)} \; h_{\{5\},2}^{(i)} \; \cdots \; h_{\{5\},n_c}^{(i)}\right]^T$, is calculated by:

$$h_{\{5\},j}^{(i)} = \begin{cases} 0 & , if \; j \notin \mathbb{G} \\ 1 & , otherwise \end{cases} \tag{7}$$

For ease of use, the built-in Matlab (Mathworks) function *rmoutlier('gesd')* was used for this purpose.

C. Post-Flagging Decision Procedure

At this point, each gait cycle of the data has accumulated a number of possible flags noting it as a potential outlier. After applying the above 5 procedures for each of the data submatrices $\boldsymbol{A}^{(i)}$, $i = 1, 2, \cdots, n_s$ coming from all the possible n_s sensor sources, we can finally define the final flagging vector $\boldsymbol{F} \in \mathbb{R}^{n_c}$ using the following expression:

$$\boldsymbol{F} = [F_1 \; F_2 \; \cdots \; F_{n_c}]^T = \sum_{i=1}^{n_s} \sum_{k=1}^{5} \boldsymbol{H}_{\{k\}}^{(i)} \tag{8}$$

Each element of this vector has the total number of potential outlier flags assigned to each of the n_c gait cycles after applying the aforementioned five procedures. $F_j \in \{0, 1, 2, \cdots, 5n_s\}$, $j = 0, 1, 2, \cdots, n_c$, i.e. the largest number of flags per gait cycle is $5n_s$, while the higher the number of flags for a given gait cycle, the higher the likelihood that gait cycle is an outlier.

The Post-Flagging Decision Procedure (PFDP) is the last necessary procedure for final outlier detection. It is an iterative process that defines the optimum threshold number of possible outlier flags that would determine whether a gait cycle is an outlier or not. The procedure starts by defining a cutoff number of flags $C \in \{1, 2, \cdots, 5n_s\}$ (Step 1). Based on that choice of C, the gait cycles are separated into two groups. Group $\mathbb{N}$ includes data from the gait cycles that have a total number of flags less than C, while group $\mathbb{O}$ includes data from the gait cycles that have a total number of flags greater than or equal to C (Step 2). All data from all sensor sources from the corresponding gait cycles are included, i.e. each set $\mathbb{N}$ and $\mathbb{O}$ includes the corresponding rows of $\boldsymbol{A}$ as defined in (1). Let $\boldsymbol{N}$ and $\boldsymbol{O}$ be the $n_1 \times n_p$ and $n_2 \times n_p$ matrices with these data respectively, where $n_1 + n_2 = n_s n_c$. Then, a piece-wise variant of a two-tailed t-test with unequal means is implemented between the data in the rows of $\boldsymbol{N}$ and $\boldsymbol{O}$, for each of the n_p columns (Step 3). Let $q \in \{1, 2, \cdots, n_1\}$ denote a row in $\boldsymbol{N}$, and $r \in \{1, 2, \cdots, n_2\}$ denote a row in $\boldsymbol{O}$. The t-test performs a test of the hypothesis that the data from the l-th column of $\boldsymbol{N}$ and the l-th column of $\boldsymbol{O}$ come from distributions with equal means, and returns the result η_l, $l = 1, 2, \cdots, n_p$. A result $\eta_l = 0$ indicates that the null hypothesis of the means being equal cannot be rejected at the 5% significance level, while $\eta_l = 1$ indicates that the null hypothesis can be rejected at the 5% level. A total number of $n_p n_s$ tests are performed , and the results are summed up in a variable η given by $\eta = \sum_{\xi=1}^{n_p n_s} \eta_\xi$ (Step 4) . Then the process is iterated for another choice of C (back to Step 1), until all possible values of $C \in \{1, 2, \cdots, 5n_s\}$ are selected. The resulted values η for each possible value of C are included in a vector $\boldsymbol{\eta} = [\eta_1 \; \eta_2 \; \cdots \; \eta_{5n_s}]^T$. The value of C that results in the maximum η_v, $v = \{1, 2, \cdots, 5n_s\}$ is selected as the final cutoff number of flags C_f for outlier determination, i.e.

$$C_f \in \{1, 2, \cdots, 5n_s\} : \eta_f = max\left(\boldsymbol{\eta}\right) \tag{9}$$

where $max\left(\right)$ is the function that returns the maximum value of a vector.

III. Results

Objective performance of any outlier detection method is difficult to determine since true data outliers are not always obvious, and it is especially dependent on which lens the data is being observed from. To overcome this limitation, with the goal of quantifying performance metrics of the proposed algorithm, synthetic periodic data is proposed with intelligent creation of outliers within the dataset. The goal of this dataset is to be as realistic as possible, and as such, will contain outliers that will be challenging for any algorithm to find. This realistic dataset is inspired by real human gait data taken from vastus medialis EMG recordings, shown in Fig. 1, with natural variations that are expected and observed in real human trials.

A. Creation of Realistic Synthetic Dataset

The artificial but realistic dataset is created so that it has a total of 500 gait cycles. Out of the 500 cycles, 460 are normal, while the remaining 40 will be designed as outliers. All gait cycles have 100 samples taken at a frequency of 100 Hz, i.e. total duration of 1 s per gait cycle. A normal, i.e. non-outlier, gait cycle vector is generated by the following function:

$$x[t] = 1 - Qcos\left(2\pi f_c t\right) + Qsin\left(\pi f_c t\right) + \beta \tag{10}$$

where $t = 1, 2, \cdots, 100$ is the sample number, $x[t]$ is the sampled signal, $Q = 1$ is an amplitude factor, $f_c = 2Hz$ is the chosen frequency and β is a uniformly distributed random noise with values in the range $[0, 0.9]$. The resulted signal $x[t]$ is finally low-pass filtered with a low-pass 4-th order Butterworth filter and a cut-off frequency of 10 Hz in order to smooth the signal while not significantly affecting the contribution of β. The set of 460 normal artificial gait cycles generated is shown in Fig. 2 (black lines). Finally, an outlier gait cycle is generated by: $y[t] = x'[t] + Q' sin\left(\pi f'_c t\right)$, where $y[t]$ is the sampled outlier signal, $x'[t]$ is the low-pass filtered version of $x[t]$ i.e. the non-outlier signal, Q' is a scalar taking uniformly distributed random values in the range $[0.1, 1]$, and f'_c is a scalar taking uniformly distributed random values in the range $[0.5, 1.05]$. As it is seen, the artificial outlier gait cycles are generated from the normal gait cycles by adding a randomly conditioned sinusoidal function. Furthermore, the variation between artificial outliers include visible differences in time, frequency, amplitude and overall shape. These variations are intentionally designed to be subtle, and quite similar to the non-outliers, making it more challenging to separate them. The set of 40 artificial outlier gait cycles generated is shown in Fig. 2 (red lines). Because this work is an explanation of the proposed ODA procedure, it is not necessary to validate with an artificial dataset containing mutliple sources. In fact, as more sources are added to the algorithm, the accuracy and validity of the proposed ODA increases, and as a result, performance

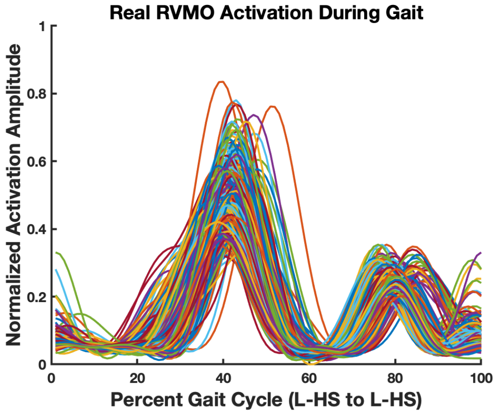

Fig. 1: Real Right Vastus Medialis Oblique (RVMO) EMG activation during treadmill walking. 500 gait cycles are plotted, while the duration of each gait cycle is normalized from Left Heel Strike (LHS) to the next LHS.

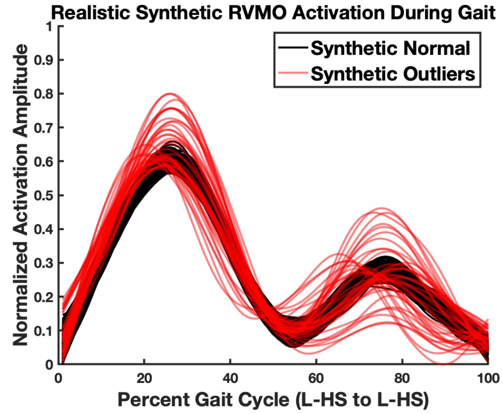

Fig. 2: Synthetic data of EMG activation, resembling real data shown in Fig. 1. 500 gait cycles are plotted. Both normal (460 black lines) and outlier (40 red lines) gait cycle data are shown. The duration of each gait cycle is normalized from Left Heel Strike (LHS) to the next LHS.

metrics from this particular dataset is considered a worst-case scenario. As such, this synthetic dataset is considered to be from one sensor source only.

B. Evaluation Using Realistic Synthetic Dataset

The five flagging procedures are applied to the above synthetic dataset of 500 gait cycles, and their results are input to the PFDP. For analysis purposes, in addition to the evaluation of the proposed method that includes all five flagging procedures, each flagging procedure is evaluated separately, the results of which are again input to the PFDP. The performance of each case is determined by counting how many outliers were found, as well as how many false positives (FPs) and false negatives (FNs) resulted. When the method flagged a gait cycle that was one of the 460 non-outlier gait cycles, that flag is considered a FP. Similarly, when the method failed to flag a gait cycle that was one of the 40 gait cycles created to be an outlier, that absence of flag is considered a FN. Finally, in order to evaluate the proposed method more extensively, 100 different sets of 500 synthetic gait cycles were generated using the procedure outlined in Section III-A, each randomly generated as analyzed above. The proposed method was evaluated in all of those 100 different datasets.

Table I summarizes the results of the evaluation across all the 100 generated datasets. Since each dataset has 40 outlier gait cycles, there are a total of 4000 outlier gait cycles. As it can be seen, the Shape-based method alone detected more outlier gait cycles (3728 (93.2%)) compared to the proposed ODA (3648 (91.2%)), however it flagged over four-times the amount of False Positives (263), compared to how many the proposed method did (60). This highlights the value of the proposed ODA in balancing the trade-off between maximizing the number of found outliers and minimizing the number of mischaracterized non-outliers. Table I also shows that the other methods alone did not perform as well as the proposed ODA. Overall, the proposed method performed very well (91.2% accuracy) in an extensive dataset, while not over-flagging the data, resulting in only 60 false positives out of 50,000 total gait cycles (0.1%), and less combined FN's and FP's (412) than all other methods individually. We expect that the method will be even more accurate if additional sensor sources data are used, as in real gait data analysis.

The performance of the individual flagging methods and how their combination is effective in detecting outliers is further analyzed. For brevity, an example with two flagging methods is provided. The performance of the Feature-Based Procedure that utilized the PCA for dimensionality reduction in one set of 500 synthetic gait cycles is shown in Fig. 3. As it can be seen, most of the outliers were correctly flagged, except from a few (7 out of 40), that were very close to the normal gait cycles, when projected in the low-dimensional space. However, most of those missed outliers (False Negatives) were flagged by the other flagging procedures. Specifically for this dataset, the Amplitude-Based Procedure that utilized the Integral values was able to detect all 7 of those gait cycles that the PCA-based method failed to detect. The results are shown in Fig. 4, where all the gait cycles are represented with their computed integral values. Although these values are scalars, i.e. one-dimensional data, a 2D scatter plot is created with both axes being the single-dimensional integral score to better visualize the gait cycle selections and to be able to compare them to Fig. 3. As seen in Fig. 4, although this procedure had many false negatives, all 7 gait cycles that were missed by the PCA-method were flagged as outliers using this method, noted with the magenta squares. This is an example of the importance of analyzing the data from different perspectives in order to improve the overall accuracy, a strength of the proposed method.

IV. Conclusion

This paper introduces a unique algorithm for outlier detection in periodic gait data using multiple sources and multiple procedures to improve the overall accuracy. It is shown that the proposed method is able to detect 91.2% of the true outliers in an extensive synthetic dataset, while only producing false positives at a rate of 0.1%, outperforming other procedures usually utilized in gait data outlier detection. The proposed method is a systematic way of removing outliers from gait data, with direct applications to human

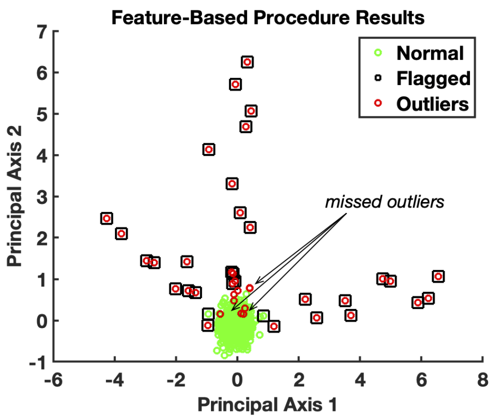

Fig. 3: Results using the Feature-Based Procedure using PCA for dimensionality reduction in a synthetic dataset. The low-dimensional representation of the data using only the first two principal components (see Sec. II.B.2) is shown. Green and red circles (○) represent normal (non-outlier) and outlier gait cycles respectively. Gait cycles in black squares (□) represent the gait cycles flagged as outliers by this method.

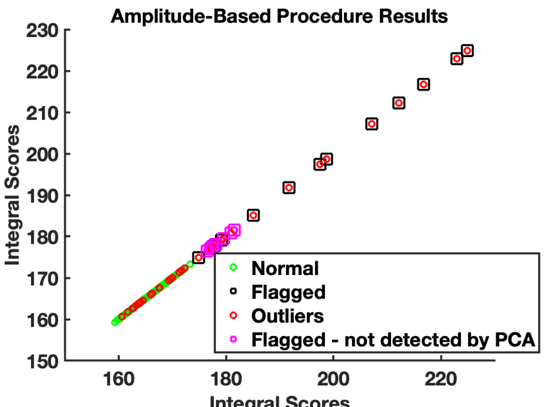

Fig. 4: Results using the Amplitude-Based Procedure using the Integral values in a synthetic dataset. The scalar integral values for each gait cycle are shown in two dimensions for clarity. Green and red circles (○) represent normal (non-outlier) and outlier gait cycles respectively. Gait cycles in black squares (□) represent the gait cycles flagged as outliers by this method. Gait cycles in magenta squares (□) represent the outliers that were missed by the Feature-Based Procedure noted in Fig. 3.

biomechanics, rehabilitation and robotics, while it can be applied to other scientific fields dealing with periodic data.

References

[1] V. J. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial Intelligence Review 2004 22:2*, vol. 22, pp. 85–126, 10 2004.

[2] F. E. Grubbs, "Procedures for detecting outlying observations in samples," *Technometrics*, vol. 11, pp. 1–21, 1969.

[3] D. M. Hawkins, "Identification of outliers," *Identification of Outliers*, 1980.

[4] E. A. Clancy, E. L. Morin, and R. Merletti, "Sampling, noise-reduction and amplitude estimation issues in surface electromyography," *Journal of electromyography and kinesiology*, vol. 12, no. 1, pp. 1–16, 2002.

[5] E. Stålberg, C. Bischoff, and B. Falck, "Outliers, a way to detect abnormality in quantitative emg," *Muscle and Nerve*, vol. 17, pp. 392–399, 1994.

[6] J. I. Furukawa, T. Noda, T. Teramae, and J. Morimoto, "Estimating joint movements from observed emg signals with multiple electrodes under sensor failure situations toward safe assistive robot control," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, pp. 4985–4991, 6 2015.

[7] A. Ghebleh and M. E. Moghaddam, "Clothing-invariant human gait recognition using an adaptive outlier detection method," *Multimedia Tools and Applications*, vol. 77, pp. 8237–8257, 4 2018.

[8] P. Reberšek, D. Novak, J. Podobnik, and M. Munih, "Intention detection during gait initiation using supervised learning," *IEEE-RAS International Conference on Humanoid Robots*, pp. 34–39, 2011.

[9] M. J. Vera, B. Dubravka, J. Nikola, I. Vojin, and P. B. Bojana, "Detecting and removing outlier(s) in electromyographic gait-related patterns," *Journal of Applied Statistics*, vol. 40, pp. 1319–1332, 6 2013.

[10] V. Barnett and T. Lewis, "Outliers in statistical data 3rd edition," *Biometrical Journal*, vol. 37, pp. 256–256, 1 1995.

[11] F. E. Grubbs, "Sample criteria for testing outlying observations," *The Annals of Mathematical Statistics*, vol. 21, pp. 27–58, 3 1950.

[12] D. van Hieu and P. Meesad, "A fast outlier detection algorithm for big datasets," *Advances in Intelligent Systems and Computing*, vol. 463, pp. 159–169, 2016.

[13] X. T. Wang, D. R. Shen, M. Bai, T. Z. Nie, Y. Kou, and G. Yu, "An efficient algorithm for distributed outlier detection in large multi-dimensional datasets," *Journal of Computer Science and Technology 2015 30:6*, vol. 30, pp. 1233–1248, 11 2015.

[14] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *IEEE Access*, vol. 7, pp. 107 964–108 000, 2019.

[15] E. Ramirez, M. Wimmer, and M. Atzmueller, "A computational framework for interpretable anomaly detection and classification of multivariate time series with application to human gait data analysis," *Lecture Notes in Computer Science*, vol. 11979 LNAI, pp. 132–147, 6 2019.

[16] A. Zhang, S. Yang, X. Zhang, J. Zhang, and W. Zhang, "Abnormal gait detection in surveillance videos with fft-based analysis on walking rhythm," *Lecture Notes in Computer Science*, vol. 10666 LNCS, pp. 108–117, 2017.

[17] B. A. Sanford, A. R. Zucker-Levin, J. L. Williams, W. M. Mihalko, and E. L. Jacobs, "Principal component analysis of knee kinematics and kinetics after anterior cruciate ligament reconstruction," *Gait and Posture*, vol. 36, pp. 609–613, 7 2012.

[18] K. Schweizer, P. C. Cattin, R. Brunner, B. Müller, C. Huber, and J. Romkes, "Automatic selection of a representative trial from multiple measurements using principle component analysis," *Journal of Biomechanics*, vol. 45, pp. 2306–2309, 8 2012.

[19] M. Sangeux and J. Polak, "A simple method to choose the most representative stride and detect outliers," *Gait and Posture*, vol. 41, pp. 726–730, 2 2015.

[20] C. Karakasis and P. Artemiadis, "Real-time kinematic-based detection of foot-strike during walking," *Journal of Biomechanics*, vol. 129, p. 110849, 2021.

[21] C. Karakasis and P. Artemiadis, "F-VESPA: A kinematic-based algorithm for real-time heel-strike detection during walking," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 5098–5103.

[22] F. E. Zajac, "Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control." *Critical reviews in biomedical engineering*, vol. 17, no. 4, pp. 359–411, 1989.

[23] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

	Shape-based	**Feature-based**	**Time-based**	**Amplitude-based**	**Statistics-based**	**Proposed ODA**
Detected outliers (of 4000)	**3728 (93.2%)**	3047 (76.2%)	3185 (80%)	1829 (45.8%)	3566 (89.2%)	3648 (91.2%)
False Negatives (FN's) (of 4000)	**272 (6.8%)**	953 (23.8%)	815 (20.4%)	2171 (54.3%)	434 (10.9%)	352 (8.8%)
False Positives (FP's) (of 46000)	263 (0.57%)	113 (0.25%)	366 (0.80%)	72 (0.16%)	**12 (0.03%)**	60 (0.13%)
Total FN's and FP's (of 50000)	535 (1.07%)	1066 (2.13%)	1181 (2.36%)	2243 (4.49%)	446 (0.89%)	**412 (0.82%)**

TABLE I: Evaluation of each separate flagging method and the proposed outlier detection algorithm (ODA).