Muscle activation patterns during gait: A hierarchical clustering analysis

Samanta Rosati, Valentina Agostini*, Marco Knaflitz, Gabriella Balestra

Department of Electronics and Telecommunications, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy

A R T I C L E   I N F O

Article history:
Received 29 February 2016
Received in revised form 8 September 2016
Accepted 23 September 2016

Keywords:
Gait analysis
Surface EMG
Muscle activation patterns
Hierarchical clustering
Dendrogram

A B S T R A C T

Human gait is characterized by a large stride-to-stride variability of the muscle activation patterns (onset-offset timings). For this reason prolonged walking sessions lasting several minutes are analyzed. To interpret correctly the electromyographic (EMG) data collected during gait, it is important to group strides sharing similar EMG activation patterns. The aim of this work is to present and validate a method, based on hierarchical clustering, able to group strides showing homogeneous onset-offset activation intervals. Results show that the variability of the onset-offset timing is significantly reduced after clustering, for all of the five lower limb muscles considered to test this method. A by-product of the clustering procedure is the possibility to define and extract the principal activations of a muscle during gait. We define principal activations those activations that are necessary for the specific muscle contribution to the biomechanical function of walking. This concept may be useful whenever the dynamic performance of the muscle has to be compared in subsequent times, such as in patient’s follow-up or when the performance of a specific subject is to be compared to that of a group of selected subjects. The contribution presented in this work could be beneficial in implementing a personalized medicine approach to rehabilitation. Clinical gait analysis, enriched by hierarchical clustering of EMG patterns as well as by the quantitative assessment of muscles principal activations, could greatly contribute to the design of therapeutic treatments tailored on the patient’s needs.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Gait analysis is used to quantitatively assess the normal and pathological functions of human walking [1]. The electromyographic (EMG) signal provides information on muscle activity during gait [2]. Muscle activation intervals can be obtained from the raw EMG signal, detecting when the EMG signal is above the background-noise level (ON) or when it is below (OFF) [3]. However, human locomotion is characterized by a large intra-subject variability [4]: muscle activation onsets and offsets markedly vary from stride-to-stride also in a specific individual with normal gait [5,6]. For this reason, previous literature highlighted the importance of analyzing prolonged walks, lasting at least 2–3 min, instead of considering a few strides, to reach reliable and repeatable measurements, both in normal and pathological gait [5–13].

The study of prolonged walks requires the use of powerful tools, able to automatically analyze hundreds of strides, like “statistical gait analysis” [5–13]. Typically, the EMG signal is segmented into separate strides, identifying the start and end of each stride by timing gait events through foot-switches [9]. In an attempt to control the large stride-to-stride variability of EMG patterns, only the strides sharing the same foot-floor contact sequences are processed together [5–8,10–13]. On the contrary, the strides in which the foot contacts the ground in a non-standard manner are analyzed separately, or treated as outlier strides (e.g., in presence of forefoot strikes instead of heel strikes) [9]. The reason for this choice is that strides with different foot-floor contacts are not expected to share the same EMG ON-OFF patterns [8,10]. In normal gait, only the strides that show the standard sequence of gait phases (Heel contact, Flat foot contact, Push-off, Swing, i.e. HPFS cycles) are considered [5–8,11–13].

However, even if the EMG signal is processed separately for each foot-floor contact sequence, e.g. considering only HPFS cycles for normal subjects, there are other sources of EMG intra-subject variability that must be accounted for. More specifically, literature reports that, even in normal gait cycles of normal subjects, a single subject’s muscle does not show a single “preferred” pattern of activation, but up to 4–5 distinct EMG patterns, each characterized by a different number of activation intervals occurring within the stride [5,11]. An example of intra-subject EMG variability in a

* Corresponding author.
E-mail address: valentina.agostini@polito.it (V. Agostini).

http://dx.doi.org/10.1016/j.bspc.2016.09.017
1746-8094/© 2016 Elsevier Ltd. All rights reserved.
young healthy subject is reported in Fig. 1. In the effort of looking for homogeneity, only the strides showing the same number of activation intervals \((m)\) are processed together, grouped in the so-called \(m\)th-activation modality \([5,12,13]\). Even focusing on a specific modality, it is possible that the activation intervals differ among all the strides belonging to the modality. There may be patterns with a distinct biomechanical meaning within the same activation modality. Hence, to correctly interpret EMG patterns it is important to adopt proper tools for grouping cycles that share the same timing pattern, while keeping separate cycles with different onset-offset timing.

Clustering designates a class of methods able to identify homogeneous groups of unlabeled elements. These methods are based on a distance metric that gives a measure of the similarity or dissimilarity among objects. A good clustering is obtained when the variability among the objects belonging to the same cluster is as low as possible, while clusters are clearly distinct. In literature, applications of clustering methods to EMG signal analysis are very limited and are ascribable to three main applications: analysis of subjects' populations \([14–16]\), segmentation \([17]\) or decomposition of EMG signals \([18,19]\), and analysis of within-session and day-to-day-session subject variability \([20]\). The stride-to-stride variability in the onset-offset EMG patterns was never approached with a clustering algorithm before. In particular, to the best of our knowledge, dendrogram clustering was never applied to EMG data processing.

Dendrogram is an agglomerative hierarchical clustering method that organizes data in a hierarchical tree based on a proximity measure \([21]\). This algorithm begins by assigning each object to a separate cluster and proceeds by joining, at each step, the two most similar clusters, until a single cluster is obtained containing all the objects. The final clusters are obtained by “cutting the tree” at a certain level according to criteria related to the distance among clusters. The main advantage of this method, with respect to partitional clustering algorithms such as k-means, is that there is no need to set a-priori the final number of clusters: this automatically emerges from the cutting of the tree.

The aim of this work is to present and validate a method, based on hierarchical clustering, able to group strides showing homogeneous onset-offset activation intervals. An important spin-off of the method herein presented is the possibility of extracting the principal activations of a muscle in a repeatable and user independent way.

2. Materials and methods

2.1. Subjects and experimental protocol

Nine healthy subjects (age range: 20–25 years) were asked to walk barefooted at self-selected speed back and forth along a 15-m walkway for 3 min, in order to collect at least 150 gait cycles for each lower limb, allowing for a statistical analysis of gait \([5,11,22,23]\).

The system STEP32 (Medical Technology, Italy) was used to acquire foot-switch signals and surface EMG signals \([5,11–13]\). Foot-switches were placed under the foot-soles (size: 10 mm × 10 mm × 0.5 mm; activation force: 3 N), beneath the first and fifth metatarsal heads, and beneath the back portion of the heel. After skin preparation, surface EMG probes were placed over the muscle’s belly of tibialis anterior (TA), gastrocnemius lateralis (GL), peroneus longus (PL), rectus femoris (RF), and lateral hamstring (LH), bilaterally \([11]\). These muscles were selected in order to study at least a pair of agonist-antagonist muscles acting at each joint of the lower limb. EMG probes were active and utilized AgCl-disks as electrodes (probe size: 27 mm × 19 mm × 7.5 mm, inter-electrode distance: 12 mm). The signal amplifier had a gain of 1000 and a 3-dB bandwidth from 10 Hz to 400 Hz. The sampling frequency was 2 kHz and signals were converted by a 12-bit analog to digital converter.

The experimental protocol was approved by the local ethical committee and all participants gave their written informed consent to participate in the study.

2.2. Proposed method

We present a method for grouping strides with similar EMG patterns. The method consists of two steps: 1) pre-processing, 2) application of the CIMAP algorithm (Clustering for Identification of Muscle Activation Patterns).

2.2.1. EMG signal pre-processing

EMG signals are pre-processed as follows.

A) Signal segmentation

The start and end of each stride is identified by timing gait events through foot-switches \([9]\). EMG signals are then segmented into separate strides (Fig. 1) and time-normalized to the stride duration. B) Stride extraction

Only strides belonging to a specific foot-floor contact type are considered. As an example, in healthy subjects we consider only the strides showing the normal sequence of gait phases (Heel contact, Flat foot contact, Push off, Swing), discarding those with non-standard foot-floor contact sequences \([9]\). This is performed by using the 4-level foot-switch signal, as it is detailed in \([9]\).

Among the selected strides, those related to deceleration, turning, and acceleration in correspondence of direction changes are discarded using a multivariate statistical filter \([7,11]\).

C) Extraction of activation intervals from each stride

For each stride, the ON/OFF activation intervals (see Fig. 1) are detected by means of the double threshold statistical detector described in \([3]\). The two thresholds are selected taking into account the background noise level and signal-to-noise ratio of the EMG signal under study. These signal parameters are automatically estimated as described in \([24]\).

Hence, we obtain a collection of strides and, for each stride, we compute the onset-offset timing of EMG activity.

2.2.2. CIMAP algorithm: clustering for identification of muscle activation patterns

Considering a specific muscle, the algorithm consists of two phases: a) datasets preparation, and b) clustering of each dataset.

A) Datasets preparation

Left and right strides are pooled together. Strides are divided in several datasets grouping strides that show the same activation modality.

Every stride \(i\) of the dataset is characterized by \(m\) couples of time instants \((ON_i, OFF_i)\) where \(m\) is the number of activations within the stride:

\[
\text{stride}_i = \{ ON_{i,1}, OFF_{i,1}, \ldots, ON_{i,j}, OFF_{i,j}, \ldots, ON_{i,m}, OFF_{i,m} \} ,
\]

Fig. 1. Surface EMG signal from the tibialis anterior muscle of a representative subject (right side): the picture displays three consecutive gait cycles. The first gait cycle shows 3 activations (ON1, OFF1, ON2, OFF2, ON3, OFF3), the second gait cycle shows 4 activations, the third gait cycle shows 2 activations.
with \( i = 1:n \), being \( n \) is the total number of strides enclosed in the dataset. The total number of strides varies from dataset to dataset. Only datasets consisting of at least 50 strides are considered.

b) Clustering of each dataset

We apply dendrogram clustering to every dataset separately. At the beginning, each stride is considered as a single-element cluster. Then, after each iteration, the two closest clusters are merged.

More specifically, every cluster \( A \) is represented by its centroid \( CLC_A \) that is defined as a 2\( m \)-dimensional vector containing the mean of the ON and OFF timings of the strides inside the cluster:

\[
CLC_A = \frac{1}{|A|} \sum_{a_i \in A} d_i = \left\{ ON_1, OFF_1, \ldots, ON_j, OFF_j, \ldots, ON_m, OFF_m \right\},
\]

where \( |A| \) represents the cardinality (number of strides) of cluster \( A \) and \( a_i \) is the \( i \)-th element (stride) belonging to the considered cluster.

Then, at each iteration, the two closest clusters are merged. The similarity measure chosen to merge two clusters \( A \) and \( B \) is the L-infinity norm, calculated as:

\[
d( CLCA, CLCB ) = \max \left( \frac{ \max_{m} ( ON_{A} - ON_{B} ) }{ ON_{m} , OFF_{m} } \right)
\]

L-infinity norm was preferred to other norms since it avoids merging strides showing (overall) similar timings except for a single, markedly different, onset (or offset).

To cut the dendrogram and identify the appropriate number of clusters we applied an automatic rule. This rule is based on the idea that the dendrogram must be cut when are merged two clusters having a comparable number of elements, small intra-cluster variability, but great distance one with respect to the other.

In this study, the intra-cluster variability is computed as the sum of all the Euclidean distances between each element (stride) belonging to the cluster and the cluster’s centroid.

We defined an index \( R_k \), associated to each iteration \( k \), that takes into account only the two clusters \( A \) and \( B \) that are merged in that iteration:

\[
R_k = \frac{ \sum_{ \text{strides} \in \{ A,B \} } \text{dist} ( \text{stride}, CLC_A/B ) }{ \max \left( \sum_{\text{strides} \in A} \text{dist} (\text{stride}, CLC_A), \sum_{\text{strides} \in B} \text{dist} (\text{stride}, CLC_B) \right) }
\]

It represents the ratio between the intra-cluster variability of the new formed cluster \( CLC_A/B \) and the maximum between the intra-cluster variability of the two original clusters \( A \) and \( B \).

Since at the beginning the algorithm merges strides that are similar, the highest \( R_k \) values are obtained in the final iterations. For this reason, we compute \( R_k \) only for the last 20% iterations. We identify the iteration \( k \) with the maximum \( R_k \) value and cut the dendrogram just before it, considering the clusters obtained for iteration \( k-1 \).

An example of EMG activation clustering is reported in Fig. 2, for the TA muscle, applied to the dataset involving strides with 3 activation intervals.

Once the clustering of the considered dataset is concluded, for each cluster, left and right strides are separated into two different clusters. For each side (left/right), clusters with less than 5% of the original number of strides (of the side under consideration) are discarded.

2.3. Extraction of the principal activations

The clustering procedure can provide many different activation patterns. However, in certain applications it may be useful to extract the muscle principal activations common to all clusters. By principal activation we mean an activation that is necessary for the biomechanical task that is being actuated by the specific muscle. This is complementary to the concept of secondary activations, which are activations present only in some strides, and have an auxiliary function, e.g., to correct muscle activation under a slightly modified environmental input.

Hence, after clustering, a post-processing phase follows in which all valid clusters from one side (left or right) are considered together to define the principal activations of a specific muscle. In particular, for each cluster, the cluster prototype is calculated as the cluster centroid. Then, the principal activations are extracted by intersection of the clusters centroids.

2.4. Validation

To validate the proposed method we demonstrate that it is able to group similar strides, achieving small intra-cluster variability. The method is validated on the EMG datasets of the 9 healthy subjects comparing the average intra-cluster variability among strides for the analyzed muscles (TA, GL, PL, RF, LH) with the average variability among strides when no clustering is applied.

More specifically, for each cluster we defined the intra-cluster variability as the average standard deviation (SD) of the ON and OFF timings among the cluster’s strides:

\[
\text{Cluster SD} = [ SD( ON_1 ), SD( OFF_1 ), \ldots, SD( ON_m ), SD( OFF_m ) ]
\]
Onsets at 0% of the stride and offsets at 100% of the stride were excluded since they showed null SD. Then, for each subject’s lower limb, the average Cluster_SD was calculated across all modalities and clusters. Then, the left and right side values were also averaged.

The variability among strides sharing the same activation modality was calculated in the same manner. For each modality, the intra-modality variability was defined as the average SD of the ON and OFF timings among the modality’s strides:

\[
\text{Modal}_{SD} = [\text{SD}(\text{ON}_1), \text{SD}(\text{OFF}_1), \ldots, \text{SD}(\text{ON}_m), \text{SD}(\text{OFF}_m)]
\]

For each subject’s lower limb, the average Modal_SD was calculated across all modalities. Then, the left and right side values were also averaged.

Boxplots obtained with clustering (Cluster_SD) and without clustering (Modal_SD) were used to compare the distributions of the average variability across the 9-subject population, for each muscle.

Moreover, for each muscle, a t-test was performed to evaluate the differences between the average variability obtained with and without clustering. The significance level (\(\alpha = 0.05\)) was adjusted for multiple comparisons (5 muscles) with Bonferroni’s correction (\(\alpha = 0.01\)).

3. Results

Subjects walked at 1.2 ± 0.1 m/s (self-selected speed). For each lower limb, an average of 164 ± 9 strides showing the standard foot-floor contact was analyzed. After removing outlier strides an average of 110 ± 15 valid strides was obtained for each lower limb. For each subject, the complete left/right EMG datasets were clustered (220 ± 15 strides).

3.1. Clustering results

Fig. 3 shows results from a representative subject for two antagonist muscles: TA and GL. The number of clusters and the number of stride-elements within each cluster varies from dataset to dataset. For each cluster, the cluster prototype was calculated as the centroid. Then, the muscle principal activations were obtained by intersection of the cluster prototypes.
Table 1

Variability in the Onset-offset Activation Timings Among Strides Belonging To a Cluster or a Modality.

<table>
<thead>
<tr>
<th>Muscle</th>
<th>Variability with clustering (CLUSTER_SD; % stride)</th>
<th>Variability without clustering (MODAL_SD; % stride)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.7 ± 0.8'</td>
<td>10.1 ± 2.3'</td>
<td>0.000003</td>
</tr>
<tr>
<td>TA</td>
<td>3.2 ± 1.0'</td>
<td>14.4 ± 3.1'</td>
<td>0.00002</td>
</tr>
<tr>
<td>LGS</td>
<td>3.0 ± 0.6'</td>
<td>7.0 ± 1.3'</td>
<td>0.000001</td>
</tr>
<tr>
<td>RF</td>
<td>3.8 ± 1.3'</td>
<td>11.8 ± 2.6'</td>
<td>0.0002</td>
</tr>
<tr>
<td>PL</td>
<td>3.3 ± 0.8'</td>
<td>9.2 ± 3.3'</td>
<td>0.0001</td>
</tr>
<tr>
<td>LH</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Values reported are mean ± SD over the 9-subject population (left and right muscles were averaged). *Significant differences between the variabilities obtained with and without clustering (p < 0.01).

Fig. 4. Comparison of the variability in the activation onset/offset among strides grouped in (a) activation modalities (no clustering applied) (light gray) or (b) clusters (dark gray), for five lower limb muscles (tibialis anterior (TA), gastrocnemius lateralis (LGS), rectus femoris (RF), peroneus longus (PL), lateral hamstrings (LH). Each boxplot represents median, 25th and 75th percentile, maximum and minimum across the sample of 9 healthy subjects. Left and right muscles were averaged.

Fig. 4 shows the boxplots of the variability before and after clustering. The mean value and standard deviation of the variability obtained with and without clustering, is reported in Table 1. The variability is significantly reduced after clustering (p < 0.01), for all the muscles considered. It is evident that the clustering procedure was successful in grouping strides with homogeneous activation patterns: the average intra-cluster variability (Cluster_SD) was, overall, smaller than 5% of stride duration.

3.2. Choice of parameters

We performed a series of tests to properly choose the CIMAP parameter values and we obtained evidence supporting the following choices:

- minimum number of strides to be considered for each modality of activation (50 strides, considering left and right strides together);
- percentage of iterations from which to start calculating $R_k$ (last 20% iterations);
- minimum cluster size, expressed as percentage of original number of strides in each dataset (5%), to discard non-representative clusters.

Notice that the left and right strides have been pooled together to obtain a higher number of strides for the clustering algorithm. However, the algorithm can work also considering the strides of a single lower limb at a time, if a sufficient number of strides are present.

4. Discussion

In general, gait pattern classification may assist in clinical decision making and cluster analysis has been often adopted to this aim [15–17,25–28]. However, cluster analysis was seldom applied to surface EMG signal processing [14,17–20]. In particular, some works focused on the decomposition of surface EMG signal into motor unit action potentials (MUAPs), based on hierarchical clustering algorithm [18], or partially based on the $k$-means algorithm, and implementing a supervised classification using a certainty-based algorithm [19]. In the past, a technique for clustering linear envelopes of the EMG during gait was proposed by [14]. Another study also analyzed EMG linear envelopes, considering a multi-dimensional representation of walking dynamics which used a hierarchical clustering procedure [20]. They observed that relatively brief sessions, containing about 20 gait cycles, may be insufficient to capture the complete variety of gait patterns that may be produced by a subject, even when the gait condition is tightly controlled. More recently, abandoning the linear envelope approach, $k$-means clustering was applied to the onset-offset detection of muscle activity [17].

Although previous research recognized the importance of studying the various patterns of activation of a muscle during a subject’s walk [5,8,12], a systematic approach for grouping strides with homogeneous onset-offset timing patterns was lacking. This contribution fills the gap, presenting a method for clustering muscle activation patterns recorded from numerous strides during human gait.

4.1. Clustering muscle activation patterns

The method presented is relevant for a correct interpretation of EMG data in gait analysis. This requires to take into account the different activation modalities separately, and to characterize the different timing patterns of each modality. This method allows for obtaining all the representative activation patterns of a subject’s muscle, each corresponding to a cluster prototype. Furthermore, the cluster size (the number of strides composing the cluster), normalized to the total number of strides, may be an additional relevant clinical information, since it indicates how frequently a specific EMG pattern is found [7]. On the other hand, the cluster size is used by the CIMAP algorithm to remove outlier cycles. Clusters with a small size, corresponding to EMG patterns that are very rarely observed, are automatically discarded.

For each activation modality, we found that various valid cluster prototypes may be obtained, confirming the richness of muscle activation patterns already documented in human gait [5–8,11–13]. Recently, the neuroscience community has paid considerable attention to the so called “muscle synergies”, to represent the CNS neural strategies for simplifying multi-muscle control (see [29] for a review). In this framework, it is not surprising that a single muscle shows different activation patterns while performing the same task. Nevertheless, in traditional approaches to gait analysis, ensemble averages are carried out across strides [see 30] for a review), mixing different (unrecognized) activation patterns. This can lead to misleading interpretations of EMG data. By contrast, the methodology presented here, allowing for clustering strides that share similar timing patterns, is ideally suited to group strides in which the muscle’s actuation by the CNS neural command is similar.
4.2. Principal and secondary activations

Principal activations are those muscle activations intrinsically necessary to the motor task under consideration. In the framework of the presented method, we defined principal activations as the activations common to all clusters prototypes. As an example, in the results reported on a representative subject (see Fig. 3), the principal activations of the TA muscle include a burst of activity around initial contact to control plantar flexion. In this specific case, the burst ranged between 86% and 5% of the stride (for both lower limbs). In literature this is known to be a strong activation aimed at decelerating the passive plantar flexion of the ankle caused by the weight of the body being applied on the heel [1]. A second burst of activity occurs during the swing phase with the purpose of producing a sufficient dorsiflexion of the ankle to clear the toes from the ground. In the specific case analyzed, this burst ranged between 57% and 76% of stride, for the left side, and between 56% and 74% of stride, for the right side, demonstrating a highly symmetrical activity. The GL muscle shows a single principal activation to produce propulsion just before toe-off, between 23% and 45% of stride, for the left side, and between 29% and 47% of stride, for the right side. Again the biomechanical relevance of this burst is well documented in literature [1].

However, both in TA and GL (as well as in the other muscles analyzed), it may also be noticed the presence of secondary or auxiliary activations. These are activations that are present only in some of the clusters, but not in all of them. This finding can be explained by the redundancy inherent to motor control, which makes it possible to execute movements in a multitude of different ways while achieving equivalent outcomes [31].

In the clinical management of patients and, more specifically, in clinical gait analysis, it is important to follow-up a patient’s performance in time, or to compare some outcome measures on a cohort population before and after a certain treatment or intervention. In these types of applications, it may be useful to obtain global indexes [32] or scores quantifying the patient’s performance at a specific time [33]. This can be particularly challenging when EMG measures are involved, because of the multiple patterns present. From this perspective, it may become useful the concept of principal activations that we have introduced here. In fact, it allows for quantifying, by means of a single string of ON/OFF intervals, the “main” or “necessary” bursts of activation, representative of the essential functions of the muscle under study. This may help the comparison of the muscle’s function of a subject in successive times, or against a reference group.

5. Conclusions

This contribution introduced a clustering procedure successful in grouping strides with homogeneous EMG onset-offset activation patterns. The method is independent from the muscle considered and it may provide a correct interpretation of EMG data collected during a walk lasting several minutes.

The concept of muscle principal activations was introduced for the first time, in contrast to secondary (or auxiliary) activations. In clinical gait analysis, the quantitative evaluation of muscles principal activations can be a useful tool for patient’s follow-up, to personalize therapeutic interventions or to compare a subject’s performance against a reference group.

Furthermore, both the clustering method and the concept of principal activations may be extended to the study of other EMG cyclostationary signals, and applied not only in the rehabilitation setting, but also in human movement science, ergonomics and sport.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jsbpc.2016.09.017.

References


